SOCIAL COMPUTING IN BLOGOSPHERE

by

Nitin Agarwal

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

August 2009
SOCIAL COMPUTING IN BLOGOSPHERE

by

Nitin Agarwal

has been approved

August 2009

Graduate Supervisory Committee:

Huan Liu, Chair
Yi Chen
Hasan Davulcu
Jieping Ye

ACCEPTED BY THE GRADUATE COLLEGE
ABSTRACT

Social computing is defined as computing through social media. It refers to the endeavor to understand complex human interactions in social media like blogs, social networking services, wikis, social bookmarking (folksonomies), and online media sharing through computational means. Research in social computing builds on participatory Web characterized by rich Web applications, user generated contents, user enriched contents, user developed widgets, and collaborative environment of participatory web and citizen journalism. Social media has observed a phenomenal growth in past few years. This work focuses on studying the categories of social media and characteristics that make social media immensely popular. Social computing presents both challenges and opportunities. It is a vibrant and fledgling field with many research challenges including phenomenal growth, dynamism, long tail phenomenon, sparse link structure, lack of ground truth, information quality, and data collection. This thesis focuses on the research in Blogosphere – the network of Web logs. Research towards these challenges is presented in the context of the blogosphere, and motivated by the need for identifying influential bloggers in communities, extracting clusters in blogs, and searching for “familiar strangers” in egocentric networks. Novel solutions such as using the network and content information available at the blogosphere simultaneously, leveraging the invaluable and extremely dynamic collective wisdom of the bloggers, constructing social identity of bloggers based on their group affiliations, and providing an evaluation framework that leverages the power of social media to these problems are presented aiming at understanding individuals, communities and their interactions, paving the way for further research and development.
To My Dear Parents
ACKNOWLEDGMENTS

It has been a great pleasure working with the faculty, staff, and students at the Arizona State University during my tenure as a doctoral student. This would have never been possible without the freedom I was given to pursue my research interests by my mentor and advisor, Prof. Huan Liu. I would like to thank him for his strong belief in me, invaluable feedback, kind advice, and incessant support all these years.

I would also like to thank all the members of Data Mining and Machine Learning group at Arizona State University. Each and every member of the group has helped me in different ways by providing useful feedback and insightful discussions.

My research has also benefited tremendously from various collaborations over the years. I would particularly like to thank Prof. Yi Chen (at Arizona State University), Dr. John J. Salerno (at Air Force Research Laboratory), Prof. Arunabha Sen (at Arizona State University), Prof. Mark Woodward (at Center of Religious Studies, Arizona State University), Prof. Philip S. Yu (at University of Illinois at Chicago), and Dr. Jianping Zhang (at MITRE) for many thoughtful conversations.

I would like to thank Michael J. Sullivan, Executive Director of Hispanic Research Center at Arizona State University whose inspiring words and tremendous help is one of the major factors that contributed towards this achievement. I would particularly like to thank him for the thoughtful discussions not only about my research but also an all-round development.

In large part, my dissertation research has been sponsored by grants from Air Force Office of Scientific Research (FA95500810132) and Office of Naval Research (N000140810477 and N000140910165).
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ix</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>xi</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 1</th>
<th>INTRODUCTION TO SOCIAL COMPUTING</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 2</th>
<th>IDENTIFYING INFLUENTIAL BLOGGERS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

2.1. Introduction

2.1.1. Applications of the Influentials

2.2. Influential Bloggers: Problem and Definition

2.3. Identifying the Influentials

2.3.1. An Initial Set of Intuitive Properties

2.3.2. Developing the Model

2.3.3. iFinder - a Preliminary Model

2.3.4. Computing Blogger Influence with Matrix Operations

2.3.5. Issues of identifying the influentials

2.4. Experiments & Further Study

2.4.1. Data Collection

2.4.2. Results and Discussions

2.4.2.1. Influential Bloggers and Active Bloggers

2.4.2.2. Evaluating the Model

2.4.2.3. Influential vs. Non-Influential Blog Posts

2.4.2.4. Effects and Usages of Weights

2.4.2.5. iFinder vs. PageRank

2.4.2.6. Temporal Patterns of the Influentials

2.4.2.7. Further Experiments

2.5. Related Work

vi
CHAPTER 3  CLUSTERING BLOGS BY LEVERAGING COLLECTIVE WISDOM  

3.1. Introduction  
3.2. Related Work  
3.2.1. Blog Clustering  
3.2.2. Leveraging Tag Information  
3.3. Problem Definition  
3.4. Generating Similarity for Blog Clustering  
3.4.1. Leveraging Collective Wisdom  
3.4.2. Baseline Approach  
3.5. Experiments and Discussion  
3.5.1. Experiment: Design and Methodology  
3.5.2. Results and Analysis  
3.5.2.1. Dynamics of Collective Wisdom  
3.5.2.2. Link Strength  
3.5.2.3. Label Hierarchy  
3.5.2.4. Visualizations - Pajek  
3.5.2.5. k-Means vs. Hierarchical Results  
3.6. Summary  

CHAPTER 4  DISCOVERING FAMILIAR STRANGERS IN BLOGOSPHERE  

4.1. Introduction  
4.2. Problem Formulation  
4.3. Social Identity Theory
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Social Media Sites Grouped Under Categories Based on Their Functionality.</td>
<td>2</td>
</tr>
<tr>
<td>2. Top 20 Most Visited Websites Globally According to Traffic Report Generated by Alexa on June 18th, 2009.</td>
<td>3</td>
</tr>
<tr>
<td>3. Two Lists of the Top 5 Bloggers According to Tuaw and iFinder, Respectively.</td>
<td>25</td>
</tr>
<tr>
<td>4. Comparison of Statistics Between Different Bloggers.</td>
<td>26</td>
</tr>
<tr>
<td>5. Intersection of Digg and Top 20 from iFinder.</td>
<td>27</td>
</tr>
<tr>
<td>6. Distribution of 100 Digg Blog Posts.</td>
<td>27</td>
</tr>
<tr>
<td>7. Distribution of 535 TUAW Blog Posts.</td>
<td>27</td>
</tr>
<tr>
<td>8. Overlap Between Top 20 Blog Posts at Digg and iFinder for Last 6 Months for Different Configurations.</td>
<td>29</td>
</tr>
<tr>
<td>9. Comparison of Statistics Between Influential and Non-Influential Blog Posts.</td>
<td>32</td>
</tr>
<tr>
<td>10. Overlap Between iFinder, Google PageRank, and Digg (Top 20 Blog Posts from Each Model).</td>
<td>33</td>
</tr>
<tr>
<td>11. Link Strength Statistics for ALL Labels.</td>
<td>57</td>
</tr>
<tr>
<td>12. Various Statistics to Compare Clustering Results for Different Threshold Values for WisColl.</td>
<td>59</td>
</tr>
<tr>
<td>13. Various Statistics to Compare Clustering Results for Different Label Structure for WisColl.</td>
<td>62</td>
</tr>
<tr>
<td>14. Baseline Link Strength Statistics.</td>
<td>65</td>
</tr>
<tr>
<td>15. Hierarchical Clustering Table with Clustering Assignment for Link Strength ≥ 5 for All-Label.</td>
<td>68</td>
</tr>
<tr>
<td>16. Comparing WisColl with Baseline approach Using k-Means and Hierarchical Clustering.</td>
<td>72</td>
</tr>
<tr>
<td>17. Summary of BlogCatalog and DBLP Datasets.</td>
<td>92</td>
</tr>
<tr>
<td>18. Clustering Coefficient Results for Both Datasets.</td>
<td>94</td>
</tr>
<tr>
<td>19. Within Similarity and Between Similarity by Different Clustering Methods.</td>
<td>95</td>
</tr>
<tr>
<td>20. Comparison of the Approaches in Terms of Accuracy and Search Space Complexity for BlogCatalog Dataset.</td>
<td>97</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>21. Comparison of the Approaches in Terms of Accuracy and Search Space Complexity for DBLP Dataset.</td>
<td>98</td>
</tr>
<tr>
<td>22. Summary of the Results of the Reaction of Three Different Blogs to the Events.</td>
<td>111</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><em>i-graph</em> Showing the <em>InfluenceFlow</em> Across Blog Post <em>p</em></td>
<td>17</td>
</tr>
<tr>
<td>2.</td>
<td>Log-log Plot of Digg Scores of Blog Posts That Appear on Digg in January 2007</td>
<td>30</td>
</tr>
<tr>
<td>3.</td>
<td>Log-log Plot of Influence Scores of Blog Posts Computed Using iFinder in January 2007</td>
<td>30</td>
</tr>
<tr>
<td>4.</td>
<td>Influential Bloggers’ Blogging Behavior Over the Whole TUAW Blog History.</td>
<td>33</td>
</tr>
<tr>
<td>5.</td>
<td>Evaluating Significance of Each of the Parameters Through Lesion Study.</td>
<td>36</td>
</tr>
<tr>
<td>6.</td>
<td>Pairwise Correlation Plots of the Four Parameters (<em>ι</em>, <em>θ</em>, <em>λ</em>, and <em>γ</em>) of the Blog Posts</td>
<td>37</td>
</tr>
<tr>
<td>7.</td>
<td>Spiky Comments Reaction on a Blog Post Related to iPhone.</td>
<td>38</td>
</tr>
<tr>
<td>8.</td>
<td>“Flat” Comments Reaction on a Blog Post Related to Some Competition in Apple Inc.</td>
<td>39</td>
</tr>
<tr>
<td>9.</td>
<td>An Instance of Label Relation Graph.</td>
<td>49</td>
</tr>
<tr>
<td>10.</td>
<td>Analysis Tree.</td>
<td>52</td>
</tr>
<tr>
<td>11.</td>
<td>Distribution of Blog Sites with Respect to the Labels.</td>
<td>54</td>
</tr>
<tr>
<td>12.</td>
<td>CRGs for Different Datasets Containing 10,642 Bloggers (10k) and 12,308 Bloggers (12k)</td>
<td>56</td>
</tr>
<tr>
<td>13.</td>
<td>All Label Cluster Frequency Count (Y-axis) by Cluster Size (X-axis) per Corresponding Threshold Value.</td>
<td>58</td>
</tr>
<tr>
<td>14.</td>
<td>All Label Cluster Histogram for Small Size Clusters (Size 10 or Less) per Corresponding Threshold Value for Figure 13.</td>
<td>58</td>
</tr>
<tr>
<td>15.</td>
<td>WisColl Results for Link Strength ≥ 3 for All-Label Dataset.</td>
<td>60</td>
</tr>
<tr>
<td>16.</td>
<td>WisColl Results for Link Strength ≥ 5 for All-Label Dataset.</td>
<td>60</td>
</tr>
<tr>
<td>17.</td>
<td>WisColl Results for Link Strength ≥ 7 for All-Label Dataset.</td>
<td>61</td>
</tr>
<tr>
<td>18.</td>
<td>WisColl Results for Link Strength ≥ 3 for Top-Level Label Dataset.</td>
<td>62</td>
</tr>
<tr>
<td>19.</td>
<td>WisColl Results for Link Strength ≥ 3 for Personal Label Dataset.</td>
<td>63</td>
</tr>
<tr>
<td>20.</td>
<td>Baseline Cluster Frequency by Cluster Size per Corresponding Threshold Value.</td>
<td>65</td>
</tr>
<tr>
<td>21.</td>
<td>Baseline Cluster Histogram for Small Size Clusters (Size 10 or Less) per Corresponding Threshold Value for Figure 20.</td>
<td>66</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>22. Results for Link Strength ≥ 0.80 for Baseline Dataset.</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>23. Hierarchical Clustering for Link Strength ≥ 5 for All-Label Dataset Value for Indexes per Table 15.</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>24. k-Means k-Analysis for Baseline Dataset.</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>25. WisColl Results for Link Strength ≥ 3 for Top-Level Label Dataset.</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>26. WisColl Results for Link Strength ≥ 5 for Top-Level Label Dataset.</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>27. WisColl Results for Link Strength ≥ 7 for Top-Level Label Dataset.</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>28. WisColl Results for Link Strength ≥ 9 for Top-Level Label Dataset.</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>29. WisColl Results for Link Strength ≥ 3 for Personal Label Dataset.</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>30. WisColl Results for Link Strength ≥ 5 for Personal Label Dataset.</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>31. WisColl Results for Link Strength ≥ 7 for Personal Label Dataset.</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>32. WisColl Results for Link Strength ≥ 9 for Personal Label Dataset.</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>33. Results for Link Strength ≥ 0.80 for Baseline Dataset.</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>34. Results for Link Strength ≥ 0.85 for Baseline Dataset.</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>35. Results for Link Strength ≥ 0.88 for Baseline Dataset.</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>36. Results for Link Strength ≥ 0.90 for Baseline Dataset.</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>37. Results for Link Strength ≥ 0.92 for Baseline Dataset.</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>38. Results for Link Strength ≥ 0.95 for Baseline Dataset.</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>39. Searching Familiar Strangers for a Node u Given the Local Network Information That u Has and the Goal γ.</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40. Log-log Plot of Degree Distribution for BlogCatalog.</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>41. Log-log Plot of Degree Distribution for DBLP.</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>42. Differential of the Ratio of Within Similarity and Between Similarity vs. k.</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>43. Accuracy vs. Search Steps for BlogCatalog Dataset.</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>44. Accuracy vs. Log of Search Steps for BlogCatalog Dataset.</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>45.</td>
<td>Accuracy vs. Search Steps for DBLP Dataset</td>
<td>100</td>
</tr>
<tr>
<td>46.</td>
<td>Accuracy vs. Log of Search Steps for DBLP Dataset</td>
<td>100</td>
</tr>
<tr>
<td>47.</td>
<td>Selectivity vs. Accuracy for BlogCatalog Dataset</td>
<td>101</td>
</tr>
<tr>
<td>48.</td>
<td>Selectivity vs. Accuracy for DBLP Dataset</td>
<td>101</td>
</tr>
<tr>
<td>49.</td>
<td>Selectivity vs. Search Steps for BlogCatalog Dataset</td>
<td>102</td>
</tr>
<tr>
<td>50.</td>
<td>Selectivity vs. Search Steps for DBLP Dataset</td>
<td>102</td>
</tr>
<tr>
<td>51.</td>
<td>Accuracy vs. Search Steps for BlogCatalog for Random Search for Varying $\sigma$</td>
<td>103</td>
</tr>
<tr>
<td>52.</td>
<td>Accuracy vs. Search Steps for DBLP for Random Search for Varying $\sigma$</td>
<td>103</td>
</tr>
<tr>
<td>53.</td>
<td>Types of Reactions of Community Blogs to an Event</td>
<td>107</td>
</tr>
<tr>
<td>54.</td>
<td>Flowchart of the Various Components of the Proposed Approach</td>
<td>108</td>
</tr>
<tr>
<td>55.</td>
<td>Blog Reactions to Saddam Hussein’s Verdict</td>
<td>109</td>
</tr>
</tbody>
</table>
1. INTRODUCTION TO SOCIAL COMPUTING

Social computing is a multi-disciplinary research program that focuses on human, cultural, and behavioral aspects. It brings together experts from various disciplines like: anthropology, cognitive science, computer science, economics, linguistics, mathematics, neuroscience, political science, psychology, sociology, statistics, and theology. Social computing refers to the intersection of social behavior and computational systems. Social computing is often defined as modeling complex human interactions that are expressed on a variety of social media. Social media, or commonly known as the Social Web, consists of an ant-colony of services including blogs, media sharing, micro blogging, social bookmarking, social news, social friendship networking websites, and wikis. Different social media sites could be alike or different in terms of functionality. We briefly describe each category and the functionalities:

- **Blogs**, or *web logs*, is a collection of articles written by people arranged in reverse chronological order. These articles are known as *blog posts*. The collection of all the blogs is referred to as *Blogosphere*. Blogs allow people to share their views, express their opinions, interact and discuss with each other through linking to other blogs or posting comments. A blog can be maintained by an individual known as an *individual blog* or by a group of people known as a *community blog*. The authors of blogs are known as *bloggers*. Some blogs such as BlogCatalog (http://www.blogcatalog.com/) also allow users to create their friendship networks.

- **Media Sharing** sites allow people to upload and share their multimedia content on the web, including, images, videos, audio, etc. with other people. People can watch the content shared by others, enrich them with tags, and share their thoughts through comments. Some media sharing sites allow users to create friendship networks.

- **Micro Blogging** sites, as the name suggests, are similar to blogs except the fact that the articles can only be of certain length. In case of Twitter (http://www.twitter.com/), the articles can be 140 characters in length. These articles are also called messages (or tweets in the case of Twitter) because of the short length. These sites are typically used to share what you are doing. Besides posting messages people can also create friendship networks. They can follow or become followers of other users.
TABLE 1
Social Media Sites Grouped Under Categories Based on Their Functionality.

<table>
<thead>
<tr>
<th>Category</th>
<th>Social Media Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blogs</td>
<td>Wordpress, Blogger, Blogcatalog, MyBlogLog</td>
</tr>
<tr>
<td>Media Sharing</td>
<td>Flickr, Photobucket, YouTube, Multiply, Justin.tv, Ustream</td>
</tr>
<tr>
<td>Micro Blogging</td>
<td>Twitter, SixApart</td>
</tr>
<tr>
<td>Social Bookmarking</td>
<td>Del.icio.us, StumbleUpon</td>
</tr>
<tr>
<td>Social Friendship Network</td>
<td>MySpace, Facebook, Friendfeed, Bebo, Orkut, LinkedIn, PatientsLikeMe, DailyStrength</td>
</tr>
<tr>
<td>Social News</td>
<td>Digg, Reddit</td>
</tr>
<tr>
<td>Wikis</td>
<td>Wikipedia, Wikiversity, Scholarpedia, Ganfyd, AskDrWiki</td>
</tr>
</tbody>
</table>

- **Social Bookmarking** sites allow people to tag their favorite webpages or websites and share it with the other users. This generates a good amount of metadata for the webpages. People can search through this metadata to find relevant or most favorite webpages/websites. People can also see the most popular tags or the most freshly used tags and freshly favored website/webpage. Some social bookmarking sites like StumbleUpon (http://www.stumbleupon.com/) allow people to create friendship networks.

- **Social Friendship Networks** allow people to stay in touch with their friends and also create new friends. Individuals create their profile on these sites based on their interests, location, education, work, etc. Usually the ties are non-directional, which means that there is a need to reciprocate the friendship relation between two nodes.

- **Social News** sites allow people to share news with others and let others vote on these stories. News that are voted the most emerge as the most popular news stories. People can tag various news stories. They can get the most popular stories, fastest upcoming stories for different time periods, and share their thoughts by providing comment.
• Wikis are publicly edited encyclopedias. Anyone can contribute articles to wikis or edit existing ones.

However, most of the wikis are moderated to protect them from vandalism. Wikis provide a great technology for content management, where people with a very basic knowledge of formatting contribute and produce rich sources of information. Wikis also maintain the history of changes and have the capability to rollback to any previous version. Popular wiki like Wikipedia (http://www.wikipedia.org/) also allow people to classify the articles under one of the following categories: Featured, Good, Cleanup, and Stub.

Table 1 presents a categorization of various social media sites in terms of functionality.

The popularity of these social media sites can be gauged by looking at the traffic report generated by Alexa\(^1\). The list of top-20 most visited websites globally according to Alexa generated on June 18th, 2009 is shown in Table 2. According to this report, 35% of 20 most visited websites are social media sites (denoted in bold font in Table 2). This clearly shows the high amount of traffic that is driven towards the social media sites which reflects the popularity of these sites among the masses.

Next we will look at the characteristics of social media that made it gain popularity among the masses rapidly in a short time period as compared to the industrial or traditional media. Some of these characteristics are:

\(^1\)http://www.alexa.com/
1. **Accessibility** - Social media sites are publicly available for almost free or at no cost. Whereas, industrial media is usually privately owned or by government and is not freely available to people.

2. **Permanence** - Social media sites can be altered anytime. Individuals can edit their blogs, profile, preferences, etc. anytime they wish by providing comments. Whereas, industrial media cannot be altered once created, e.g., a magazine article that is published cannot be altered instantaneously by providing comments.

3. **Reach** - Like industrial media, social media sites also has a global audience.

4. **Recency** - The time lag between communications produced by social media sites can be almost zero. The communication on social media sites can be instantaneous. Whereas, the communications on industrial media can take days, weeks or even months.

5. **Usability** - Most social media sites do not require any special skills to create content. Social media sites offer technologies with almost zero operational cost. Whereas, industrial media requires specialized skills and training.

With these characteristics social media is an extremely fledgling domain with people not only generating content but also enriching it by providing metadata like tags, labels, categories, etc. Some social media sites also provide the capability to be contextually mashed up with some other sites, generating user developed widgets. For instance, people can combine Flickr’s images with Google’s maps to display images on a map creating an image map. Such mashups or widgets aim to combine disparate data sources under a single context and add more meaning to the data. This collaborative environment gives rise to a phenomenon referred to as the *participatory web or citizen journalism*. The above characteristics of social media sites lead to various challenges like:

- **Size** - Through reactive interfaces, low barrier to publication, and zero operational costs, which are all made possible by the new paradigm of Web 2.0, social media has observed a phenomenal growth in user participation. Blogosphere has consistently doubled every five months for the last 4 years. Its growth is
reported by Technorati (http://www.technorati.com/) in a report\(^2\) which mentions approximately 18.6 new blog posts are created every second. Technorati has tracked 133 million blogs till December 2008. Other social media sites like Facebook have approximately 200 million active users as recorded in May 2009. Similarly, Twitter, another micro blogging site has observed a phenomenal growth of 95% in one month\(^3\) by amassing nearly 19.1 million users in March 2009. With this rate it is expected to accrue 50 million users by summers 2009. Other social media sites like Digg, Del.icio.us, Stumbleupon, Flickr, YouTube, etc. are also growing at terrific pace. With such a rapid pace of content generation, it gets really difficult to follow what is currently happening in social media. The information quickly overwhelms the individuals. Search engines are often faced with the dilemma of choosing freshness of results over accuracy. To handle this issue in the blogosphere, we identify influential bloggers who stand out as representatives and can be followed to glean the insight of the current affairs in the blogosphere. These influential bloggers are the sources of good content. The work [1] is covered in more details in Chapter 2.

- **Dynamism** - As mentioned earlier, social media sites encourage instantaneous response with almost zero time lag in communication, the environment is highly dynamic. It can be observed from the blogosphere that people have varied interests and their interest in one topic is short-lived [2, 3]. This causes a drift not only in people’s interests but also as a whole in the blogosphere. However, people tend to categorize their blog posts using the same category descriptors they used to categorize their old blog posts, a phenomenon referred to as *path dependence* [4]. This is because either they are ignorant of the category structure (also because the taxonomy structure is highly dynamic and keeps evolving), or they are lazy to submit their blogs to more focused or refined categories. To handle this issue in the blogosphere, we tap the collective wisdom of the people and propose a clustering algorithm that reflects the dynamics of social media sites. The work [5] is covered in more details in Chapter 3.

- **Search in Long Tail** - Social media sites have started a surge of open-source intelligence. Since

\(^2\)http://technorati.com/blogging/state-of-the-blogosphere/

more and more people are participating in Web 2.0 activities, it has generated enormous amounts of intelligently crafted content. Web 2.0 has allowed the mass not only to contribute and edit posts/articles through blogs and wikis, but also enrich the existing content by providing tags or labels, hence turning the former information consumers to the new producers. Allowing the mass to contribute or edit has also increased collaboration among the people unlike Web 1.0 where the access to the content was limited to a chosen few. Blogs are invigorating this process by encouraging the mass to document their ideas, thoughts, opinions, and views in the form of blog posts, and share them with other bloggers. This largely results in a power law distribution of the number of blogs vs. their popularity, meaning, only a very few blogs are extremely popular or known (the Short Head) and a vast number of blogs are largely unknown (the Long Tail) [6]. The popularity could be gauged through numerous parameters such as connections or links, readership, etc. In particular, many bloggers are active locally with limited connections to other bloggers. The bloggers in the long tail have niche interests and present exciting business opportunities. Here is the dilemma: Before a blogger becomes prominent or in the Short Head, it is not worth paying particularly customized attention to the blogger; and the blogger cannot be well targeted for otherwise potential business opportunities (i.e., niches). To handle this issue, we propose a social identity based search approach that identifies bloggers in the long tail who are similar and disconnected (or, familiar strangers [7]); and aggregates them for increased visibility and better personalized services (such as customization and recommendation). The work [8,9] is covered in more details in Chapter 4.

- **Sparse Link Structure** - Due to the casual environment of social media especially Blogosphere, people usually skip to link the source they were inspired from to write their blog post [10]. This creates an extremely sparse link structure of the blog network. This presents challenges in understanding social interactions evolving in online communities especially in the blogosphere through link analysis. Understanding social interactions would lead to better understanding of the socio-cultural ties between these communities to foster collaboration, better personalization, predictive modeling, and enable tracking.

---

4A detailed analysis is presented in Chapter 2.
and monitoring. To handle this issue in the blogosphere, we study an event based community interaction approach [11] and perform an empirical analysis on a controlled environment presented in Chapter 5. We plan to explore it further as a future direction.

- **Information Quality** - As a consequence of ease of use and low barrier to publication social media suffers from the challenge of information quality. Often due to the casual nature of the blogosphere, bloggers use colloquial forms of language. Apparently features that seem noisy might be informative. Such a casual environment nurtures sentiments, expressions, and emotions through writing; it is much more prevalent to observe intentionally modified spellings such as “what’s uppppp?” and “this is so cooooool..”. These instances demonstrate examples of intonation\(^5\) in written texts. These examples through misspellings clearly emphasize stress on the emotions and convey more information than the regular text. It would be undesirable to consider them as sheer misspellings and replace them with the correct spellings. Services like UrbanDictionary (http://www.urbandictionary.com) can be used to unravel the informative content in the slangs, abbreviations, and/or colloquial forms of language used by the bloggers. Besides colloquial usage, intentional misspellings, and slang text, there is a lot of off-topic chatter or noise that could distort the analysis. It has a tremendous potential and presents a great promise for further exploration.

- **Evaluation of Algorithms/Models** - In order to measure the difference an algorithm makes over the existing counterparts, it is necessary to systematically evaluate the merit, worthiness, and significance of the algorithm. This process is called evaluation. The algorithms or the proposed models are evaluated using a set of standard criteria. These standard criteria are often used in various domains. However, sometimes the standard criteria could not be used in evaluating algorithms or techniques in the context of the blogosphere. For instance, evaluating concepts like influence presents a big challenge due to the absence of ground truth. Evaluation models based on training and testing data fail in such situations. Performing human evaluations through surveys looks like the only solution for this challenge.

---

\(^5\)Intonation is a linguistic concept that refers to the different meanings conveyed by the different ways of pronunciation of a word [12]. The listener could interpret different meanings based on the prosodic utterance. Intonation is as common in written texts as it is in spoken language.
However, human evaluations present bigger challenges such as funding and recruiting unbiased and representative users. High costs and long time are another constraints. To handle this issue in the blogosphere, we discuss some novel and avant-garde evaluation strategies for evaluating concepts like influence, where conventional evaluation criteria fall short in Chapter 2 and Chapter 4.

Data is an essential asset for performing any sort of analysis. Data collection is equally important and challenging. Moreover, since research in the blogosphere is a relatively new domain, there are not many benchmark datasets available and especially the ones available are not rigorous to support the work presented here. These datasets miss out important elements, such as, link information, comment information, social network information among bloggers. Often these datasets are multi-lingual which presents a whole different set of challenges which are out of the scope of this work. In this work, we crawled data from different blogs such as The Unofficial Apple Weblog (TUAW)\(^6\), and BlogCatalog\(^7\). These datasets are available for download at Social Computing group’s webpage (http://socialcomputing.asu.edu/) at Arizona State University. We also tested our model on the available benchmark datasets such as DBLP\(^8\). Different research problems involve different challenges and issues, to be tackled which necessitates the use of datasets with different characteristics. For example, in Chapter 2, we present an analysis of influential bloggers which is possible only at a multi-authored blog. So we used TUAW which is a multi-authored blog, instead of BlogCatalog which is a portal of single-authored blogs. However, in Chapter 4, we used BlogCatalog since it provides social network among bloggers required to analyze the approaches presented in this paper. This is not achievable in TUAW dataset.

\(^{6}\)http://www.tuaw.com/
\(^{7}\)http://www.blogcatalog.com/
\(^{8}\)http://kdl.cs.umass.edu/data/dblp/dblp-info.html
2. IDENTIFYING INFLUENTIAL BLOGGERS

2.1. Introduction

The advent of participatory Web applications (or Web 2.0 [13]) has created online media that turn the former mass information consumers to the present information producers [14]. Examples include blogs, wikis, social annotation and tagging, media sharing, and other such services. A blog site or simply blog (short for web log) is a collection of entries by individuals displayed in reverse chronological order. These entries, known as the blog posts, can typically combine text, images, and links to other blogs, blog posts, and/or to Web pages. Blogging is becoming a popular means for mass Web users to express, communicate, share, collaborate, debate, and reflect. Blogosphere is the virtual universe that contains all blogs. Bloggers, the blog writers, loosely form their special interest communities where they share thoughts, express opinions, debate ideas, and offer suggestions interactively. Blogosphere provides a conducive platform to build the virtual communities of special interests. It reshapes business models [15], inspires viral marketing [16], provides trend analysis and sales prediction [17,18], aids counter-terrorism efforts [19] and acts as grassroots information sources [20].

In a real-world world, according to [21], 83% of people prefer consulting family, friends or an expert over traditional advertising before trying a new restaurant, 71% of people do the same before buying a prescription drug or visiting a place, and 61% of people talk to family, friends or an expert before watching a movie. In short, before people buy or make decisions, they talk, and they listen to other’s experience, opinions, and suggestions. The latter affect the former in their decision making, and are aptly termed as the influentials [21]. Influence has always been an unabated interest in business and society. With the pervasive presence and ease of use of the Web, an increasing number of people with different backgrounds flock to the Web - a virtual world to conduct many previously inconceivable activities from shopping, to making friends, and to publishing. As we draw parallels between real-world and virtual communities, among citizens of the blogosphere, we are intrigued by the questions like whether there exist the influentials in a virtual community (a blog), who they are, and how to find them.

Blogs can be categorized into two major types: individual and community blogs. Individual blogs are single-authored who record their thoughts, express their opinions, and offer suggestions or ideas. Others can comment on a blog post, but cannot start a new line of blog posts. These are more like diary entries or
personal experiences. Examples of individual blogs are Sifry’s Alerts: David Sifry’s musings\(^1\) (Founder & CEO, Technorati), Ratcliffe Blog–Mitch’s Open Notebook\(^2\), The Webquarters\(^3\), etc. A community blog is where each blogger can not only comment on some blog posts, but also start some topic lines. Examples of community blog sites are Google’s Official Blog\(^4\), The Unofficial Apple Weblog\(^5\), Boing Boing: A Directory of Wonderful Things\(^6\) etc. For an individual blog, the host is the only one who initiates and leads the discussions and thus is naturally the influential blogger of his/her site. For a community blog where many have equal opportunities to participate, we study who are the influentials in a virtual community. Henceforth, blogs refer to community blogs.

2.1.1. Applications of the Influentials

Since the bloggers can be connected in a virtual community anywhere anytime, the identification of the influential bloggers can benefit all in developing innovative business opportunities, forging political agendas, discussing social and societal issues, and lead to many interesting applications. For example, the influentials are often *market-movers*. Since they can influence buying decisions of the fellow bloggers, identifying them can help companies better understand the key concerns and new trends about products interesting to them, and smartly affect them with additional information and consultation to turn them into unofficial spokesmen. As reported in [22], approximately 64% advertising companies have acknowledged this phenomenon and are shifting their focus toward blog advertising.

The influentials could also *sway* opinions in political campaigns, elections, and affect reactions to government policies [23]. Tapping on the influentials can help understand the changing interests, foresee potential pitfalls and likely gains, and adapt plans timely and pro-actively (not just reactively). The influentials can also help in customer support and troubleshooting since their solutions are trustworthy because of the sense of authority these influentials possess. For example, Macromedia\(^7\) aggregates, categorizes and searches the

\(^{1}\)http://www.sifry.com/alerts/
\(^{2}\)http://www.ratcliffeblog.com/
\(^{3}\)http://webquarters.blogspot.com/
\(^{4}\)http://googleblog.blogspot.com/
\(^{5}\)http://www.tuaw.com/
\(^{6}\)http://boingboing.net/
\(^{7}\)http://weblogs.macromedia.com/
blog posts of 500 people who write about Macromedia’s technology. Instead of going through every blog post, an excellent entry point is to start with the influentials’ posts.

Some recent numbers from Technorati\(^8\) show a 100% increase in the size of Blogosphere every six months, “..., about 1.6 Million postings per day, or about 18.6 posts per second”\(^9\). Blogosphere has grown over 60 times during the past three years. With such a phenomenal growth, novel ways have to be developed in order to keep track of the developments in the blogosphere.

Researchers have studied the influence in the blogosphere from the perspective of influential blog sites (more in Section 2.5). Regardless of a blog being influential or not, it can have its influential bloggers. Influential bloggers of a blog have impact on the fellow bloggers as in a real-world community. In this work, we address the novel problem of identifying influential bloggers in a blog and investigate its issues and challenges.

- Are there influential bloggers as in a real-world community? Are they simply active bloggers?

- What measures should be used to define influential bloggers? A solution can be subjective, depending on the need for identifying influential bloggers.

- How to find influential bloggers? As there is no training data to tell us who are influential bloggers or not, it is infeasible to apply classification. Combining the statistics collected for each blogger, can we create a robust model that quantitatively tells how influential a blogger is?

- Can we tune/adjust the model to identify different classes of influential bloggers to satisfy various needs?

Specifically, we make the following contributions:

- Identified the collectable statistics in the Blogosphere that is used to quantify a blogger’s influence.

- Defined and formulated the influence of a blogger in terms of the collectable statistics.

\(^8\)http://technorati.com/
\(^9\)http://www.sifry.com/alerts/archives/000436.html
• Proposed an algorithm - iFinder [1] - that computes the influence score of each blogger.

• Evaluated the proposed algorithm to identify various categories of influential bloggers, their temporal patterns, relative importance of collectable statistics and other interesting observations.

• Designed a novel evaluation framework to validate the model in absence of the ground truth.

In the following chapter, we first review the literature and differentiate this work from the existing ones. In Section 2.2, we study the statistics collectable from a blog site, and define the problem of identifying influential bloggers. In Section 2.3, we propose a preliminary model that allows for evaluating different key measures for identifying the influentials and can be adapted to look for different types of influential bloggers. In Section 4.7, we conduct an empirical study to evaluate many aspects of the proposed approach and its effectiveness, and observe how the key measures work with a correlation study. Section 2.5 reviews the existing work in this domain. Finally we summarize the chapter with some possible future directions in Section 4.9.

2.2. Influential Bloggers: Problem and Definition

Each blog post is often associated with some metadata like post’s author, post annotations, post’s date and time, number of comments. In addition, one can also collect certain statistics from the blog website for example, outlinks - posts or articles to which the author has referred; inlinks - other posts that refer to this post, post length; average length of comments per post; and the rate at which comments are posted on a blog post. Since a long blog post can simply contain many outlinks, outlinks are normalized by the length of the blog post. Inlinks are collected using Technorati API\(^{10}\).

In the simplest case, one can approximate an influential blogger with an active blogger who posts frequently. Since this is not the case in a physical world where a voluble person is not necessarily or seldom influential, we are inquisitive whether we can employ the above metadata and statistics to identify influential bloggers. The search for influential bloggers boils down to the question on how to define an influential blogger. First, active bloggers are not necessarily influential and influential bloggers can be inactive. Hence,

\(^{10}\)http://technorati.com/developers/api/cosmos.html
we categorically divide bloggers into four types: active and influential, active and non-influential, inactive and influential, and inactive and non-influential. Second, while active bloggers can be simply defined by how frequently a blogger posts, it is a more complex matter how to define an influential blogger with the aid of the above mentioned statistics.

Recognizing the subjective nature of defining an influence blogger, we propose a preliminary model to quantify the properties of the influential bloggers by combining various statistics collectable from a blog site and assigning influence scores to each blogger and their blog posts. Next, we investigate how these statistics can be used in various ways to adjust the model for different purposes. In this work, we first develop an intuitive model that goes beyond the post frequency and allows the use of the combination of statistics. Then we demonstrate how to use this model to identify influential bloggers who may or may not be active, and further investigate how to further refine and evolve the preliminary model in finding various types of influential bloggers.

An intuitive way of defining an influential blogger is to check if the blogger has any influential blog post, i.e., A blogger can be influential if s/he has more than one influential blog post. Assume we have an influence score\(^{11}\) for a post \(p_i\), \(I(p_i)\). For a blogger \(b_k\) who has \(N\) blog posts, \(\{p_1, p_2, ..., p_N\}\), their influence scores can be ranked in descending order, and her influence index, \(iIndex(b_k)\) can be defined as \(\max(I(p_i))\), where \(1 \leq i \leq N\). Given a set \(U\) of \(M\) bloggers, \(\{b_1, b_2, ..., b_M\}\), the problem of identifying influential bloggers is defined as determining an ordered subset \(V\) of \(K^{12}\) bloggers, \(\{b_{j_1}, b_{j_2}, ..., b_{j_K}\}\) that are ordered according to their \(iIndex\) such that \(V \subseteq U\) and \(K \leq M\), i.e. \(iIndex(b_{j_1}) \geq iIndex(b_{j_2}) \geq ... \geq iIndex(b_{j_K})\). \(V\) contains \(K\) most influential bloggers. For all the blog posts \(\{p_1, p_2, ..., p_L\}\) by all \(M\) bloggers, influential blog posts are those whose influence scores are greater than \(iIndex(b_{j_K})\) or, \(I(p_l) \geq iIndex(b_{j_K})\) for \(1 \leq l \leq L\). Hence, we have the following corollary: those bloggers who published blog posts that satisfy \(I(p_l) \geq iIndex(b_{j_K})\), for \(1 \leq l \leq L\) will be called influential bloggers because their \(iIndex\) will be greater than or equal to \(iIndex(b_{j_K})\).

\(^{11}\)These concepts are defined mathematically in Section 2.3.3.

\(^{12}\)Note that \(K\) is a user specified parameter.
We now study the intuitive characteristics that help define iIndex and I so as to build an experimental model that can gauge the influence to distinguish between “influential” and “activeness” properties of bloggers.

2.3. Identifying the Influentials

We first present some desirable properties related to blog-post influence which can be approximately defined by collectable statistics, next propose a preliminary model of identifying the influentials using these statistics, then discuss some interesting issues that can be evaluated by experimenting the preliminary model.

2.3.1. An Initial Set of Intuitive Properties

Following [21], one is influential if s/he is recognized by fellow citizens, can generate follow-up activities, has novel perspectives or ideas, and is often eloquent. Below we examine how these social gestures that forms the characteristic properties of an influential can be approximated by some collectable statistics.

- **Recognition** - An influential blog post is recognized by many. This can be equated to the case that an influential post $p$ is referenced in many other posts. The influence of those posts that refer to $p$ can have different impact: the more influential the referring posts are, the more influential the referred post becomes. Recognition of a blog post is measured through the inlinks ($i$) to the blog post.

- **Activity Generation** - A blog post’s capability of generating activity can be indirectly measured by how many comments it receives, the amount of discussion it initiates. In other words, few or no comment suggests little interest of fellow bloggers, thus non-influential. Hence, a large number of comments ($\gamma$) indicates that the post affects many such that they care to write comments, and therefore, the post can be influential. There are increasing concerns over spam comments that do not add any value to the blog posts or blogger’s influence. Fighting spam is outside the scope of this work and recent research can be found in [24, 25].

- **Novelty** - Novel ideas exert more influence as suggested in [21]. Hence, the outlinks ($\theta$) is an indicator of a post’s novelty. If a post refers to many other blog posts or articles it indicates that it is less likely to be novel. The number of outlinks is negatively correlated with the number of comments which means more outlinks reduces people’s attention. This is confirmed later in Section 5.2.5.
• Eloquence - An influential is often eloquent [21]. This property is most difficult to approximate using some statistics. Given the informal nature of the blogosphere, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so. Therefore, we use the length of a post ($\lambda$) as a heuristic measure for checking if a post is influential or not. The blog post length is positively correlated with number of comments which means longer blog posts attract people’s attention. This is confirmed later in Section 5.2.5. Eloquence of a blog post could be gauged using more sophisticated linguistic based measures which are intended to be pursued as a future work.

The above four form an initial set of properties possessed by an influential post. There are certainly some other potential properties. It is also evident that each of the above four may not be sufficient on its own, and they should be used jointly in identifying influential bloggers. For example, a high $\theta$ and a poor $\lambda$ could identify a “hub” blog post. Starting with this initial set, we next build a preliminary model that allows us to examine, analyze, modify, and extend the model.

2.3.2. Developing the Model

First we study a model that only uses links to rank the bloggers and then improve on it to include other statistics. For this purpose we consider PageRank [26] algorithm. Akin to webpages, PageRank assigns numerical weighting for each blog post to “measure” its relative importance. The PageRank score of a blog post ($p_i$) is a probability ($PR(p_i)$) that represents the likelihood of a random surfer clicking on links will arrive on this blog post and is represented as:

$$PR(p_i) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

(2.1)

where $d$ is the damping factor that the random surfer stops clicking at some time, $M(p_i)$ is the set of all the blog posts that link to $p_i$, $L(p_j)$ is the total number of outbound links on blog post $p_j$, and $N$ is the total number of blog posts. The PageRank values $\mathbf{R}$ could be computed as the entries of the dominant eigenvector
of the modified adjacency matrix,

\[
R = \begin{pmatrix}
\frac{1-d}{N} & l(p_1, p_1) & l(p_1, p_2) & \cdots & l(p_1, p_N) \\
\frac{1-d}{N} & l(p_2, p_1) & \ddots & \vdots & \vdots \\
\vdots & \vdots & \ddots & l(p_i, p_j) \\
\frac{1-d}{N} & l(p_N, p_1) & \cdots & l(p_N, p_N)
\end{pmatrix}
\]

where the function \( l(p_i, p_j) \) is 1 if blog post \( p_j \) links to blog post \( p_i \), and 0 otherwise.

As pointed out in [10], blog sites in the blogosphere are very sparsely linked and it is not suitable to rank blog sites using Web ranking algorithms. The Random Surfer model of webpage ranking algorithms [26] does not work well for sparsely linked structures. The temporal aspect is most significant in blog domain. While a webpage may acquire authority over time (its adjacency matrix gets denser), the older a blog post gets the lesser people care about it and hence its influence diminishes over time. This is due to the fact that the adjacency matrix of blogs (considered as a graph) will get sparser as thousands of new sparsely-linked blog posts appear every day. Next we propose a model - iFinder - that uses all the above-mentioned statistics.

We perform experiments to compare iFinder and PageRank algorithm and report our findings in Section 4.7.

2.3.3. iFinder - a Preliminary Model

Blog-post influence can be visualized in terms of an influence graph or \( i \)-graph in which the influence of a blog post flows among the nodes. Each node of an \( i \)-graph represents a single blog post characterized by the four properties (or parameters): \( \iota, \theta, \gamma \) and \( \lambda \). \( i \)-graph is a directed graph with \( \iota \) and \( \theta \) representing the incoming and outgoing influence flows of a node, respectively. Hence, if \( I \) denotes the influence of a node (or blog post \( p \)), then \( \text{InfluenceFlow} \) across that node is given by,

\[
\text{InfluenceFlow}(p) = w_{in} \sum_{m=1}^{\mid \iota \mid} I(p_m) - w_{out} \sum_{n=1}^{\mid \theta \mid} I(p_n)
\]  

(2.2)

where \( w_{in} \) and \( w_{out} \) are the weights that can be used to adjust the contribution of incoming and outgoing influence, respectively. \( p_m \) denotes all the blog posts that link to the blog post \( p \), where \( 1 \leq m \leq \mid \iota \mid \); and \( p_n \) denotes all the blog posts that are referred by the blog post \( p \), where \( 1 \leq n \leq \mid \theta \mid \). \( \mid \iota \mid \) and \( \mid \theta \mid \) are the total numbers of inlinks and outlinks of post \( p \). \( \text{InfluenceFlow} \) measures the difference between the
total incoming influence of all inlinks and the total outgoing influence by all outlinks of the blog post \( p \). InfluenceFlow accounts for the part of influence of a blog post that depends upon inlinks and outlinks. From Eq. 2.2, it is clear that the more inlinks a blog post acquires the more recognized it is, hence the more influential it gets; and an excessive number of outlinks jeopardizes the novelty of a blog post which affects its influence. We illustrate the concept of InfluenceFlow in the i-graph displayed in Figure 1. This shows an instance of the i-graph with a single blog post. Here we are measuring the InfluenceFlow across blog post \( p \). Towards the right of \( p \) are the inlinks and outlinks are towards the left of \( p \). We add up the influence “coming into” \( p \) and add up the influence “going out” of \( p \) and take the difference of these two quantities to get the influence that \( p \) has generated.

As discussed earlier, the influence \((I)\) of a blog post is also proportional to the number of comments \((\gamma_p)\) posted on that blog post. We can define the influence of a blog post, \( p \) as,

\[
I(p) \propto w_{com}\gamma_p + \text{InfluenceFlow}(p)
\]  

(2.3)

where \( w_{com} \) denotes the weight that can be used to regulate the contribution of the number of comments \((\gamma_p)\) towards the influence of the blog post \( p \). We consider an additive model because additive function is good to determine the combined value of each alternative [27]. It also supports preferential independence of all the parameters involved in the final decision. Since most decision problems like the one at hand are
multi-objective, a way to evaluate trade-offs between the objectives is needed. A weighted additive function can be used for this purpose [28].

From the discussion in Section 2.3.1, we consider blog post quality as one of the parameters that may affect influence of the blog post. Although there are many measures that quantify the goodness of a blog post such as fluency, rhetoric skills, vocabulary usage, and blog content analysis\(^{13}\), for the sake of simplicity, we here use the length of the blog post as a heuristic measure of the goodness of a blog post in the context of blogging. We define a weight function, \(w\), which rewards or penalizes the influence score of a blog post depending on the length \(\lambda\) of the post. The weight function could be replaced with appropriate content and literary analysis tools. Combining Eq. 2.2 and Eq. 2.3, the influence of a blog post, \(p\), can thus be defined as,

\[
I(p) = w(\lambda) \times (w_{com}\gamma_p + \text{InfluenceFlow}(p))
\]  

(2.4)

The above equation gives an influence score to each blog post. Note that the four weights can take more complex forms and can be tuned. We will evaluate and discuss their effects further in the empirical study.

Now we consider how to use \(I\) to determine whether a blogger is influential or not. According to the definition of influential blogger in Section 2.2, a blogger can be considered influential if s/he has at least one influential blog post. We use the blog post with maximum influence score as the representative\(^{14}\) and assign its influence score as the blogger influence index or \(iIndex\). For a blogger \(B\), we can calculate the influence score for each of \(B\)’s \(N\) posts and use the maximum influence score as the blogger’s \(iIndex\), or

\[
iIndex(B) = \max(I(p_i))
\]  

(2.5)

where \(1 \leq i \leq N\). With \(iIndex\), we can rank bloggers on a blog site. The top \(k\) among the total bloggers are the most influential ones. Thresholding is another way to find influential bloggers whose \(iIndices\) are greater than a threshold. However, determining a proper threshold is crucial to the success of such a strategy and requires more research. Blog posts whose influence score is higher than the influence score of the top-\(k^{th}\) influential blogger could be termed as influential blog posts.

\(^{13}\)A reason we did not adopt any of these is their computation is beyond the scope of this work. We use some simpler measure to examine its effect in determining influence.

\(^{14}\)There could be other ways. For example, if one wants to differentiate a productive influential blogger from non-prolific one, one might use another measure.
2.3.4. Computing Blogger Influence with Matrix Operations

We have described the iFinder model and how to compute the influence of a blog post using the social gestures. Here we convert the computational procedure into basic matrix operations for convenient and efficient implementation.

We define the inlinks and outlinks to the blog posts using a link adjacency matrix $A$ where the entry $A_{ij}$ is 1 if $p_i$ links to $p_j$ and 0 otherwise, defined as

$$
A_{ij} = \begin{cases} 
1 & p_i \rightarrow p_j \\
0 & p_i \not\rightarrow p_j 
\end{cases}
$$

Matrix $A$ denotes the outlinks between the blog posts. Consequently, $A^T$ denotes the inlinks between the blog posts. We define the vectors for blog post length, comments, influence, and influence flow as,

$$
\overrightarrow{\lambda} = (w(\lambda_{p_1}), ..., w(\lambda_{p_N}))^T,
$$

$$
\overrightarrow{\gamma} = (\gamma_{p_1}, ..., \gamma_{p_N})^T,
$$

$$
\overrightarrow{i} = (I(p_1), ..., I(p_N))^T,
$$

$$
\overrightarrow{f} = (f(p_1), ..., f(p_N))^T
$$

respectively.

Now, Eq. 2.2 can be rewritten in terms of the above vectors as,

$$
\overrightarrow{f} = w_{in} A^T \overrightarrow{i} - w_{out} A \overrightarrow{i} = (w_{in} A^T - w_{out} A) \overrightarrow{i} \tag{2.6}
$$

and Eq. 2.4 can be rewritten as,

$$
\overrightarrow{i} = \overrightarrow{\lambda} (w_c \overrightarrow{\gamma} + \overrightarrow{f}) \tag{2.7}
$$

Eq. 2.7 can be rewritten using Eq. 2.6 which can then be solved iteratively,

$$
\overrightarrow{i} = \overrightarrow{\lambda} (w_c \overrightarrow{\gamma} + (w_{in} A^T - w_{out} A) \overrightarrow{i}) \tag{2.8}
$$

The above equations requires $A$ to be stochastic matrix [29] which means all the blog posts must have at least one outlink. In other words, none of the rows in $A$ has all the entries as 0. Otherwise the influence score for such a blog post would be directly proportional to the number of comments. However, in the blogosphere,
this assumption does not hold well. Blog posts are sparsely connected. This problem can be fixed by making $A$ stochastic. This can be achieved by:

- Removing those blog posts with no outlinks and the edges that point to these blog posts while computing influence scores. This does not affect the influence scores of other blog posts, since the blog posts with no outlink do not contribute to the influence score of other blog posts.

- Assigning $1/N$ in all the entries of the rows of such blog posts in $A$. This implies a dummy edge with uniform probability to all the blog posts from those blog posts which do not have a single outlink.

For a stable solution of Eq. 2.8, $A$ must be aperiodic and irreducible [29]. A graph is aperiodic if all the paths leading from node $i$ back to $i$ have a length with highest common divisor as 1. One can only link to a blog post which has already been published and even if the blog post is modified later, the original posting date still remains the same. We use this observation to remove cycles in the blog posts by deleting those links that are part of a cycle and point to the blog posts which were posted later than the referring post. This guarantees that there would be no cycles in $A$, which makes $A$ aperiodic. A graph is irreducible if there exists a path from any node to any node. Using the second strategy mentioned above by adding dummy edges to make $A$ stochastic, ensures that $A$ is also irreducible.

As in [26,30,31], iFinder adopts an iterative method to compute the influence scores of blog posts. iFinder starts with little knowledge and at each iteration iFinder tries to improve the knowledge about the influence of the blog posts until it reaches a stable state or a fixed number of iterations specified a priori. The knowledge that iFinder starts with is the initialization of the vector $\vec{i}$. There are several heuristics that could be used to initialize $\vec{i}$. One way to initialize the influence score of all the blog posts is to assign each blog post uniformly a number, such as 0.5. Another way could be to use inlink and outlink counts in some linear combination as the initial values for $\vec{i}$. In our work, we used authority scores from Technorati which are available through their API\(^{15}\). One could also use PageRank values to initialize $\vec{i}$ but since we compare our results with PageRank algorithm we do not use it as the initial scores to maintain a fair comparison.

\(^{15}\)http://technorati.com/developers/api/cosmos.html
The computation of influence score of blog posts can be done using the well known power iteration method [32]. The underlying algorithm of iFinder can be described as: Given the set of blog posts $P$, $\{p_1, p_2, ..., p_N\}$, we compute the adjacency matrix $A$, and vectors $\overrightarrow{\lambda}$ and $\overrightarrow{\gamma}$. The influence vector $\overrightarrow{i}$ is initialized to $\overrightarrow{i}_0$ using Technorati’s authority values. Using Eq. 2.8 and $\overrightarrow{i}_0$, $\overrightarrow{i}$ is computed. At every iteration we use the old value of $\overrightarrow{i}$ to compute the new value $\overrightarrow{i}'$. iFinder stops iterating when a stable state is reached or the user specified iterations are exhausted, whichever is earlier. The stable state is judged by the difference in $\overrightarrow{i}$ and $\overrightarrow{i}'$, measured by cosine similarity. The overall algorithm is presented in Algorithm 1.

**Algorithm 1:** Compute Influence Scores of a Set of Blog Posts.

**Input:** Given a set of blog posts $P$, number of iterations $iter$, Similarity threshold $\tau$

**Output:** The influence vector, $\overrightarrow{i}$ which represents the influence scores of all the blog posts in $P$.

1. Compute the adjacency matrix $A$;
2. Compute vectors $\overrightarrow{\lambda}$, $\overrightarrow{\gamma}$;
3. Initialize $\overrightarrow{i} \leftarrow \overrightarrow{i}_0$;
4. repeat
   5. $\overrightarrow{i}' = \overrightarrow{i} (w_c \overrightarrow{\gamma} + (w_{in}A^T - w_{out}A) \overrightarrow{i})$;
   6. $iter \leftarrow iter - 1$;
   7. until $(\text{cosine_similarity}(\overrightarrow{i}, \overrightarrow{i}') < \tau) \lor (iter \geq 0)$

2.3.5. Issues of identifying the influentials

The preliminary model presents a palpable way of identifying influential bloggers and allows us to address many relevant issues such as evaluation, feasibility, efficacy, subjectivity, and extension.

- Can we use this model to differentiate influential bloggers from active bloggers? We study the existence of influential bloggers at a blog site by applying iFinder.

- How can we evaluate iFinder’s performance in identifying the influential bloggers? Are influential blog posts indeed different from non-influential blog posts?

- How can we properly determine the weights when combining the four parameters in $iIndex$? If one
changes the value of a weight, will the change significantly affect the ranking of influential bloggers?

How these weights can help find special influential bloggers?

- How does iFinder perform when compared against other models to find authoritative blogposts like PageRank [26]?

- How do we handle the subjectivity aspect of the problem of identifying influential bloggers as different people may have disparate preferences? Since we have access to the whole history of the blog site, we look into these questions by consecutively studying the influencers in multiple 30-day windows. Can we also employ the model to find any temporal patterns of the influential bloggers?

- Are all the four parameters necessary? We design and perform a lesion study and a correlation study. Some of the parameters may be correlated with each other, so one of them may be redundant. Lesion study is conducted taking one parameter out each time and comparing the results. Pairwise correlation analysis among all the four parameters is also conducted.

- How can we extend the preliminary model? Are there any other parameters that can be incorporated in a refined model?

In the next, we set out to use the proposed model in an empirical study, attempt to experimentally address these issues, report preliminary results, and suggest new lines of research in finding influential bloggers.

2.4. Experiments & Further Study

We first discuss the need for experimental data, and select a real-world blog site for experiments; and second, we design various experiments with the preliminary model using $iIndex$, and answer the questions raised in Section 2.3.5 based on the experimental results. In the process, we develop and elaborate an evaluation procedure for effective comparison.
2.4.1. Data Collection

Data collection is one of the critical tasks in this work. To our best knowledge, our effort is the first attempt to find influential bloggers. Hence, there are no available blog data sets for the purposes of our experiments. We need to collect real-world data.

There exist many blog sites. Some like Google’s Official Blog site act as a notice board for important announcements rather than for discussions, sharing opinions, ideas and thoughts; some do not provide most of the statistics needed in our work, although they can be obtained via some additional work (more explanation later). A few publicly available blog datasets like the BuzzMetric dataset\(^{16}\) were designed for different research experiments so there is no way to obtain some key statistics required in this work.

Therefore, we crawled a real-world blog site that provides the most statistics required in our experiments. The advantages of doing so include (1) minimizing our effort on figuring out ways to obtain the needed statistics, and (2) maximizing the reproducibility of our experiments independently. The Unofficial Apple Weblog (TUAW) site is such a site that satisfies these requirements. This blog site provides most needed information like blogger identification, date and time of posting, number of comments, and outlinks. The only missing piece of information at TUAW is the inlinks information, which we can obtain using Technorati API\(^{17}\). We crawled the TUAW blog site and retrieved all the blog posts published since it was set up. We have collected over 10,000 posts till January 31, 2007. We keep the complete history of the TUAW blog site and update it incrementally. All the statistics obtained after crawling is stored in a relational database for fast retrieval later\(^{18}\).

2.4.2. Results and Discussions

The following subsections introduce the experiments, results, and discussions corresponding to the questions raised in the Section 2.3.5.

\(^{16}\)http://www.nielsenbuzzmetrics.com/cgm.asp

\(^{17}\)http://technorati.com/developers/api/cosmos.html

\(^{18}\)This dataset will be made available upon request for research purposes.
2.4.2.1. Influential Bloggers and Active Bloggers

Many blog sites publish a list of top bloggers based on their activities on the blog site. The ranking is often made according to the number of blog posts each blogger submitted over a period of time. In this work, we call these people *active* bloggers. Since the top bloggers on the blog site TUAW are those from the last 30 days, we define our study window of 30 days as well. Using the number of posts of a blogger posted is obviously an oversimplified indicator, which basically says the most frequent blogger is an influential one. Such a status can be achieved by simply submitting many posts, as even junk posts are counted. Hence, an active blogger may not be an influential one; and in the same spirit, an influential blogger need not be an active one. In other words, the most active $k$ bloggers are not necessarily the top influential one, and an inactive blogger can still be an influential one.

In our first experiment, we generate a list of top-$k$ bloggers using the preliminary model proposed in Section 2.2. We set the default values of all the weights as 1 assuming they are equally important. An in-depth study of these weights is in Section 5.2.2. By setting $k = 5$, we compare the top 5 influential bloggers with the top 5 bloggers published at TUAW. Table 3 presents two lists of top 5 bloggers according to TUAW and based on the proposed model using iIndex: the first column contains the top 5 bloggers published by TUAW and the second column lists the top 5 influential bloggers. Names in *italics* are the bloggers present in both lists. Three out of 5 TUAW top bloggers are also among the top 5 influential bloggers identified by iFinder. This set of bloggers suggests that some of the bloggers can be both active and influential. Some active bloggers are not influential and some influential bloggers are not active. For instance, ‘Mat Lu’ and ‘Michael Rose’ in the TUAW list, so they are active; and ‘Dan Lurie’ and ‘Laurie A. Duncan’ in the list of the influentials, but they are not active.

In total, there could be four types of bloggers: both active and influential, active but non-influential, influential but inactive, inactive and non-influential. Since we have all the needed statistics, we can delve into the numbers and scrutinize their differences of the first three groups of bloggers. Their detailed statistics are presented in Table 4. *Inactive and non-influential bloggers* seldom submit blog posts and submitted posts do not influence others, so this group does not show up in Table 4.
TABLE 3

Two Lists of the Top 5 Bloggers According to Tuaw and iFinder, Respectively.

<table>
<thead>
<tr>
<th>Top 5 TUAW Bloggers</th>
<th>Top 5 Influential Bloggers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erica Sadun</td>
<td>Erica Sadun</td>
</tr>
<tr>
<td>Scott McNulty</td>
<td>Dan Lurie</td>
</tr>
<tr>
<td>Mat Lu</td>
<td>David Chartier</td>
</tr>
<tr>
<td>David Chartier</td>
<td>Scott McNulty</td>
</tr>
<tr>
<td>Michael Rose</td>
<td>Laurie A. Duncan</td>
</tr>
</tbody>
</table>

- **Active and influential bloggers** who actively post and some of them are influential posts. ‘Erica Sadun’, ‘David Chartier’ and ‘Scott McNulty’ are of this category. This can be verified by the large number of posts and the large number of comments and citations by other bloggers. For instance, ‘Erica Sadun’ submitted 152 posts in the last 30 days, among which 9 of them are influential, attracting a large number of readers evidenced by 75 comments and 80 citations.

- **Inactive but influential bloggers**. These bloggers submit a few but influential posts. ‘Dan Lurie’ published only 16 posts (much fewer than 152 posts comparing with ‘Erica Sadun’, an active influential blogger) in the last 30 days. Dan was not selected by TUAW as a top blogger. A closer look at his blog posts reveals that 4 of his blog posts are influential, i.e., 25% of the blog posts by ‘Dan Lurie’ are influential. One of his influential posts is about iPhone\(^1\)\(^9\), which attracted a large number of bloggers to comment and intrigued a heated discussion of the new product (77 comments and 33 inlinks). Its length is 1417 bytes, and there are no outlinks. All these numbers suggest that the post is detailed, innovative, and interesting to other bloggers. By reading the content, we notice that the post is a detailed account of his personal experience rather than extracts from external news sources. This kind of posts allows a reader to experience something new, thus often results in many comments and discussions.

- **Active but non-influential bloggers**. These bloggers post actively, but their posts may not generate sufficient interests to be ranked as the top 5 influentials. ‘Mat Lu’ and ‘Michael Rose’ were ranked 3\(^{rd}\) and 4\(^{th}\) top bloggers by TUAW, as they submitted 73 and 58 blog posts in the last 30 days (around 2

\(^1\)http://www.tuaw.com/2007/01/09/iphone-will-not-allow-user-installable-applications/
TABLE 4
Comparison of Statistics Between Different Bloggers.

<table>
<thead>
<tr>
<th></th>
<th>Num of Comments</th>
<th>Num of Inlinks</th>
<th>Blog Post Length</th>
<th>Num of Outlinks</th>
<th>Total Num of Blog Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max  Avg</td>
<td>Max  Avg</td>
<td>Max  Avg</td>
<td>Max  Avg</td>
<td>Max  Avg</td>
</tr>
<tr>
<td>Active &amp; Influential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erica Sadun</td>
<td>75  11.02</td>
<td>80  10.13</td>
<td>2935  830</td>
<td>15  2.53</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>David Chartier</td>
<td>56  11.31</td>
<td>32  10.25</td>
<td>3529  1055</td>
<td>14  4.35</td>
<td>68</td>
</tr>
<tr>
<td>Scott McNulty</td>
<td>112 11.56</td>
<td>33  8.925</td>
<td>2246  623</td>
<td>12  2.59</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inactive &amp; Influential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dan Lurie</td>
<td>96  19.63</td>
<td>37  10.26</td>
<td>1569  794</td>
<td>4  2.32</td>
<td>16</td>
</tr>
<tr>
<td>Laurie Duncan</td>
<td>65  16.29</td>
<td>34  10.61</td>
<td>2888  994</td>
<td>11  3.47</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active &amp; Non-Influential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mat Lu</td>
<td>42  8.029</td>
<td>29  10.01</td>
<td>1699  771</td>
<td>12  4.11</td>
<td>73</td>
</tr>
<tr>
<td>Michael Rose</td>
<td>31  8.727</td>
<td>21  9.606</td>
<td>1378  736</td>
<td>15  6.15</td>
<td>58</td>
</tr>
</tbody>
</table>

posts a day), respectively. Though these are much more than the 16 posts of ‘Dan Lurie’, they are not among the top 5 influential bloggers because their other statistics are not comparable with those of the influentials (i.e., having fewer comments and inlinks, and more outlinks).

A closer look at two influential blog posts: Here we further study the most influential blog posts by number one (‘Erica Sadun’) and number five (‘Laurie A. Duncan’) influential bloggers, respectively. The most influential blog post by ‘Erica Sadun’ is on keynote speech of Apple Inc. CEO, Steve Jobs\(^{20}\) which fostered overwhelming discussions through 63 comments and 80 inlinks. By reviewing the comments, we observe that most people appreciated her efforts and found the blog post extremely informative. The blog post was the first one dispensing a minute-by-minute description of the much-awaited keynote speech, new products, and services Apple would launch. The blog post was well-written and did not borrow information from any other sources. The most influential blog post by ‘Laurie A. Duncan’ detailed the violation of license agreements by macZOT\(^{21}\) with a developer\(^{22}\). This incident instigated a lot of discussion through 57 comments and 20 inlinks. Many people commented and cited this blog post, and agreed with the miserable state of license agreements, being appalled by how big companies could exploit small developers by finding loopholes in the laws. Similar sentiments expressed in a surge of comments are an important feature of many.


\(^{21}\)http://www.maczot.com/

TABLE 5
Intersection of Digg and Top 20 from iFinder.

<table>
<thead>
<tr>
<th>Bloggers</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influential</td>
<td>S1: 17</td>
<td>S2: 7</td>
</tr>
<tr>
<td>Non-influential</td>
<td>S3: 3</td>
<td>S4: 0/1</td>
</tr>
</tbody>
</table>

TABLE 6
Distribution of 100 Digg Blog Posts.

<table>
<thead>
<tr>
<th>Bloggers</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influential</td>
<td>S1: 71</td>
<td>S2: 14</td>
</tr>
<tr>
<td>Non-influential</td>
<td>S3: 8</td>
<td>S4: 7</td>
</tr>
</tbody>
</table>

TABLE 7
Distribution of 535 TUAW Blog Posts.

<table>
<thead>
<tr>
<th>Bloggers</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influential</td>
<td>S1: 327</td>
<td>S2: 42</td>
</tr>
<tr>
<td>Non-influential</td>
<td>S3: 131</td>
<td>S4: 35</td>
</tr>
</tbody>
</table>

influential blog posts. The above study of two most influential posts shows the efficacy of the proposed model.

2.4.2.2. Evaluating the Model

As we know, there is no training and testing data for us to evaluate the efficacy of the proposed model. The absence of ground truth about influential bloggers presents another challenge. The key issue is how to find a reasonable reference point for which four different types of bloggers can be evaluated so that we can observe their tangible differences. As an alternative to the ground truth, we resort to another Web2.0 site Digg (http://www.digg.com/) to provide a reference point. According to Digg, “Digg is all about user powered content. Everything is submitted and voted on by the Digg community. Share, discover, bookmark, and promote stuff that’s important to you!”. As people read articles or blog posts, they can give their votes in the form of digg and these votes are recorded on Digg servers. This means, blog posts that appear on Digg are liked by their readers. The higher the digg score for a blog post is, the more it is liked. In a way, Digg can be considered as a large online user survey. Though only submitted blog posts are voted, Digg offers a way for us to evaluate the blog posts of the four types. Digg provides an API to extract data from their database for a window of 30 days. We used this API to obtain the data for the month of January 2007. Given the nature of Digg, a not-liked blog post will not be submitted thus will not appear in Digg. For January 2007, there were
in total 535 blog posts submitted on TUAW. As Digg only returns top 100 voted posts, we use these 100 blog posts at Digg as our benchmark in evaluation. Next, we present the results for model efficiency based on the comparison between iFinder and Digg.

We take the four categories of bloggers, viz. 1. Active and Influential, 2. Inactive and Influential, 3. Active and Non-influential, and 4. Inactive and Non-influential and categorize their posts into S1, S2, S3, and S4, respectively. We rank the blog posts of each category based on the influence score and pick top 20 blog posts from each of the first three categories. We randomly pick 20 blog posts from the last category in which bloggers are neither active nor influential. Next we compare these four sets of 20 blog posts with the Digg set of 100 blog posts to see how many posts in each set also appear in the Digg set. The results are shown in Table 5. From the table, we can see that S1 has 17 out of 20 in the Digg set, and S4 has 0 or 1 found in the Digg set depending on randomization. The results show the differences among the four categories of bloggers and iFinder identifies the influentials whose blog posts are more liked than others according to Digg. For reference purposes, we also provide the distributions of 100 Digg and 535 TUAW blog posts in Tables 6 and 7, respectively. Note that we selected top 5 active and 5 influential bloggers (Table 3), in which 3 are both active and influential (Table 4). We observe from Tables 5, 6 and 7 that influential bloggers are more likely to be liked than active bloggers.

Some observations that could be derived from this experiment are: (1) Compare active and influential bloggers with active and non-influential bloggers in Tables 6 and 7. 21.71% (= 71/327) of blog posts from S1 were liked by people, but only 6.1% (= 8/131) of blog posts from S3 were liked by the people, according to Digg. This shows that if a blogger is influential then (s)he is liked by people and not when they are active. Results from iFinder in Table 5 are also consistent with this observation. On the other hand, compare influential and active bloggers with influential and inactive bloggers in Tables 6 and 7. 21.71% (= 71/327) of blog posts from S1 were liked by the people and 33.33% (= 14/42) of the blog posts from S2 were liked by the people. This shows that no matter if the blogger is active or inactive, if they are influential they will be liked by the people, according to Digg. Results from iFinder in Table 5 are also consistent with this observation. These two facts bring out the difference between influential and active bloggers. Influential
TABLE 8

Overlap Between Top 20 Blog Posts at Digg and iFinder for Last 6 Months for Different Configurations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>All-in</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>No Inlinks</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No Comments</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>No Outlinks</td>
<td>11</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>No Blog post length</td>
<td>12</td>
<td>14</td>
<td>11</td>
<td>15</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

bloggers are more liked as compared to active bloggers. (2) According to S3 in Table 5, Table 6, and Table 7, active bloggers are not necessarily influential while according to S1, influential bloggers may be active. (3) In Table 6, S4 has 7 blog posts liked by people even though they were non-influential and inactive. This is because one of the bloggers in S4 was ranked 6th in the list of influential bloggers and 4 of his blog posts appeared in Digg. So in such cases where the blogger is on borderline we could get good overlap values for S4 too.

This work is different from typical classification tasks, in the sense that there is no training and testing data so we cannot learn a model from the training data and evaluate it on the testing data. Since there is an absence of ground truth or the real knowledge about the influential bloggers, we treat people’s opinions as close to ground truth while evaluating iFinder. Digg\textsuperscript{23} has been used to evaluate iFinder.

Since Digg assigns score to blog posts and not bloggers, we compare the top influential blog posts from Digg\textsuperscript{24} and iFinder. We compare top 20 blog posts\textsuperscript{25} for every month for the last six months starting from January 2007 till June 2007. We report the overlap in the two lists, since there is not 100% overlap so we do not report rank correlation coefficients like Kendall-Tau rank correlation coefficient or Spearman’s rank correlation coefficient. We try different configurations of iFinder by considering 1. All-in i.e. all the four parameters, 2. No inlinks (outlinks, comments, and blog post length), 3. No comments (inlinks, outlinks,

\textsuperscript{23}http://www.digg.com/
\textsuperscript{24}We get this data using Digg API.
\textsuperscript{25}On average, 70-80 blog posts from TUAW are submitted to Digg every month, so we pick top 20 to avoid under-sampling or over-sampling.
blog post length), 4. No outlinks (inlinks, comments, blog post length), and 5. No blog post length (inlinks, outlinks, comments). We report the overlap results for all these 5 configurations with Digg in Table 8.

We get the best overlap for May 2007 i.e. 80% and 50% overlap for February 2007 for “All-in” configuration. Considering all the parameters, on average we achieve 65.83% overlap. This overlap indicates how much our proposed model aligns with people’s opinions on Digg. Being an empirical model, there is a lot of room for improvements in this model.

We also studied the contribution of different parameters and their relative importance from the experiments with the other 4 configurations. From the results in Table 8, it can be observed that configuration 2 (no inlinks) always performs the worst, configuration 3 (no comments) performs better, then comes configuration 4 (no outlinks) and then come configuration 5 (no blog post length). This gives us the order of importance of all the four parameters, i.e. inlinks > comments > outlinks > blog post length, in the decreasing order of importance to influence estimation. Given this analysis, we can adjust the weights for different parameters to achieve better than “All-in” results. Later in Section 2.4.2.4 we experiment by adjusting weights and observe how different weighting schemes generate special types of influential bloggers.

Next we study the power law characteristics ($p(x) \propto x^{-\gamma}$) exhibited by the Digg and compare it with the same characteristics of iFinder. Here $\gamma$ denotes the exponential factor of the power law. We plot the log-log graph ($\log(p(x)) \propto -\gamma \log(x)$) of Digg scores and influence scores from Digg and iFinder respectively.
Exponent \((-\gamma)\) of the power law is denoted by the slope of the curve. The results are presented in Figure 2 and Figure 3. Here the \(y\)-axis represents the Digg scores (obtained through Digg) and influence scores (obtained through iFinder) of blog posts. \(x\)-axis denotes the blog posts ranked based on Digg scores and iFinder, respectively. It is evident from the linear behavior of the curves in Figure 2 and Figure 3 that both the models obey power law. However, the exponent \(\gamma\) is higher for Digg as compared to iFinder because initial 5 blog posts in Figure 2 have more than 1000 Digg score and rest of them have less than 100 Digg score. Whereas, the influence score for all the blog posts are below 1000 for iFinder. This shows that the digg scores of the blog posts fall drastically as compared to the influence score through iFinder. This can be explained because the users can view the Digg scores which contributes more diggs to already popular blog post and helps in making rich even richer faster. However, the influence scores from iFinder are not seen by the users which explains the smaller exponential factor \(\gamma\) and smoother fall of influence scores in iFinder.

2.4.2.3. Influential vs. Non-Influential Blog Posts

Here we study the contrast in the characteristics between influential and non-influential blog posts. Using the definition of influential blog posts from Section 2.2, we pick influential blog posts submitted by the influential bloggers listed in Table 3. Rest of the blog posts are treated as non-influential blog posts. Totally we have 22 influential and 513 non-influential blog posts for January 2007. Similar to Table 4, we compare the max and average statistics for all the four parameters (comments, inlinks, blog post length, and outlinks) for both influential and non-influential blog posts and report the results in Table 9. It shows influential blog posts are much longer in length and have far more comments. There are a lot more inlinks in influential blog posts, but the number of outlinks is a weaker piece of evidence, though the influential blog posts have slightly smaller number of outlinks.

2.4.2.4. Effects and Usages of Weights

There are four weights in our preliminary model to regulate the contribution of four parameters toward the calculation of the influence score using Eq 2.2 & Eq 2.4. To recall, \(w_{in}\) is for the influence from incoming links, \(w_{out}\) for the influence from outgoing links, \(w(\lambda)\) for the “goodness” of a blog post, and \(w_{comm}\) for the
TABLE 9
Comparison of Statistics Between Influential and Non-Influential Blog Posts.

<table>
<thead>
<tr>
<th></th>
<th>Num of Comments</th>
<th>Num of Inlinks</th>
<th>Blog Post Length</th>
<th>Num of Outlinks</th>
<th>Total Num of Blog Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max</td>
<td>Avg</td>
<td>Max</td>
<td>Avg</td>
<td>Max</td>
</tr>
<tr>
<td>Influential Blog Posts</td>
<td>112</td>
<td>74.18</td>
<td>80</td>
<td>38.63</td>
<td>3529</td>
</tr>
<tr>
<td>Non-Influential Blog Posts</td>
<td>69</td>
<td>10.84</td>
<td>39</td>
<td>8.96</td>
<td>1930</td>
</tr>
</tbody>
</table>

number of comments. All weights take real values in $[0, 1]$. We now study how the change of their values will affect the ranking of the influentials.

One may notice that $w(\lambda)$ simply scales the influence score of a blog post, so varying $w(\lambda)$ is not expected to affect the ranking of influential bloggers, but to scale up or down the influence scores. This is verified by conducting experiments in which the other three weights are fixed and only $w(\lambda)$ is varied. We observe that the relative ordering of the influential bloggers remain the same while their influence score is scaled up or down. Although this weight is immaterial for identifying the influentials at one blog site, it can be used in comparing the influential bloggers of different blog sites for normalization purposes (outside the scope of this work).

For the remaining three weights, $w_{comm}$, $w_{in}$ and $w_{out}$, we fix two of them and observe how the ranking changes by varying the third weight. Fixing $w_{in}$ and $w_{out}$ and varying $w_{comm}$ from 0.0 to 1.0 in steps of 0.1, we observe that the model stabilizes for $w_{comm} \geq 0.6$, i.e., it does not change the ranking of the influential bloggers. While varying $w_{in}$ and $w_{out}$ respectively, we observe that the model stabilizes when $w_{in} \geq 0.9$ and $w_{out} \geq 0.2$. To summarize, we obtain the same ranking of influential bloggers as shown in the right column of Table 3 for $w_{comm} \geq 0.6$, $w_{in} \geq 0.9$, $w_{out} \geq 0.2$.

Clearly, the value change of the above three weights can lead to different rankings. This allows one to adjust the weights of the model to attain different goals. With the preliminary model of default setting, we can tune these weights in identifying influential bloggers with different characteristics. For example, by
setting $w_{in}$ and $w_{out}$ to 0, we can obtain influential bloggers based on the number of comments a blogger’s post obtained. Similarly we can obtain the blog post that received most citations or the blog post including the least outlinks. If one wants to emphasize one aspect, one can tune weights and obtain ranking to reflect that aspect. The increase of $w_{out}$ is one way to discourage the citations of other blog posts, in a way, encouraging a post with independent ideas. In short, these weights provide a means to further evolve and expand the preliminary model for a wide range of applications.

2.4.2.5. iFinder vs. PageRank

We compared iFinder with Google PageRank. We used Google’s blog search interface to obtain the ranked list of blog posts according to the PageRank values because of two primary reasons. First, we do not implement
the PageRank algorithm to avoid concerns regarding accurate implementation. Second, Google keeps on evolving their algorithm so the search interface has the latest and most advanced version of the PageRank which is certainly better than the primitive and published version of the PageRank [26].

We compared top 20 blog posts in a pairwise fashion from iFinder, PageRank, and Digg for each month starting from January 2007 to Jun 2007. The results are reported in Table 10. Unlike Digg there is no issue of coverage with Google’s PageRank comparison. Google indexes all the blog posts available at TUAW. Since our comparison is on monthly basis, we check this by looking at the total number of results displayed by Google and the total number of blog posts submitted at TUAW for each month. It is evident from Table 10 that; first, iFinder performs better than Google’s PageRank when compared to Digg as the ground truth. Second, to rule out the possible explanation that Digg does not cover all the blog posts so Google-Digg overlap is poor; we study the overlap between Google and iFinder, since both covers all the blog posts submitted. Results show that there is an insignificant overlap between Google and iFinder. Third, Google’s model is less aligned with Digg as compared to iFinder, which shows that Google’s blogpost relevance ranking doesn’t fit well with the taste of the people.

2.4.2.6. Temporal Patterns of the Influentials

Above, we study the influential bloggers with a time window of 30 days (or monthly). For a blog site that has a reasonably long history, we can also study the temporal patterns of its influential bloggers. The blog site TUAW provides blogging data since its inception February 2004. We hence apply iFinder to identify top 5 influential bloggers with a moving 30-day window until January 2007, and there is no overlap between two consecutive windows. In total, there are 26 influential bloggers during Feb.2004-Jan.2007. The temporal patterns of the influentials can be observed from a matrix in Figure 4. Influential bloggers are ordered according to the time they were recognized as influential vertically(column-wise), and the rows represent the progression of time. The \((i, j)\)-th cell in this matrix stores the rank of the \(j^{th}\) blogger in the \(i^{th}\) time window. For example, the first cell \((sean bonner, Feb-04)\) shows that Sean Bonner was ranked top 1 among the influential
bloggers list in February 2004\textsuperscript{26}. Black cells represent that the particular blogger was not among the top 5 for that time period. The color gradient represents rank of a influential blogger, a darker color representing a better rank.

We can observe some different temporal patterns for the influentials in Figure 4. Among all the 26 bloggers, 17 are influential for at least 4 months. We broadly categorize the influential bloggers into the following:

**Long-term influentials** They steadily maintain the status of being influential for a very long time. Scott McNulty is the best example of this category. Scott McNulty is steadily influential from Jan-05 till Jan-07. They can be considered “authority” in the community.

**Average-term influentials** They maintain their influence status for 4-5 months. Examples of such bloggers from Figure 4 are “Sean Bonner”, “Gregory Han”, and “Barb Dybward”.

**Transient influentials** They are influential for a very short time period (only one or two months). Examples are Michael Sciannamea, Fabienne Serriere and Dan Pourhadi. For instance, Fabienne Serriere was influential in Jan-06 and never became influential again.

**Burgeoning influentials** They are emerging as influential bloggers recently. Bloggers that belong to this category are Dan Lurie and Erica Sadun. They are the influentials worthy more follow-up examinations.

Disparate bloggers can present different temporal patterns. Long-term influentials are more influential than other bloggers as they are more trustworthy as compared to other bloggers based on a long time of history. Burgeoning influentials have potential to become long-term ones. But it is difficult to say these things about transient influentials as they might become influential by chance. Certainly, there could be many other temporal patterns depending on a particular application. The categories presented here are some examples. Many potential applications can be developed using categories. When we want to know about a new blog site, the best way to approach it is to look at its long-term influentials as they have lasting influence.

\textsuperscript{26}In early stage of the blog site, there are a few cases in which there was little blogging activity such as Feb-04, Oct-04, and Nov-04, resulting in fewer than 5 influentials.
in the community. The blog posts of those average-term influentials can be used to understand the changing topics. The blog posts of burgeoning influentials might contain the trendy buzz. With accumulated blogging data, we can also learn to predict if a burgeoning influential will more likely become long-term, average-term, or transient.

2.4.2.7. Further Experiments

We conduct more experiments to (1) verify if any of the four factors (number of comments, inlinks, outlinks, and length of a blog post) can be eliminated via a lesion study; (2) examine the pairwise correlations of the four factors; and (3) study another statistics - the rate of comments to extend iFinder.

Lesion study. We study the performance of the model by removing one parameter in turn. That is, we compute the influence scores using only the remaining three parameters. We rank the top 5 bloggers by leaving one parameter out and thus obtain 4 ranking results, comparing with the result of “All-in” (with all four parameters). Had there been a parameter that did not contribute to the influence score, removing it would not result in any difference in the ranking. The results are presented in Figure 5. The x-axis denotes different ranking schemes to find the influentials. For example, “No outlinks” signifies the ranking of influential bloggers computed using inlinks, comments, and post length, but leaving outlinks out. Interestingly, all the top 5 influentials remain unchanged, but their relative ranks vary. It is evident that no blogger maintains the same rank in all the five variations and no two ranked lists are the same. Thus, the four parameters contribute in the preliminary model in determining influential blogger. As discussed in Section 2.4.2.4, the trade-off
between the parameters can be achieved by adjusting their associated weights to accommodate different needs.

**Correlation analysis.** We perform pairwise correlation analysis between the parameters to further examine whether there is any redundant parameter. With four parameters, there are 6 pairwise correlations as shown in Figure 6(a)-(f). The number below each scatter plot is the correlation coefficient. We observe that there is no strong correlation between any pair of parameters. In other words, none of the parameters can be covered by another one. We notice that 5 of 6 scatter plots show positive correlations, but the (d) scatter plot shows some negative correlation, which suggests that more outlinks in a blog post somehow mean fewer comments the post receives, and vice versa. This supports that links among blog posts are different from web links (Section 2.5).

**Rate of comments.** This parameter seems a good indicator on how influential a post is. If a post receives many comments in a short period (i.e., it exhibits a spike), it has apparently generated a lot of response,
indicating that the post is potentially influential. However, is the opposite true too, i.e., the observation of a flat distribution of comment rates of a blog post implies a non-influential post? We conduct a case study and present the results in Figures 7 and 8 with comment rates of two influential blog posts: one related to the newly publicized iPhone release and the other about a competition held at Apple Inc. Figure 7 exhibits a spiky type of user response. Most of the comments were submitted during the first hour (over 50) after the blog post was published. On the other hand, comment rates in Figure 8 are relatively “flat”, around 10 comments per hour even after 7 or 8 hours of the blog post submission. Since the spiky pattern is not a necessary characteristic of an influential post, more research is needed to explore how to incorporate the comment rate. We envision that this parameter can be used to build a more refined model for special time-critical applications like disaster prevention and management, emergency handling.

Other extensions to the preliminary model include 1). study of spam comments filtering to prevent spam attacks using techniques mentioned in [24, 25], 2). study more appropriate blog post quality estimation techniques involving content and literary analysis, and 3). study different functions to non-linearly penalize influence due to outlinks. This basically means assigning negligibly small penalty if few outlinks are present and very high penalty for outrageous number of outlinks. This is required to avoid penalizing those novel
blog posts that refer to a few blog posts to support their explanation. One such function could be exponential which would replace \( w_{out} \sum_{n=1}^{l} I(p_{n}) \) in Eq. 2.2 with \( \exp(w_{out} \sum_{n=1}^{l} I(p_{n})) \). We would have to investigate thoroughly the role of \( w_{out} \) in such a scenario.

2.5. Related Work

Blogosphere has been expanding speedily since its inception. This has attracted a surge of research on Blogosphere. Authors in [33] consider influence a characteristic of virtual communities, among others like membership, reinforcement of needs, shared emotional connection, whose presence governs the establishment of a community. Link structures and overlap between different sub-communities are used to help identify influence between them. Next we review briefly existing works in the area of influential blog sites and blog leaders. We compare and contrast these approaches with the proposed model in this work.

2.5.1. Influential Blog Sites

Finding influential blog sites in the blogosphere is an important research problem, which studies how some blog sites influence the external world and within the blogosphere [34]. It is different than the problem of identifying influential bloggers in a community. Blogosphere follows a power law distribution [35] with very few influential blog sites forming the short head of the distribution and a large number of non-influential
sites form the long tail where abundant new business, marketing, and development opportunities can be explored [36]. Our work is about identifying influential bloggers at a blog site regardless of the site being influential or not. We briefly review some work on influential blog sites.

Researchers have studied the blog graph from the perspective of information diffusion and identify the key players who maximize the spread [37]. Gruhl et al [38] study information diffusion of various topics in the blogosphere from individual to individual, drawing on the theory of infectious diseases. A general cascade model [39] is adopted. They associate ‘read’ probability and ‘copy’ probability with each edge of the blogger graph indicating the tendency to read one’s blog post and copy it, respectively. They also parameterize the stickiness of a topic which is analogous to the virulence of a disease.

An interesting problem related to viral marketing [16, 40] is how to maximize the total influence in the network (of blog sites) by selecting a fixed number of nodes in the network. A greedy approach can be adopted to select the most influential node in each iteration after removing the selected nodes. This greedy approach outperforms PageRank, HITS and ranking by number of citations, and is robust in filtering splogs (spam blogs) [41]. Leskovec et al. [42] proposed a submodularity based approach to identify the most important blogs which outperforms the greedy approach. Nakajima et al. [43] attempts to find agitators, who stimulate discussions; and summarizers, who summarize discussions, by thread detection. Watts and Dodds [44] [45] studied the “influentials hypothesis” using computer simulations of interpersonal influence processes and found that large cascades of influence are driven by a critical mass of easily influenced individuals. The phenomenal growth of the blogosphere along with increasing link sparsity presents significant challenges to employ purely link analysis based approaches.

In this work we deal with identifying influential bloggers at one blog site which differs from those briefly reviewed above. A blog site is a special type of social network that contains information such as outlinks (other blog posts it is referring to), inlinks (other blog posts that are citing this blog post), comments which is not present in a general social network. Identifying the influential bloggers at a blog site requires the integrated use of the information specific to a blog site.
2.5.2. Blog Leaders

In the past, researchers have studied various forms of blog leaders in the blogosphere. Here we analyze these different types of blog leaders and compare with influential bloggers.

Authors in [46] categorize the blogs into two classes, “Affective Blogs” and “Informative Blogs”. Affective blogs are those that are more like personal accounts and diaries form of writings. Informative blogs are more technology oriented, news related, commonsense knowledge, objective, and high-quality informative blogs. Training a binary classifier on a hand-labeled set of blogs using Naïve Bayes, SVM, and Rocchio classifier they separate the affective blogs from the informative blogs. However, there could be influential bloggers who write affective blogs which would be missed by such an approach.

Kavanaugh et al. [47] study the existence of “Opinion Leaders” in the physical world and observe there role when they engage in some form of blogging. Song et al. [48] define “Opinion Leaders” on the blogosphere as those who bring in new information, ideas, and opinions, then disseminate them down to the masses. They rank the blogs using a novelty score measured by the difference in the content of a given blog post and ones that refer to this blog post. First the blog posts are reduced to topic space using Latent Dirichlet Allocation (LDA) and then using cosine similarity measure between these transformed blog posts novelty score is computed. However, opinion leaders are different from influential bloggers. There could be a blogger who is not very novel in his content but attracts a lot of attention to his posts through comments and feedback. These type of bloggers will not be captured by novelty based approach. Moreover, not many blogs refer to the blogs they borrowed their content from, due to the casual nature of the blogosphere.

Many blog sites list “Active Bloggers” or top blog posts in some time frame (e.g., monthly). Those top lists are usually based on some traffic information (e.g., how many posts a blogger posted, or how many comments a blog post received) [34]. Certainly these statistics would leave out those blog sites or bloggers who were not active. Moreover, influential bloggers are not necessarily active bloggers at a blog site.

2.6. Summary

Blogosphere is one of the fastest growing, social media. The virtual communities in the blogosphere are not constrained by physical proximity and allow for a new form of efficient communications. The influential
bloggers naturally exert their influence on other members, lead trends, and affect group interests in a community. They are the conduits of information in their communities. With many great successes of Web 2.0 applications, more and more people take part in one form or another of activities in virtual communities. Finding the influential bloggers will not only allow us to better understand interesting activities happening in a virtual world, but also present unique opportunities for industry, sales, and advertisements. With the speedy expansion of the blogosphere, it is vital to develop novel tools that facilitate people to participate, connect, and explore.

We address a novel problem of identifying influential bloggers at a blog site by presenting a preliminary model of identifying influential bloggers of a community blog site. Our work differs from existing works on blogosphere influence over traditional media, influential blog sites, and influence maximization within the blogosphere. Influential bloggers can exist at many blog sites, regardless of these sites being influential or not. We examine essential issues of identifying influential bloggers, evaluate the effects of various collectable statistics from a blog site on determining blog-post influence, develop unique experiments using another Web2.0 application, and conduct experiments by using the whole history of blog posts of a real-world blog site. The results demonstrate that (1) influential bloggers are not necessarily active bloggers, (2) iFinder can effectively find influential bloggers, (3) by tuning the weights associated with the parameters of the preliminary model, one can examine how different parameters impact the influence ranking for different needs, and (4) the preliminary model can serve as a baseline in identifying influential bloggers and can be extended by incorporating additional parameters to discover different patterns. We expect that the preliminary model will evolve to address many new needs arising from the real (or rather virtual) world.

In this work, we study a novel problem of identifying influential bloggers at community blogs. Nevertheless, the blogosphere consists of more individual blogs than community blogs. So as a part of future work we would like to extend the study to include individual blogs as well. However, since individual blogs are single authored, it is insensible to find influential bloggers. This could be achieved by synthesizing a virtual community of similar individual blogs. We have published some works that aggregate the individual blogs
that are similar and were not connected previously [5, 8]. These aggregated individual blogs could be treat as a virtual community.
3. CLUSTERING BLOGS BY LEVERAGING COLLECTIVE WISDOM

3.1. Introduction

The advent of Web 2.0 [13] has created a surge of content via online media such as blogs, wikis, social bookmarking like del.icio.us, online photo sharing like Flickr and other such services. Although blogging is not a new phenomenon, it has been there for the past 7-8 years, Web 2.0 has helped make it more accessible and permeate the society rapidly. Journalism “by the people” or citizen journalism began to flourish. In 2006 Time named “You” as Person of the Year, due to the growth of user-generated-content.

People not only generate content but also enrich the existing media (both text and non-text). A popular example is Google’s image labeler, which is a social game that involves human subjects labeling images and retaining the most commonly used tags for a particular image. Other examples include tagging blog sites and blog posts with relevant labels, providing feedback for search results, etc. This is known as a new form of collective wisdom. Collective wisdom (also known as group wisdom, wisdom of crowds, open source intelligence and co-intelligence) is defined as the shared knowledge arrived at by a group of individuals, used to obtain the best possible approximation to the perfect solution. Here, the terms blog site and blog are used interchangeably, where a blog site refers to the collection of blog posts.

Till August 2007, Technorati was tracking over 70 million weblogs, and were seeing about 120,000 new weblogs being created worldwide each day. That’s about 1.4 blogs created every second of every day. By August 2008, the amount of weblogs tracked had increased by another 7.7 million, as reported by Technorati¹. With such a significant growth of the blogosphere there is a need for automatic and dynamic organization of the blog sites. Clustering of these blog sites is a promising way to achieve the automatic organization of the content. In this work we focus on the challenges involved in clustering blog sites by leveraging the available collective wisdom.

Blog site clustering not only helps better organize the information but also aids convenient accessibility to the content. Clustering blog sites helps in optimizing the search engine by reducing the search space. We only need to search the relevant cluster and not the entire blogosphere.

The Blogosphere follows a scale-free model and obeys Long Tail distribution [36]. A vast majority of bloggers reside in the Long Tail and are currently not well targeted for otherwise potential business opportu-

¹http://technorati.com/blogging/state-of-the-blogosphere/
nities (i.e., niches). To do better requires a good number of bloggers that can provide more data for targeted marketing. This warrants a need for aggregating the Long Tail bloggers. Clustering various Long Tail bloggers to form a critical mass will not only potentially expand a blogger's social network, but also increase participation so as to move them from the Long Tail towards the Short Head. This could help the search engines to expand their result space and include results beyond just the Short Head. Including relevant clusters from the Long Tail in the search space would help in identifying those niches.

Clustering blog sites leads to connecting the bloggers. Connecting the bloggers in the Long Tail helps in identifying familiar strangers [49, 50]. The underlying concept of familiar strangers is that they share some patterns and routines (or commonalities), although they are not directly connected. Clustering blog sites also helps promote the Web 2.0 new marketing 4Ps [51]: personalization, participation, peer-to-peer, and predictive modeling.

The rest of the chapter is organized as follows. Section 3.2 covers related work. Section 3.3 presents the problem definition with the challenges. Section 3.4 describes the proposed approach. Section 3.5 presents the experimental design and results. We summarize the work in Section 3.6 with possible future directions.

3.2. Related Work

3.2.1. Blog Clustering

Blogs have only recently become a subject of research. Authors in [52], [53], [33] and [54] explore clustering of the blogs to identify communities. Blog community extraction based on clustering relies on user induced connections in the underlying blog network. However such blog clusters only identify the community structure of the blogosphere and may not help in clustering of blogs with similar contents. A closer attempt to cluster blogs based on similar content is done by [55] utilizing the blog tags for hierarchical clustering. Work by [56] performs semantic analysis in order to discover topic trends, with the goal of identifying clusters that persists over time. The clusters are based on identifying biconnected components in a graph.

Though content based clustering of blogs has not been studied widely, content or topic based clustering of web documents and text has been widely studied. In [57] and [58] web document clustering has been done based on the k-means algorithm [59]. Apart from k-means, agglomerative and hierarchical clustering has
also been used for document clustering. Such is the case for [60], which uses a hierarchical structure for linkage based clustering measured by similarities of other objects linked to a pair of objects, where objects can refer to authors, papers, links, and web sites. Similar work is accomplished by [61], where the authors use hierarchical clustering to identify communities by establishing connections per the co-occurrence of words in entities such as web pages and blogs. Authors in [62], [63], and [64] present reviews of clustering algorithms for a collection of documents.

The above mentioned clustering algorithms can be directly applied to blogs by considering blogs as web documents. The blogs represented using the term vector-space model as term frequency vectors can then be clustered from a similarity matrix. However by doing this we would be ignoring the many unique characteristics of blogs which would aid us in obtaining a better clustering. Blog sites are not as rich in text as web documents. Most blog sites are personal accounts, opinions, ideas, thoughts, and expressions that has less content and not well-authored. However, labels or tags assigned by humans (both bloggers as well as readers) also known as the collective wisdom make them special and different from web documents. The Web and the blogosphere are often compared and the existing approaches in webpage clustering or web community discovery are explored for their use in similar tasks in the blogosphere. But there is a key difference between the blogosphere and the Web in terms of the lifetime of the content (pages and links) posted on both the media. Blogs are mainly used as a tool for communication. So the entries in the blogs are very short-lived. Most of them become obsolete and are never referred later. Thus the links in the blogs have significant temporal locality. However in the Web, new pages may refer to very old pages (e.g., an authoritative webpage) creating a longer lifetime of content in the Web. We may also aggregate the links in the blogs over time but that misses the key temporal dynamics observed in the blogosphere. Bloggers’ interests shift over time. For instance, a blogger is initially interested in Politics so he interacts more often with fellow bloggers who are also interested in Politics, forming a community. Later his (her) interest shifts to Economics so (s)he shifts the interactions with the bloggers who are also interested in Economics. Now if all the interactions are aggregated over time we would lose this temporal dynamics in the interaction patterns and community evolutions. Based on this key difference between the Web and the blogosphere, conventional webpage clustering
or community discovery algorithms do not work well in the blogosphere domain. In this chapter, we will also explore some of the techniques that focus on the dynamics of interactions or the evolution of blog communities. Traditional keyword clustering algorithms mentioned above would fail to return good results due to sparsity and curse of dimensionality, therefore novel techniques are required that leverages the enormous collective wisdom available.

3.2.2. Leveraging Tag Information

Collective wisdom as represented by the labels or tags provided by humans have been previously used for various tasks like search and retrieval and recommender systems. The human annotations provided for web pages and blogs provide valuable metadata for use in search. Websites like ‘del.icio.us’, and ‘Flickr’ use such user provided metadata in the form of collaborative tagging for search and retrieval. Since large amount of such metadata is available even in the blogosphere, it can be leveraged for search and retrieval operations. Authors in [65] provide an algorithm to search using the tag information. In [2] the authors have used the tag information for a blog recommendation algorithm.

Though the use of ‘collective wisdom’ has been studied as mentioned above, there is still an opportunity for improvements in terms of using a greater variety of user generated data (like user provided labels in blogs) and for more kinds of applications in the blogosphere (like blog clustering) and the web in general. The algorithm presented in this work leverages the collective wisdom extracted from the user-provided annotations.

3.3. Problem Definition

In this work we consider the problem of clustering blog sites. More formally, given a set \( T \) of \( m \) blog sites, \( \{b_1, b_2, \ldots, b_m\} \), we construct \( k \) clusters \( \{T_1, T_2, \ldots, T_k\} \) of the \( m \) blog sites, such that \( k \leq m \) and \( \bigcup_{i=1}^{k} T_i = T \). We exploit the collective wisdom while forming clusters of these blog sites. The collective wisdom is available in the form of predefined labels for each blog site. A single blog site could be tagged under multiple labels. More on how these labels are used to cluster the blog sites will be discussed in Section 3.4.2.

With the proposed framework we intend to explore new ways for clustering. We also show that the clusters
thus obtained are more meaningful as compared to traditional ways for clustering. Moreover, conventional approaches for clustering have inherent shortcomings like,

- Text clustering suffers from the curse of dimensionality and sparsity [66].
- The similarity measure may not capture the semantic similarity very well [67].
- The clusters thus obtained are sometimes not very meaningful [67].
- User needs to specify the number of clusters \textit{a priori} which could be hard to anticipate [68].

Next, we propose a novel clustering approach that leverages collective wisdom and tackles the challenges in conventional clustering approaches listed above.

3.4. Generating Similarity for Blog Clustering

3.4.1. Leveraging Collective Wisdom

We leverage the label information available with each blog site in clustering them. A naïve way could be to treat all the blog sites that have same label as one cluster but that would result in too many clusters and moreover some of the clusters thus obtained might be related and would be better if they are merged into one cluster. Also many blog sites are tagged under more than one label, which makes it difficult to form clusters in the naïve way. To achieve this, we cluster similar labels.

Clustering the similar labels can be formulated as an optimization problem. Assume we have \( t \) labels, \( l_1, l_2, ..., l_t \) and are clustered into \( k \) clusters, \( C_1, C_2, ..., C_k \), then optimal clustering is obtained if, for any two labels \( l_i \) and \( l_j \),

\[
\min \sum d(l_i, l_j), \forall (l_i, l_j) \in C_m, 1 \leq m \leq k, i \neq j
\] (3.1)

\[
\max \sum d(l_i, l_j), \forall l_i \in C_m, \forall l_j \in C_n, 1 \leq (m, n) \leq k, m \neq n
\] (3.2)

Here \( d(l_i, l_j) \) refers to a distance metric between the labels \( l_i \) and \( l_j \). Formulation 3.1 minimizes the within-cluster distance between the cluster members and Equation 3.2 maximizes the between-cluster distance. Finding efficiently an optimal solution for the above min-max conditions is infeasible.
Existing work like [69] proposes a method for clustering based on maximum margin hyperplanes through the data by posing the problem as a convex integer program. The hard clustering constraint is relaxed to a soft clustering Equation that can be feasibly solved with a semidefinite program. In a probabilistic approach, data is considered to be identically and independently drawn from a mixture model of several probability distributions [70]. An expectation maximization (EM) based approach is used to first estimate conditional probabilities of a data point \( x \) given a cluster \( C \) by \( P(x|C) \) and then find an approximation to a mixture model given the cluster assignments. k-means is an approximation to EM based clustering approach. Another approach to cluster the blog sites is based on the tags assigned to the blog posts and the blog site. Each blog site can be profiled based on these accumulated tags. A simple cosine similarity distance metric could be used to find similarity between different blog sites. However, the vector space model of the blog sites based on the tags is high-dimensional and sparse. We use a SVD based clustering algorithm as the baseline to avoid the curse of dimensionality. More will be discussed later in Section 3.4.2.

Based on the above discussion and limitations with the vector space model, we propose an approach to achieve blog site clustering leveraging on the “collective wisdom” that can be inferred from the blogger’s label entries. Often bloggers specify more than one predefined labels for their blog site. Such action on these blog sites help in establishing a link between these labels. These links are captured in a Label Relation Graph, an instance of which is depicted in Figure 9. For example, labels like Computers and Technology; Computers and Internet; Computers and Blogging were linked by the bloggers when categorizing their blogs. A naïve solution is to cluster the blogs simply by the labels. However, this naïve clustering approach may not be suitable for clustering the blogs. As illustrated in Figure 11, bloggers
often use *Personal* as the label descriptor for a variety of personal interests. Using this descriptor may not be sufficiently helpful in capturing the nuances of bloggers’ interests and we need to refine the label descriptor by identifying and aggregating the related labels. This is also referred as the topic irregularity problem where bloggers use the same label descriptor to define their blog which in fact contains blog posts of varied interests. That would require that different labels with similar themes be connected even when a blogger does not list his/her blog under all these labels. For example, a blog named *Words From Iraq* is labeled as *Iraq* and *Society*; another blog, *Iraq’s Inconvenient Truth*, is labeled as *Political*, and *News and Media*. Although they are related blogs, they are annotated by completely different labels. Connecting these labels through some sort of link or relation would make more sense. The number of blog sites that create the links between various labels is termed as *link strength*, which could be treated as the edge weights of the label relation graph. Using this label relation graph, different labels can be clustered or merged. We call this collective wisdom based approach, *WisColl*. We experiment with different thresholds for the link strength in Section 3.5.1. We visualize the label relation graph thus obtained using a visualization and analysis tool, Pajek\(^2\) described in Section 3.5.2.4. Once the label relation graph is computed we perform clustering using k-means and hierarchical clustering algorithms and compare their results. The results are presented in Section 3.5.2.5.

It can be observed from the blogosphere that people have varied interests and their interest in one topic is short-lived [2, 3]. This causes a drift not only in people’s interests but also as a whole in the blogosphere. However, people tend to categorize their blog posts using the same category descriptors they used to categorize their old blog posts, a phenomenon referred to as *path dependence* [4]. This is because either they are ignorant of the category structure (also because the taxonomy structure is highly dynamic and keeps evolving), or they are lazy to submit their blogs to more focused or refined categories. Desirably, *WisColl* is dynamic and adaptive to the current interests, since the tags of a blog site could change depending on what the blogger is blogging about. This results in dynamic as well as adaptive clustering. Hence, every time new blog sites appear, there will be new edges appearing into the label relation graph as well as changes to the link

strength whenever a blogger specifies different labels. In this scenario, the clustering results would change. Since the blogosphere provides more emphasis on the freshness of the content, the proposed clustering approach would also reflect similar dynamics. We demonstrate how WisColl captures this dynamic behavior in Section 3.5.2.1.

3.4.2. Baseline Approach

As a baseline clustering algorithm, we cluster the blogs using the blog post text and then find the predominant label for each cluster. The ‘Vector-Space’ model is used to encode the blogs with each blog being represented as a term frequency vector. Singular Value Decomposition (SVD) and the cosine similarity measure are then used to obtain the similarity matrix for clustering.

The vector-space representation for each blog is constructed to find the term frequencies in the blog posts of each blogger. For each blogger up to five blog posts are available and thus extracted and using these posts the ‘blog-term’ matrix is constructed. The following preprocessing steps are applied to the terms obtained by blog posts before constructing the matrix:

1. To trim the white spaces and punctuation marks, token scrubbing is performed on the blog post text;
2. All the terms are stemmed using the portal stemmer to obtain their morphological roots; and
3. The stop words are removed from the remaining list of terms.

After the preprocessing steps, using the resulting normalized terms the blog-term matrix $B$ ($t \times m$ matrix, with $t$ terms and $m$ bloggers/blog sites) is constructed. Latent semantic analysis [71] is performed on this matrix to obtain the lower dimensional concept space representation of each blog. This involved decomposing the blog-term matrix using SVD [71].

$$B = U\Sigma V^T$$

(3.3)

The blogger-term vectors were then projected into a reduced concept space by selecting the top $k$ eigenvalues represented by $V_k$ in eq. 3.4. The reconstructed blog-term matrix is of rank $k$.

$$B_k = U_k\Sigma_k V_k^T$$

(3.4)
In our experiments we achieved the best performance by selecting top 25 eigenvectors. In the resulting matrix $V_k$ each row corresponds to one blog and is represented by the vector $d_i = (d_{i1}, d_{i2}, ..., d_{ik})$ $1 \leq i \leq m$. The $m \times m$ similarity matrix $S$ was then constructed by finding ‘cosine similarity’ between the reduced concept space vectors corresponding to each pair of blogs. The $(i, j)^{th}$ element of $S$ gives the similarity between blogs $i$ and $j$ and is given by,

$$S(i, j) = (d_i \times d_j)/(|d_i| \times |d_j|)$$

(3.5)

Once we have the similarity matrix clusters of bloggers/blogs can be easily visualized. Clustering is achieved by setting a threshold $\tau$ for similarity. A link between two nodes is considered weak if the similarity is less than $\tau$. When the weak links are removed, clusters emerge. By identifying the predominant labels for the nodes in each cluster we identify the cluster labels. We visualize these clusters using Pajek in Section 3.5.2.4. We also performed k-means and hierarchical clustering using the computed similarity matrix $S$ and report the results in Section 3.5.2.5.

3.5. Experiments and Discussion

Figure 10 illustrates our approach to data collection and experimentation. We initiate our technique by drawing from a pool of blogs. To this set of blogs, we apply typical baseline clustering approach using the term vector space model. We also apply our collective wisdom based approach by identifying commonality among blogs according to their labels, which are mapped afterwards to the corresponding term vector in the blog
space that will allow the two sets to be compared. We use agglomerative hierarchical and k-means clustering techniques to compare the baseline and WisColl, and we use Pajek to help visualize our results.

Next we discuss how we collect the data and control the structural information granularity of the labels.

3.5.1. Experiment: Design and Methodology

We test WisColl with the sample data collected from a blog site directory available at BlogCatalog\(^3\), which will serve as template to further test other blog sources. BlogCatalog is a directory of blog sites that allows bloggers to label the blog sites under a given hierarchy. The directory structure of BlogCatalog is relatively shallow, with 33 nodes having no children. The maximum depth of the hierarchy is 3 and only two nodes have that depth. We experiment with varying granularity of structural information. Bloggers submit the blog sites to BlogCatalog. Each site is authored by a blogger. Each blog site contains some blog posts of which last 5 are displayed on the BlogCatalog.

To collect the BlogCatalog data we started with 4 bloggers from different labels as the starting points and crawl their social networks, recursively in a breadth-first fashion using the API available at BlogCatalog that allows access to their blogger friends. These bloggers belong to the most popular labels (i.e., having largest number of blog sites) at BlogCatalog and had the largest number of friends. For each blogger thus crawled (uniquely identified by their blogger IDs), we collect their blog site URL, blog site title, blog site labels, blog post tags, blog post snippets, blog post title, blog post permalink, and the blogger’s social network information, i.e., his/her friends. The data will be made available upon request.

Along with evaluating collective wisdom we also evaluate the structural properties of the labels, i.e., since the labels have a nested hierarchical structure, what level gives the best clustering results. To perform this we construct three different datasets:

1. **Top-level**: The hierarchical structure of the labels is known *a priori*. For this dataset we abstract the labels of all the blog sites to their parent level labels. For example, Family is a child of Personal.

   So all the blog sites that are labeled Family are relabeled as Personal, thus abstracting their labels

\(^3\)http://www.blogcatalog.com/
to the parent level. Note that the maximum depth of this hierarchical structure of labels is 3 and we abstract the labels to the highest parent level label. There are in total 56 labels.

2. All-label: This variant of the dataset does not abstract the label information. It considers the full hierarchical structure of the labels. There are in total 110 labels at all the levels of the hierarchy.

3. One node-split: According to the distribution of blog sites in various top level labels, illustrated in Figure 11, Personal has the largest number of blog sites\(^4\). Hence, we split Personal into its child labels, to reduce the skewed distribution.

Note that the approach presented here would work for any blog dataset with user specified metadata like labels or tags.

3.5.2. Results and Analysis

In this section we present the experiments results and analysis. The experiments are designed to evaluate the following:

\(^4\)For the sake of space constraint and the analysis presented here, we limit the labels in this chart that have at least 1000 blog sites.
• What granularity of label hierarchical structural information generates best clustering? For this, we study the clustering results for the three variants of the dataset mentioned in Section 3.5.1.

• Which one of the two clustering approaches: WisColl, that leverages collective wisdom, or the baseline approach, performs best per k-means, hierarchical, or visualization?

Before we delve into the parameter tuning of various clustering methods, we study dynamics of collective wisdom and the effect of different thresholds for link strength on WisColl. Based on the following study, we fix the threshold for link strength for rest of the experiments. For threshold experiments we use the All-label variant of the dataset.

3.5.2.1. Dynamics of Collective Wisdom

In order to study the dynamic characteristics of collective wisdom, we look at the changes resulted from data containing different number of bloggers i.e., 7023, 10642, 11947, and 12308 bloggers. Increase in data causes the change of collective wisdom which can be captured in category relation graphs (CRGs). This suggests how the tagging behaviors of various bloggers change over time. To study the dynamics of the collective wisdom we visualize the CRGs obtained from the data containing 10,462 bloggers and 12,308 bloggers, depicted in Figure 12 as 10k and 12k, respectively.

Comparing the CRGs of data with 10k and 12k bloggers in Figure 12, we observe that “Computers Science Technology” and “Internet Business Blogging Blog Resource” are merged in 12k. This is due to a new link that emerges between “Technology” and “Internet”. Similarly we notice a new link between “Society” and “News & Media” categories. “Political News & Media Society” in 10k expands to “Political News & Media Society Humor Writing” in 12k. This transformation connects categories, “Society” with “Humor” and “Humor” with “Writing”. Another instance of expansion is “Coaching Education & Training” in 10k to “Career & Jobs Coaching Education & Training” in 12k. This creates a new link between “Career & Jobs” and “Coaching”. We also notice changes to existing links like “Travel Vacation” which transforms to “Photo Blog Travel Vacation”. This establishes a new link between “Photo Blog” and “Travel” categories. Some of the category links remain unchanged in 10k and 12k like, “Investing Finance Meme”, “Shopping
Fig. 12. CRGs for Different Datasets Containing 10,642 Bloggers (10k) and 12,308 Bloggers (12k).


3.5.2.2. Link Strength

We experiment with different thresholds for the All-label variant of the dataset. These values have been captured in Table 11. The table shows the range of values, and their distribution, for all of the 456 link strength values, for all of the 110 All-label nodes.

Each link-strength, or threshold selection, results in a network re-structuring and reduction. The re-structuring occurs as a result of removing those links whose strength values are below the threshold value,
TABLE 11

Link Strength Statistics for ALL Labels.

<table>
<thead>
<tr>
<th>Line Value</th>
<th>Frequency</th>
<th>Freq%</th>
<th>CumFreq</th>
<th>CumFreq%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>320</td>
<td>70.1754</td>
<td>320</td>
<td>70.1754</td>
</tr>
<tr>
<td>2</td>
<td>62</td>
<td>13.5566</td>
<td>382</td>
<td>83.7719</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>7.2368</td>
<td>415</td>
<td>91.0088</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>3.2895</td>
<td>430</td>
<td>94.2982</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>1.5351</td>
<td>437</td>
<td>95.8333</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>1.5351</td>
<td>444</td>
<td>97.3684</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>1.3158</td>
<td>450</td>
<td>98.6942</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>450</td>
<td>98.6942</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.4286</td>
<td>452</td>
<td>99.1228</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.2193</td>
<td>453</td>
<td>99.3421</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>0.6579</td>
<td>456</td>
<td>100.0000</td>
</tr>
</tbody>
</table>

Totals 456 100.0000

which can cause a cluster to transform or break into smaller clusters. The reduction occurs by removing those nodes that are no longer connected to any other node as a result of their links being removed for falling below the threshold.

Different threshold values will result in different cluster formations, generation and deletion. Figure 13 illustrates the dependency of threshold value, and the distribution of clusters by cluster size and frequency, for the range of All-label threshold values, corresponding to those values in Table 11. This figure shows that after selecting a threshold value of 2, this will result in a network structure consisting of two clusters, one of size 2, as containing two nodes, and the other of size 72, as being formed by seventy-two nodes. For the case that the threshold is set to 3, this results in three clusters: two of them of size 2, one of size 48. In looking at the data from Figure 13, the plot shows that as we increase the threshold, the number of smaller size clusters increases, and the larger size clusters decreases, up to a certain transition point, after which also the smaller cluster size decreases, as the total number of remaining nodes in the network decreases. For the All-label case, this transition point is centered around a threshold value of 5. This is better illustrated by Figure 14 which shows cluster frequency variation and transition, with increasing threshold value, for the smaller size clusters.

The results illustrated by Figure 13 and Figure 14 represent the 90 to 99 percent cumulative-frequency for the link strengths selected from Table 11. We present the cluster visualization results for the All-label link strength range, for representative threshold values of 3, 5 and 7 in Figure 15 through Figure 17. Contour lines

\(^5\)Pajek was used to create the visualizations.
Fig. 13. All Label Cluster Frequency Count (Y-axis) by Cluster Size (X-axis) per Corresponding Threshold Value.

Fig. 14. All Label Cluster Histogram for Small Size Clusters (Size 10 or Less) per Corresponding Threshold Value for Figure 13.
In the figures, link strength is denoted by the values on the edges. Names of the nodes depict the labels assigned by the bloggers to the blog sites. Here a node represents all the blog sites that are labeled as the label of the node. A cluster of labels would represent a cluster of all the blog sites that are labeled with one of these labels. Some nodes like Internet>Web Design depict the hierarchical structure of labels. Here the blog sites are labeled Web Design which is a child of Internet. We present detailed statistics for clustering results for all the threshold values in Table 12 for comparison. For threshold \( \geq 3 \), total coverage is highest but we have a single large cluster and two very small clusters depicted by the cluster coverages. Similar is the case for threshold \( \geq 4, 7, \) and \( 10 \). This indicates highly unbalanced clusters are achieved at other thresholds as compared with threshold \( \geq 5 \). This value coincides with our previous notion that this is the transition point as shown in Figure 13 and Figure 14. Hence we set threshold=5 for rest of the experiments. More results and analysis are reported in [1].

### 3.5.2.3. Label Hierarchy

Next we study the effect of structural information of labels on WisColl. For this experiment we consider all the three variants of the dataset, i.e., Top-level, All-label, and One node-split.

A sample of the clustering results for Top-level is shown in Figure 18 (see complete progression in Appendix A). This version of the dataset is the worst among all the three variants. The cluster size is highly unbalanced. There is a single cluster to which all the nodes belong. This is the case for all threshold levels. No contour lines are drawn since no additional clusters are generated at any level. A sample clustering results for...
Fig. 15. WisColl Results for Link Strength $\geq 3$ for All-Label Dataset.

Fig. 16. WisColl Results for Link Strength $\geq 5$ for All-Label Dataset.
for One node-split is shown in Figure 19 (see complete progression in Appendix A). The results are a slightly better than Top-level but still the clustering quality is poor with unbalanced cluster size, as only in a couple of instances, more than a single cluster is present. As such, best clustering results are obtained with All-label as shown in Figures 15 through 17.

We compare the statistics of clusters obtained from WisColl for different versions of datasets in Table 13. Although the total coverage is maximum for Top-level label structure, there is only one cluster that connects all the labels. This results in 100% coverage for the 1st cluster. So there is no search space reduction at all. Every time a query comes the results are returned from the 1st and only cluster and since it contains all the labels, whole dataset needs to be searched. Similarly, results for One node-split show that the cluster size is highly unbalanced. There are only 3 clusters with the 1st cluster having majority of coverage (=76.44%) and the difference between 1st and 2nd cluster is very small. This largely affects the search space reduction. Results for All-labels has the lowest coverage but the cluster sizes are not as unbalanced. Moreover, the difference between the coverage for 1st and 2nd clusters is larger than One node-split. This leads to better search space reduction. This shows that leveraging the complete structure of collective wisdom gives best
Table 13

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of clusters</th>
<th>Highest degree</th>
<th>Lowest degree</th>
<th>Largest cluster size</th>
<th>Smallest cluster size</th>
<th>Coverage Total %</th>
<th>Coverage 1st cluster %</th>
<th>Coverage 2nd cluster %</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-categories, &gt;= 5</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>54.76</td>
<td>42.3</td>
<td>6.375</td>
</tr>
<tr>
<td>Top-level</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>22</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>One node-split</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>21</td>
<td>2</td>
<td>82.87</td>
<td>76.44</td>
<td>3.88</td>
</tr>
</tbody>
</table>

Fig. 18. WisColl Results for Link Strength $\geq$ 3 for Top-Level Label Dataset.

results as compared to exploiting a part of it. This proves that the more collective wisdom is available the better it is.

3.5.2.4. Visualizations - Pajek

In this section we visualize the Label Relation Graph using Pajek. Pajek is a Windows based local-PC visualization software tool capable of analyzing very large networks up to a 10,000,000 vertices. Our implementation of the process consists of first transforming the data set into a collection of connected blog term pairs, for the baseline, and of connected label pairs for our WisColl approach. Each of these connected node pairs represents an existing link between the corresponding node pair. A ‘text-to-Pajek’ utility listed in Pajek’s website transforms the paired data into an undirected Pajek style network for each data set. The baseline
Fig. 19. WisColl Results for Link Strength ≥ 3 for Personal Label Dataset.

approach generated 842 vertex nodes on over 700,000 linked term node-pairs. For the label data, a total of 110 vertex nodes were generated for the almost 600 linked label node-pairs for the ALL type labels. Node pairs that are frequently linked have multiple representations. The networks were simplified by replacing those node pairs instances with multiple representation, with a single representative link as the link-strength between the nodes for the node pair. The simplification resulted in 346,921 edges for the baseline network, and 456 for the ALL label network. Once the networks were simplified, a feature in Pajek that allows to transform a network into a new network by removing lines bellow a specified threshold was utilized to perform our analysis.

Hence, using these features just described, we visualize the results of WisColl algorithm (e.g. Figure 16) and baseline algorithm to study the advantages of collective wisdom. Results for the baseline clustering were generated in an analogous fashion as to what was collected for the WisColl approach. As such, we first studied the Baseline’s link strength range of values. These values are shown in Table 14. The table shows the range of values, and their distribution, for all of the 346,921 edge link strength values, corresponding to 842 Baseline nodes.
Figure 20 illustrates the dependency of threshold value, and the distribution of edge-connected clusters by size and frequency, for the Baseline range of threshold values, corresponding to those values in Table 14. This figure shows a similar behavior as was observed for Figure 13, that as we increase the threshold, the number of smaller size clusters increases, and the larger size clusters decreases, up to a certain transition point, after which the total number of remaining nodes in the network decreases. For the Baseline case, this transition point is centered around a threshold value of 0.75 to 0.80. This is better illustrated by Figure 21 which shows cluster size distribution and the transition, with increasing threshold value, for the smaller size clusters.

Figure 22 presents the edge-connected cluster visualization results for the Baseline link strength range, for representative threshold value of 0.80 (See Appendix for a progression of threshold values from 0.80 to 0.95 in Figure 33 through Figure 38). Contour lines were added to highlight the edged-connected/linked clusters. The node placement throughout the figures was maintained in order to facilitate visualizing the clusters creation and transitions.

Best visualization of cluster type clustering forming results for baseline approach were achieved with the threshold \( \tau = 0.9 \). Here nodes represent the blog sites/posts. For easier comparison we also display the labels of their blog sites besides their name. For example, a node label like, \texttt{emom=Small Business:Moms}, tells us that the blogger \texttt{emom} has a blog site with labels \texttt{Small Business} and \texttt{Moms}. Cluster quality for both the approaches could be compared by looking at the labels of the cluster members. However, we do not use the label information while clustering in baseline approach. We report the differences between the two approaches based on the results as follows:

1. There are too many clusters obtained from baseline approach and many have very small size (most of them are 2-member clusters). However, this is not the case with WisColl.

2. As a result of too many small sized clusters, clusters are too focussed. This affects the insertions of new blog site later on. Cluster configurations are highly unstable in such a focussed clustering. For example \texttt{cozimono = Music:Rock:Pop} and \texttt{billiam = Music:Rock:Pop} are clustered together. This group is highly focussed and if a new blog about Music comes is added then it won’t be assigned to this group.

3. Deeper analysis shows that some clusters obtained from baseline clustering, have members whose
TABLE 14
Baseline Link Strength Statistics.

<table>
<thead>
<tr>
<th>Line Value Range</th>
<th>Frequency</th>
<th>Freq%</th>
<th>CumFreq</th>
<th>CumFreq%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000 - 0.0417</td>
<td>53288</td>
<td>15.3603</td>
<td>53288</td>
<td>15.3603</td>
</tr>
<tr>
<td>0.0417 - 0.0833</td>
<td>51239</td>
<td>14.7964</td>
<td>58417</td>
<td>16.7567</td>
</tr>
<tr>
<td>0.0833 - 0.1250</td>
<td>57045</td>
<td>16.9373</td>
<td>115462</td>
<td>33.6940</td>
</tr>
<tr>
<td>0.1250 - 0.1667</td>
<td>40221</td>
<td>11.5361</td>
<td>155683</td>
<td>47.2302</td>
</tr>
<tr>
<td>0.1667 - 0.2083</td>
<td>33588</td>
<td>9.8817</td>
<td>289271</td>
<td>87.1119</td>
</tr>
<tr>
<td>0.2083 - 0.2500</td>
<td>27705</td>
<td>7.9860</td>
<td>316976</td>
<td>95.0979</td>
</tr>
<tr>
<td>0.2500 - 0.2917</td>
<td>22250</td>
<td>6.3136</td>
<td>539526</td>
<td>160.4115</td>
</tr>
<tr>
<td>0.2917 - 0.3333</td>
<td>18505</td>
<td>5.3341</td>
<td>724581</td>
<td>215.7456</td>
</tr>
<tr>
<td>0.3333 - 0.3750</td>
<td>16759</td>
<td>4.8543</td>
<td>892140</td>
<td>260.5999</td>
</tr>
<tr>
<td>0.3750 - 0.4167</td>
<td>11556</td>
<td>3.3863</td>
<td>100770</td>
<td>304.4862</td>
</tr>
<tr>
<td>0.4167 - 0.4583</td>
<td>8581</td>
<td>2.4735</td>
<td>109351</td>
<td>330.2197</td>
</tr>
<tr>
<td>0.4583 - 0.5000</td>
<td>6052</td>
<td>1.7445</td>
<td>115403</td>
<td>347.6642</td>
</tr>
<tr>
<td>0.5000 - 0.5417</td>
<td>4290</td>
<td>1.2366</td>
<td>119693</td>
<td>359.0008</td>
</tr>
<tr>
<td>0.5417 - 0.5833</td>
<td>2623</td>
<td>0.8117</td>
<td>122316</td>
<td>369.8125</td>
</tr>
<tr>
<td>0.5833 - 0.6250</td>
<td>2382</td>
<td>0.6866</td>
<td>124698</td>
<td>376.6787</td>
</tr>
<tr>
<td>0.6250 - 0.6667</td>
<td>1364</td>
<td>0.3932</td>
<td>138362</td>
<td>380.6119</td>
</tr>
<tr>
<td>0.6667 - 0.7083</td>
<td>851</td>
<td>0.2453</td>
<td>146913</td>
<td>383.0672</td>
</tr>
<tr>
<td>0.7083 - 0.7500</td>
<td>488</td>
<td>0.1407</td>
<td>151801</td>
<td>384.4679</td>
</tr>
<tr>
<td>0.7500 - 0.7917</td>
<td>279</td>
<td>0.0804</td>
<td>154590</td>
<td>384.5483</td>
</tr>
<tr>
<td>0.7917 - 0.8333</td>
<td>167</td>
<td>0.0481</td>
<td>156261</td>
<td>384.5964</td>
</tr>
<tr>
<td>0.8333 - 0.8750</td>
<td>79</td>
<td>0.0228</td>
<td>157050</td>
<td>384.6192</td>
</tr>
<tr>
<td>0.8750 - 0.9167</td>
<td>33</td>
<td>0.0095</td>
<td>157083</td>
<td>384.6287</td>
</tr>
<tr>
<td>0.9167 - 1.0000</td>
<td>23</td>
<td>0.0066</td>
<td>157106</td>
<td>384.6353</td>
</tr>
<tr>
<td>1.0000 - 1.0000</td>
<td>15</td>
<td>0.0043</td>
<td>157121</td>
<td>384.6396</td>
</tr>
</tbody>
</table>

Totals: 346921, 100.0000

Fig. 20. Baseline Cluster Frequency by Cluster Size per Corresponding Threshold Value.
Fig. 21. Baseline Cluster Histogram for Small Size Clusters (Size 10 or Less) per Corresponding Threshold Value for Figure 20.

Fig. 22. Results for Link Strength $\geq 0.80$ for Baseline Dataset.
blog site labels are not semantically related. For example, bluemonkey jammies = Humor:Personal and emperoranton = SEO: Marketing are clustered together. However, the labels are totally different and are not at all semantically related. There are several such clusters obtained from baseline clustering approach. This shows that baseline clustering does not give semantically coherent clusters. This is because vector space clustering using blog posts are susceptible to text noise, and blogs are usually noisy. Also blogs are dynamic in nature with the blogger occasionally posting about different topics. Such off topic posts affect the clustering using vector space methods. However WisColl gives high-quality, semantically coherent clusters. For example, clusters having members like Internet> Web Design and Internet> Web Development; Food & Drink and Food & Drink> Recipes; Internet, Computers, Technology, and Technology> Gadgets etc. are semantically related.

4. Several clusters obtained from baseline approach have members that have exactly same labels. For example, the cluster with bloggers emom and geraelindsey have the exact same labels, i.e., Small Business and Moms. This does not help in identifying relationships between blog sites that have different themes. Clustering blog sites that have different yet related theme/topics are more helpful. WisColl generates clusters that have blog sites with topics like, Technology, Computers, Internet, and Technology> Gadgets. Such a cluster serves a better purpose for various applications like search, organization of information, etc.

3.5.2.5. k-Means vs. Hierarchical Results

A hierarchical clustering was generated for the 27 labels identified by WisColl for link strength 5 for the network illustrated in Figure 16. The resulting hierarchical clustering, built using using Pajek’s Hierarchical Ward method feature, is illustrated in Figure 23. The nodes in the figure are represented by an index identifier, for which the corresponding labels are included in Table 15. From the hierarchical diagram, we identify/select 7 major clusters, which are tabulated in the Table 15 by label node and cluster ID. We then generate a k-means clustering for k=7 for the 27 labels to analyze how well k-means and hierarchical clustering compared between the two methods.

In order to compare how WisColl fares with regards to the baseline blogger clustering, we generated
### TABLE 15

Hierarchical Clustering Table with Clustering Assignment for Link Strength $\geq 5$ for All-Label.

<table>
<thead>
<tr>
<th>Index</th>
<th>Clus ID</th>
<th>Cluster Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Philosophy</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Arts &amp; Entertainment</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Society</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Personal Relationships</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>Photo Blog</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>Photo Blog Photograph</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>Humor</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>Writing</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>Internet&gt;Web Design</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>Internet&gt;Web Development</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>Personal&gt;Family</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>Personal&gt;Parenting</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>News &amp; Media</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>Political</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>Food &amp; Drink</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>Food &amp; Drink&gt;Recipes</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>Business&gt;Small Business</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>Internet&gt;SEO</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>Business&gt;Marketing</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>Blog Resources</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>Personal&gt;Development and Growth</td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>Blogging</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td>Personal</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
<td>Computers</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>Internet</td>
</tr>
<tr>
<td>26</td>
<td>7</td>
<td>Technology&gt;Gadgets</td>
</tr>
<tr>
<td>27</td>
<td>7</td>
<td>Technology</td>
</tr>
</tbody>
</table>
Fig. 23. Hierarchical Clustering for Link Strength $\geq 5$ for All-Label Dataset Value for Indexes per Table 15.
Fig. 24. k-Means k-Analysis for Baseline Dataset.

similar hierarchical and k-means clusters for the baseline’s blogger space, which is as follows. In order to be able to compare between the results obtained from the label space, into the blogger space, we “map” each vector in the label space to its corresponding matching vector(s) in the blogger space. This was accomplished by associating the blogs to each of the labels in the label cluster, given the blogger had used that label in his/her blog. This transforms the clusters in the label space to an equivalent cluster made out of blogs in the blogger space. The mapping was generated for both hierarchical and k-means results. In more than one occasion, the same blogger had used more than one label. So if the between-cluster distance is computed using single-link, then many of the between-cluster distances would be 0. This will skew the distribution of the cluster distances. Therefore we calculate the between-cluster and within-cluster distances using the average-link. The mapping generates same number of clusters in the blogger space as the number of clusters obtained in label space. Since we selected 7 clusters for hierarchical clustering in the label space, after mapping we would get 7 clusters of blogs.

Next we try to observe the best value for k in k-means. For the analysis and comparison, we followed the premise discussed in Section 3.4.1, where we presented the min and max criteria based on Equation 3.1, which minimizes the within-cluster distance (i.e., more cohesive clusters) and Equation 3.2 that maximizes the between-cluster distance (i.e., well separated clusters). Figure 24 shows these results for the blog clusters. To assess the value of k, we take advantage from the results obtained from Figure 20, which suggests that
based on link strength analysis, the highest cluster count is for link strength 0.75, which is slightly above 60. Hence, we use 60 as the maximum value of k. We perform clustering of blogs for k = 5 to 60. As shown in the figure 24, we observe a lot of fluctuations in the range k = 5 to k = 12. Upon deeper analysis it shows that we obtain lowest within-cluster distance for k = 7. This results is in accordance to the hierarchical clustering as it also generated 7 clusters. Note that for k > 25 even though it looks that the clustering is better (large between cluster distance and small within cluster distance) but in reality we obtain a lot of 2 member clusters for higher value of k, just like in Figure 22.

The results comparing hierarchical and k-means clusters are summarized in Table 16. The table shows three categories: a) “WisColl - Label Space”, evaluates the clustering generated using label relation graph in the label space; b) In “WisColl - Blogger Space” we transformed the label space clusters to blogger space as described previously using Baseline approach; and c) “Baseline - Blogger Space” refers to the clustering generated by clustering the bloggers. It is in the “b” case, when we project the labels into the bloggers space so that we can make a fair comparison between the “b” and “c” results.

From the results in Table 16, we make two observations:

1. WisColl (category type ‘b’) performs better than Baseline (category type ‘c’) in both clustering algorithms: k-means and hierarchical. Although we achieve lower ‘within-cluster’ distance for Baseline approach than WisColl, the ‘between-cluster’ distance is higher for baseline as compared to WisColl. This implies that though we have cohesive/tight clusters for Baseline approach, the clusters are not as well separated since their between clusters results indicate that the Baseline clusters are significantly closer than those generated using WisColl. In addition, the variance in the between-cluster distance of Baseline approach is much higher than that of WisColl. Note that this comparison is made in blogger space for a fair comparison.

2. k-means performs better than hierarchical for WisColl in label space, since k-means has lower within-cluster and higher between-cluster distance. However, in blogger space for WisColl, hierarchical clus-
TABLE 16
Comparing WisColl with Baseline approach Using k-Means and Hierarchical Clustering.

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Within</th>
<th>Between</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) WisColl-Label space</td>
<td>k-means</td>
<td>0.0615 ± 0.1643</td>
<td>0.2860 ± 0.0536</td>
</tr>
<tr>
<td></td>
<td>hierarchical</td>
<td>0.0857 ± 0.1672</td>
<td>0.2761 ± 0.0571</td>
</tr>
<tr>
<td>(b) WisColl-Blogger space</td>
<td>k-means</td>
<td>0.0844 ± 0.0995</td>
<td>0.7090 ± 0.0143</td>
</tr>
<tr>
<td></td>
<td>hierarchical</td>
<td>0.0849 ± 0.0943</td>
<td>0.8118 ± 0.0047</td>
</tr>
<tr>
<td>(c) Baseline-Blogger space</td>
<td>k-means</td>
<td>0.0363 ± 0.1264</td>
<td>0.2194 ± 0.1301</td>
</tr>
<tr>
<td></td>
<td>hierarchical</td>
<td>0.0890 ± 0.1186</td>
<td>0.3644 ± 0.0903</td>
</tr>
</tbody>
</table>

Clustering performs a little better than k-means clustering with higher between-cluster distance and comparable within-cluster distance.

To summarize the results, we can clearly see that WisColl, when compared with baseline after mapping the labels to blogger space, has a better “separation” and less variation.

3.6. Summary

Clustering blog sites is a challenging task with many real-world applications. Classic clustering methods are not designed to take advantage of some of the characteristics of the blogosphere, like user-specified metadata. In this work, we propose to cluster blog sites by employing collective wisdom. We investigate various types of information available in a blog catalog site, like user specified tags and labels. We propose to leverage collective wisdom to generate label relation graph that represents similarity between labels. We call this collective wisdom based approach, WisColl, which is further used to perform clustering using conventional clustering approaches such as k-means and hierarchical clustering algorithms. This approach differs from the classic algorithms and uses the collective wisdom in the form of label information. We compare WisColl with a representative SVD-based approach that does not use collective wisdom, to discern their differences. We argue that since WisColl mainly relies on the label information by the bloggers, it is nimble to automatically adapt to the dynamic changes of the label information. Furthermore, in search of an effective clustering algorithm with collective wisdom, we evaluate different values of link strengths and various levels of label hierarchies, present results statistically and visually, and summarize findings that deepen our understanding of this new type of clustering.
The work of WisColl is a proof-of-concept project in using various forms of collective wisdom in the Social Web. It is just a beginning. A lot of work remains to be explored. In the short term, our future work includes the integrative use of multiple types of information such as labels, tags, and posts in clustering across different social media websites such as del.icio.us and Flickr. We would like to explore the possibility of adaptive thresholding for clusters with different link strength value distribution. Future tests will include an expanded base beyond the original data collected from BlogCatalog, and analyze how well our method conforms with the expanded data set. On the long term, subsequent to this work, we will explore how effectively bloggers belonging to the Long Tail are aggregated and represented as part of the critical mass. This corresponds to understanding how much we can close the gap with respect to the Short Head, which would result in a more effective search.

Fig. 25. WisColl Results for Link Strength $\geq 3$ for Top-Level Label Dataset.
Fig. 26. WisColl Results for Link Strength $\geq 5$ for Top-Level Label Dataset.

Fig. 27. WisColl Results for Link Strength $\geq 7$ for Top-Level Label Dataset.
Fig. 28. WisColl Results for Link Strength $\geq 9$ for Top-Level Label Dataset.

Fig. 29. WisColl Results for Link Strength $\geq 3$ for Personal Label Dataset.
Fig. 30. WisColl Results for Link Strength $\geq 5$ for Personal Label Dataset.

Fig. 31. WisColl Results for Link Strength $\geq 7$ for Personal Label Dataset.
Fig. 32. WisColl Results for Link Strength $\geq 9$ for Personal Label Dataset.

Fig. 33. Results for Link Strength $\geq 0.80$ for Baseline Dataset.
Fig. 34. Results for Link Strength $\geq 0.85$ for Baseline Dataset.

Fig. 35. Results for Link Strength $\geq 0.88$ for Baseline Dataset.
Fig. 36. Results for Link Strength $\geq 0.90$ for Baseline Dataset.

Fig. 37. Results for Link Strength $\geq 0.92$ for Baseline Dataset.
Fig. 38. Results for Link Strength $\geq 0.95$ for Baseline Dataset.
4. DISCOVERING FAMILIAR STRANGERS IN BLOGOSPHERE

4.1. Introduction

Familiar strangers as defined by Stanley Milgram [7] in physical world are those individuals who do not know each other but share some common attributes like interests, occupation, location etc. For instance, people taking the same train daily find familiar faces but do not know each other. Analogous to physical world, it is equally interesting and challenging to define and study the existence of familiar strangers in virtual or online world. Social networks represent a complex set of human relations through interactions expressed via a spectrum of social media websites like blogs, online friendship networks, wikis, media sharing websites, social tagging websites etc. In an online world, familiar strangers could be defined as those individuals who are not friends with each other, i.e., they are not in each other’s social network, but they share some common set of attributes like hobbies, community affiliations, workplace, location, etc. A more formal definition is given later.

Identifying familiar strangers has profound applications in online social networks. Since the online social networks are shown to have long tail distribution, i.e., most of the members have very few contacts and very few members have a large number of contacts, which means that most of these members do not know each other. Although many of them could have a lot in common but due to the long tail distribution it is quite likely that they may not know each other. Aggregating such familiar strangers could form a critical mass such that (1) the understanding of one member gives us a sensible and representative glimpse to others, (2) more data about familiar members can be collected for better customization and services (e.g., personalization and recommendation), (3) the nuances among them suggest new business opportunities, and (4) knowledge about them can facilitate predictive modeling and trend analysis in new product/market development. Connecting them to form a critical mass can potentially expand their social network, i.e., job searching, special interest group formation. Aggregating familiar strangers can encourage participation due to the crowd effect [72]. People usually trust those with similar interests. Knowledge transfer or information flow among friends and acquaintances becomes smoother and more receptive.

Identifying familiar strangers in online social networks is interesting and involves several key challenges. Individuals have only local view, i.e., individuals know their contacts but may not know their contacts’ con-
tacts and so on. Searching for all the contacts of a node, his contacts’ contacts and so on, to identify familiar strangers incurs an exponential cost. Each individual is associated with some content or attributes. The challenge lies in intelligently putting that information to the benefit of searching familiar strangers. Evaluation and validation of the proposed approaches is a big issue due to the absence of an established ground truth.

The key contributions of this work are:

1. We studied “familiar stranger” concept proposed by Stanley Milgram in online social networks.
2. We formulate the problem of identifying familiar strangers for a node given the egocentric view.
3. Upon removing the constraint of egocentric view, i.e., if the nodes have the access to global information, we show that the problem of identifying familiar strangers can be reduced to a well-known \(np\)-complete Steiner tree problem and study its \(2\)-approximation solution to estimate the lower bound on the search space.
4. The Steiner tree solution that utilizes the global information is used to generate the ground truth.
5. We study and propose a Social Identity theory based approach to search for familiar strangers given the egocentric view. Towards this goal we first construct the social identity of the individuals and then leverage their social identity to search for their familiar strangers.
6. We propose other alternative approaches that are also constrained by the local information and compare with the social identity based approach.
7. We performed extensive experiments on a real world blogger social network dataset, BlogCatalog and citation network dataset, DBLP to show that the proposed social identity based approach outperforms the other alternative approaches and is quite close to the Steiner tree based search approach in terms of search space complexity, which utilizes the global information.

The following chapter is organized as follows: We define and formulate the problem of identifying familiar strangers constrained by local information in Section 4.2. Then we present a background of social identity
4.2. Problem Formulation

Here we define familiar strangers and formulate the problem of searching them using local information. Given a social network $G$ where $V$ is the set of vertices (nodes) or the members of the social network. The nodes are associated with an attribute. The attribute can take one or more values from a domain $D = \{a_1, a_2, \ldots, a_l\}$.

We call this the attribute-value set of a node and is denoted by $A_u$ for a node $u$ ($u \in V$). Each node $u$ has a local view of the network (also known as an egocentric view [73]), that means the node only knows its adjacent nodes denoted by $C_u = \{m_1, m_2, \ldots, m_y \mid \text{edge}(u, m_p) \neq 0, 1 \leq p \leq y\}$, also known as $u$’s contacts. Here $\text{edge}(c, d) \neq 0$ denotes an edge between nodes $c$ and $d$. This is similar to a scenario where one knows his/her friends but doesn’t know his/her friends’ friends and so on. In order to define familiar strangers of $u$, it is essential to define the notion of similarity.
Definition 1 (Similarity) Nodes \( u \) and \( v \) are similar iff \( A_v \cap A_u \neq \emptyset \), where \( \gamma \) is a goal described as \( \gamma \subseteq A_u \).

Definition 2 (Familiar Strangers) Given \( u \) and \( \gamma \), \( T_u \) is the set of familiar strangers of \( u \) iff (1) for all the nodes \( v \in T_u \), \( edge(u,v) = 0 \) i.e., all the nodes \( v \) are non-adjacent to \( u \) - stranger\(^1\) and (2) all the nodes \( v \) are similar to \( u \) with respect to \( \gamma \) as defined above - familiar.

Example 1: The problem of searching for familiar strangers given a node \( u \) can be illustrated in Figure 39 where a blogger social network is presented in the left, snippet of which is presented in the middle. Here the attribute \( A \) is “Interest” and \( D \) is the domain for the values of “Interest”. \( C_u \) represents the contacts of \( u \) and \( A_u \) represents the attribute-value set of \( u \). \( A_{v_1}, A_{v_2}, A_{v_3}, A_{v_4} \) represent the attribute-value sets of \( v_1, v_2, v_3, \) and \( v_4 \) respectively. We need to find \( T_u \), familiar strangers of \( u \) for the goal \( \gamma \) (“Sports”) defined by the combination of “Exercise” and “Recreation”.

The challenge lies in searching for familiar strangers efficiently, i.e., in minimum number of edge traversals with local information. To compute the lower bound on the search space for finding the familiar strangers, consider the centralized version of the problem, in which the node \( u \) has global or whole view of the network and the objective is to find the smallest set of edges that will connect all the nodes in \( T_u \) starting at node \( u \). This centralized version of the familiar strangers problem corresponds to the Steiner tree problem. Given a subset of nodes \( V' \subset V \) in a graph \( G = (V,E) \), the Steiner tree \( (T) \) spans the node set \( V' \) with least number of edges. The node set \( V' \) is referred to as the required nodes or terminal nodes and the set of nodes in \( V \setminus V' \) is referred to as the optional nodes or Steiner vertices. It may be noted that tree \( T \) contains all the nodes in set \( V' \) and zero or more nodes in set \( V \setminus V' \). The Steiner tree in a social network that spans the node \( u \) and the familiar strangers \( T_u \) provides the least number of edges that need to be traversed to find all the familiar strangers of \( u \) and thus provides a lower bound on the search space of the familiar strangers problem.

The problem of finding the Steiner tree is known to be NP-complete [75]. We provide an Integer Linear Programming (ILP) formulation to solve the Steiner tree problem optimally. Given the undirected social

\(^1\)This definition of stranger nodes is borrowed from the famous concept of weak ties [74].
network graph $G = (V, E)$, we first construct the corresponding directed graph $H = (V, F)$, in which two directed edges $\{(v_i, v_j), (v_j, v_i)\} \in F$ for each undirected edge $(v_i, v_j) \in E$. Let the number of required nodes be denoted by $n$, i.e., $|V'| = n$ and let an arbitrary vertex say, the node $u \in V'$ be designated as the root node. The ILP views the directed graph $H$ as a flow graph, in which $(n - 1)$ units of flow are routed from the root $u$ towards the nodes in $V' \setminus \{u\}$ through minimum number of edges. Each node in $V' \setminus \{u\}$ consumes exactly one unit of flow. The edges of graph $H$ through which a positive (unit) flow exists form the minimum-edge arborescence\(^2\) in $H$ spanning the vertices $V'$. The undirected edges in graph $G$ corresponding to the arborescence edges forms the required Steiner tree in $G$.

Let indicator variables $x_{v_i v_j} = 1$, if edge $(v_i, v_j)$ belongs to the required minimum-edge arborescence $T$ in $H$, otherwise, $x_{v_i v_j} = 0$. Let variables $f_{v_i v_j} \geq 0$ represent non-negative flow on the edges. The variables $x_{v_i v_j}$ and $f_{v_i v_j}$ are defined for all edges $(v_i, v_j) \in F$. The objective is to minimize the number of edges in the arborescence in $H$,

$$\text{Minimize } \sum_{(v_i, v_j) \in F} x_{v_i v_j}$$

- There are exactly $(n - 1)$ units of flow emanating out of the root node $u$ and 0 units of flow going into it. That is,

$$\sum_{(u, v_j) \in F} f_{uv_j} = n - 1, \quad \sum_{(v_j, u) \in F} f_{v_j u} = 0$$

- Every other required node, i.e., $v_i \in V' \setminus \{u\}$ consumes 1 unit of flow. That is,

$$\forall v_i \in V' \setminus \{u\}, \quad \sum_{(v_j, v_i) \in F} f_{v_j v_i} - \sum_{(v_i, v_j) \in F} f_{v_i v_j} = 1$$

- A positive flow exists on an edge, iff the edge is selected in the arborescence which is ensured by:

$$\forall (v_i, v_j) \in F, \quad f_{v_i v_j} \leq (n - 1) x_{v_i v_j}$$

Because solving ILP in general takes exponential time, we employ a 2-Approximation algorithm based on Minimum Spanning Tree approach [75] for computing Steiner trees. The 2-Approximation algorithm produces a solution that is guaranteed to be within 2 times the optimal solution in terms of the edge traversals.

\(^2\)An arborescence $T$ of a graph $H$ is a directed, rooted tree subgraph of $H$ in which all edges point away from the root.
4.3. Social Identity Theory

Real-world social networks of people have been shown to exhibit properties of searchability, which means a target can be found quickly even in the absence of global network view [76]. Searchability in social networks has been attributed to the tendency of people to cluster their contacts into meaningful groups based on different attributes and selecting relevant cluster of contacts to advance the search at each hop which would take the search closer to the destination. This arrangement of neighbors in groups gives a sense of social identity [77].

Formally, social identity is defined as:

“that part of an individual’s self concept which derives from his knowledge of his membership of a social group (or groups) together with the value and emotional significance attached to that membership.” [77]

In other words, each individual based on his knowledge of his contacts breaks down his contacts hierarchically into layers, where the top layer represents all his contacts and each successive layer corresponds to a division of contacts into more specific groups. Theoretically, this successive division of contacts into groups is possible to a level where each group contains a single individual, however, in practice and for purpose of identification people cluster their contacts into groups of manageable size. This organization of the contacts into meaningful groups and their affiliations gives rise to a sense of identification of an individual. This forms the social identity of that individual. To summarize, a person can be defined by the contacts he has and the groups these contacts form and their affiliations in these groups.

Knowledge of a contact could be multi-faceted. An individual could know various things about his contacts, like physical location of his contacts, their occupation, their interests, etc. Each of these different facet could be used to group his contacts, resulting in different groupings for each facet. Essentially, this means that a single individual could be identified by multiple social identities based on the different groupings of his contacts. Each of these multiple facets are termed as social dimensions. So based on different social dimensions an individual could construct multiple groupings of his contacts, hence he could be defined by multiple social identities. Figure ?? illustrates 2 ways of constructing social identity of an individual $u$ based on 2 social dimensions. Contacts $v$ and $w$ of $u$ are grouped under different clusters in both the dimensions.
Social dimension $d_1$ could represent *occupation* of the contacts and $d_2$ could represent *physical location* of the contacts. Contacts $v$ and $w$ might be in same occupation but living in different physical locations. Hence in the clustering obtained by considering $d_1$ (i.e., occupation) $v$ and $w$ are grouped in same cluster whereas considering $d_2$ (i.e., physical location) $v$ and $w$ are grouped in different clusters.

Social identity theory has been widely studied in real-world social networks in terms of observing search-ability property of the network. In this work, we attempt to utilize the social identity theory in online social networks to identify familiar strangers, which is the first of its kind to the best of our knowledge. Directly connected neighbors of a node form the set of its contacts and the attribute-value set of the nodes are used to construct the social identity. More details on social identity construction using the attribute-value set is described in the Social Identity Construction.

4.4. Approaches for Egocentric View

Here we present strategies to find familiar strangers $T_u$ of a node $u$ given goal $\gamma$ using an egocentric view of the network.

4.4.1. Social Identity Approach

According to social identity theory, people cluster their contacts into meaningful groups and pick the cluster that has maximum similarity with the goal $\gamma$. So we prune some contacts at each level and propagate the search with the selected cluster of contacts to ensure that the search remains closer to the specified $\gamma$ instead of wandering away.

**Social Identity Construction:** Social identity based search relies on the ability of a node $u$ to cluster its contacts. Each node in the network is represented as a vector space model of its attributes and simple cosine similarity based measures could be used to compute affinity matrix between contacts of node $u$. Then conventional clustering algorithms like k-means could be used to cluster the contacts of node $u$. The clustering approach could be more sophisticated if more data is available about the nodes of the network besides the attribute-value set. For a blogger social network dataset along with the blogger network and their attribute-value set$^3$ we also have their blog posts and the metadata associated with the blogger like tags, categories, ...

$^3$Bloggers’ attribute-value set construction is explained in more detail in the BlogCatalog section.
and blog post text. This rich metadata about the bloggers is used to construct the vector space model for each of the contact $s$ of a node $u$ in blogger network. Here the terms of the vector space model are the words in the vocabulary after removing stop words and stemming. However, this is a very sparse and high-dimensional vector. So this sparse vector could be transformed to concept space vector using latent semantic analysis [71]. The transformed vector is less sparse and low dimensional.

Clustering of the contacts could be performed either offline or online while searching. We perform the clustering offline. So the social identity of the nodes of the network are constructed $a priori$ to speedup the search process. Online clustering takes care of the dynamics of the network, nevertheless, it increases the response time while searching. We can also bypass the construction of social identities to search for familiar strangers. Perhaps this would mean that the search phase will look at all the contacts of a node to find the most relevant nodes to propagate the search. However, by constructing social identity we cluster the contacts and pick the relevant cluster, hence pruning the search space early on. Since clustering is done offline, which does not incur clustering overhead costs while searching.

**Example 2:** To illustrate with an example, refer to Figure 39, where we need to find the familiar strangers of the node $u$ with respect to the goal, $\gamma = \{\text{Exercise, Recreation}\}$. We can either search all his contacts viz., $v_1, v_2, v_3, v_4$ to find the contacts that are similar to the $\gamma$. This would result in $v_4$ as the contact whose attribute-values match with the $\gamma$. Or we can cluster the contacts offline and pick the relevant cluster. Clustering resulted in two clusters one with $v_1, v_2, v_3$ and the other with $v_4$. Now the second cluster with $v_4$ is more similar to the $\gamma$, so we pick the contacts in this cluster, which in this case is $v_4$. The latter strategy greatly prunes the search space, especially when the nodes have much larger number of contacts\(^4\) and the clustering is performed offline.

**Social Identity based Search** for familiar strangers of a node $u$ and $\gamma$ can be summarized in the pseudo code in Algorithm 2. Given a node: $u$, its contacts: $C_u$, its contacts’ attribute-value set: $B_{C_u}$, and $\gamma$ as input, it outputs a set of node(s) $T_u$ that are the familiar stranger(s) for $u$. Algorithm first clusters the contacts $C_u$

\(^4\)It has been found that on average people have approximately over 150 contacts, also known as the Dunbar number [78].
Algorithm 2: Searching Familiar Strangers of $u$.

of node $u$ and selects the cluster that has maximum similarity with $\gamma$, i.e., $C'_u$. Then among the node(s) in $C'_u$, node(s) whose attribute-value set matches with $\gamma$ are selected and we call this set of node(s) $C''_u$. The node(s) in $C''_u$ are then added to a data structure $Q$. For each node $t$ in $Q$ search is repeated by first clustering the contacts $C_t$ of node $t$ and then selecting the cluster that has maximum similarity with $\gamma$, i.e., $C'_t$. Then further filter $C'_t$ by selecting the node(s) whose attribute-value set matches with $\gamma$. Assign these node(s) to the set $C''_t$. Node(s) in $C''_t$ are added to the set $T_u$ and $Q$. $Q$ is a FIFO data structure to ensure a breadth-first search. We do not add $C''_u$ to $T_u$ since these are the adjacent contacts of $u$ and not strangers.

A social network could be a cyclic graph so a person might get multiple requests to search his contacts to find familiar strangers. We assume that a node searches his contacts only once. This is realistic because once a person has searched his contacts and forwarded the search request to his contacts he has no incentive to do it again. This is realized by associating a `participatedFlag` to each node which is set to false by
default and is set to true once the node gets a search request and forwards it to his contacts (line 16 in the Algorithm 2). A node checks the participatedFlag before searching its contacts and propagating the search to the next hop (line 9 in the Algorithm 2).

4.4.2. Exhaustive Search Approach

Here a node explores all his contacts and his contacts explores all their contacts and so on to search for the nodes that have maximum similarity with the goal \( \gamma \). This procedure continues till all the familiar strangers \( T_u \) of the node \( u \) are found. This exhaustive search procedure incurs an exponential computational cost. Approximately, for an average degree \( d \) of the network, and \( h \) hops needed to find all the familiar strangers, the total number of edges the exhaustive approach needs to traverse is \( O(d^h) \) which is exponential to the search depth. However, exhaustive search guarantees that all the familiar strangers of a node are found.

4.4.3. Random Search

The search starts from \( u \) and propagates by randomly selecting some nodes at each hop. A user-specified selectivity fraction \( \sigma \in \mathbb{R} \) and \( \sigma \in [0, 1] \) controls the number of contacts randomly selected at each hop. This is different than the social identity based search because of (1) no clustering of the contacts of a node, and (2) no intelligent selection of contacts in random search approach. Exhaustive search is a special case of random search where \( \sigma = 1 \).

4.5. Datasets

A blogger network, BlogCatalog\(^5\) and citation network DBLP\(^6\) is used for the evaluation of different approaches.

**BlogCatalog** A blog in BlogCatalog is associated with various information pieces like the categories the blog is listed under, blog level tags, snippets of 5 most recent blog posts, and blog post level tags. Bloggers submit their blogs to BlogCatalog and specify the metadata mentioned above for improved access to their blogs. This way the blog sites are organized under pre-specified categories. A blogger also specifies his social network of other bloggers. A blogger’s interests could be gauged by the categories he publishes his

\(^5\)http://www.blogcatalog.com

\(^6\)http://kdl.cs.umass.edu/data/dblp/dblp-info.html
blogs in. There are in total 60 categories in BlogCatalog. Each blogger could list his blog under more than one categories. On average each blogger lists their blog under 1.6 categories. All the categories his blog has been published are agglomerated to construct his profile vector. This profile vector forms the attribute-value set for this blogger. However, in case where the category information is unavailable, we can use various existing author-topic model extraction approaches [79] to extract topics of the author from the text in blog posts, tags, and comments. Topics hence extracted for each blog author could be treated as their attribute-values. We use BlogCatalog API to collect bloggers’ social network and their attribute-values. We collected 23,566 bloggers with 1,165,622 number of blogger-blogger links and 0.002 link density. This shows that the dataset is highly sparse. The average degree of the blogger social network is 98. This shows that on average a blogger has as many as 98 bloggers in his/her social network due to which the diameter of the network is as small as 5. This means that even the furthest pair of nodes are 5 hops apart. Each blogger is represented by two vectors; first vector for his social network and second vector is the attribute-values. Note that the blogger’s social network vector is extremely sparse as also depicted by the average degree of nodes and link density in Table 17.

**DBLP** dataset presents information on computer science publications. We construct social network of authors using the co-author relation. Two authors are connected through an edge if they have collaborated on at least one paper. So all the co-authors of an author constitute his social network. Each author publishes his work in the choice of his venue, which also tells us about his interests. Based on the venue information of the publications we construct the attribute-value set of each author. So each author in DBLP dataset is again represented by two vectors; first vector for their social network and second vector is their attribute-value set. There are about 35,001 authors in the dataset and 3198 number of different venues. On average each author publishes in 28.7 venues. This means that on average each author has 28.7 elements in attribute-value set, much larger than the BlogCatalog dataset. There are 1,067,447 number of author-author links and 0.0009 link density. This shows that the dataset is even more sparse than the BlogCatalog dataset. An author can collaborate with only a few authors in his life time. We use a part of DBLP dataset which is the largest connected component of the graph generated using the co-author relation. The average degree of the author social network is 9. This shows that on average an author collaborates with 9 authors, much smaller than
TABLE 17
Summary of BlogCatalog and DBLP Datasets.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>BlogCatalog</th>
<th>DBLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>23,566</td>
<td>35,001</td>
</tr>
<tr>
<td>Number of node-node links</td>
<td>1,165,622</td>
<td>1,067,447</td>
</tr>
<tr>
<td>Link density</td>
<td>0.002</td>
<td>0.0009</td>
</tr>
<tr>
<td>Average degree of nodes</td>
<td>98</td>
<td>9</td>
</tr>
<tr>
<td>Diameter of the network</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Attribute name</td>
<td>Categories</td>
<td>Venues</td>
</tr>
<tr>
<td>Size of domain of the attribute</td>
<td>60</td>
<td>3198</td>
</tr>
<tr>
<td>Average size of attribute-value set per node</td>
<td>1.6</td>
<td>28.7</td>
</tr>
</tbody>
</table>

the BlogCatalog dataset, due to which the diameter is twice as large as the BlogCatalog as summarized in Table 17.

4.5.1. Dataset Characteristics

For BlogCatalog and DBLP, we investigate characteristics like power-law degree distribution and small-world assumption which are necessary for searchability in the network with local information [76].

**Degree Distribution** We study the degree distribution of the nodes in BlogCatalog and DBLP dataset. We display the log-log graph of this distribution with \( \log(\text{degree}) \) on the x-axis and \( \log(\text{frequency}) \) on the y-axis, for BlogCatalog and DBLP in Figure 40 and Figure 41, respectively. We observe that both BlogCatalog and DBLP dataset follow power law distribution \( P(x) \sim x^{-k} \) with scaling exponent \( k \) of 1.693 and 2.7896, respectively.

**Small-World Assumption** Networks conforming to small world assumption are characterized by short average path lengths and high clustering coefficient [80]. The distance between any two nodes in the network is defined as the number of edges along the shortest path connecting them. Average path length of a network is defined as follows [80]:

\[
l_G = \frac{1}{n \times (n - 1)} \times \sum_{i,j} d(v_i, v_j)
\]  

(4.1)

where \( n \) is the number of vertices in the graph \( G \) and \( d(v_i, v_j) \) denotes the shortest path between two nodes.
Fig. 40. Log-log Plot of Degree Distribution for BlogCatalog.

Fig. 41. Log-log Plot of Degree Distribution for DBLP.

$v_i$ and $v_j$. For BlogCatalog and DBLP, we computed the average path length using the above formulae and was found to be 2.379 and 5.083, respectively.

Clustering coefficient is a common property of social networks representing circles of friends in which every member knows every other member. If a node $v$ in graph $G$ is connected to $k_v$ other nodes then the clustering coefficient of node $v$ is defined as [80]:

$$C_v = \frac{2E_v}{k_v(k_v - 1)}$$

(4.2)

where $E_v$ is the actual number of edges that exist between the $k_v$ vertices. We compute $C_v$ for all the vertices $v$ of the graph $G$ and compute the average value. We compare the clustering coefficient values of the two datasets with that of random networks generated using the same set of nodes as in BlogCatalog and DBLP but the edges are rewired according to Erdős-Rényi model [81]. We report the results for clustering coefficient for both the datasets and their random network counterparts in Table 18, which shows that clustering coefficient values for the original datasets (Actual Networks) is much higher than their random counterparts (Random Networks). Low average path length and high clustering coefficient implies that the two datasets indeed exhibit small- world characteristics.
TABLE 18
Clustering Coefficient Results for Both Datasets.

<table>
<thead>
<tr>
<th></th>
<th>Actual Network</th>
<th>Random Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlogCatalog</td>
<td>0.51</td>
<td>0.001 ± 0.0002</td>
</tr>
<tr>
<td>DBLP</td>
<td>0.69</td>
<td>0.001 ± 0.0002</td>
</tr>
</tbody>
</table>

4.6. Experiments - Constructing Social Identity

Social identities of the nodes are not available in the online social networks, so we construct the social identities of the nodes using conventional clustering algorithm - *k*-Means. BlogCatalog dataset has very rich metadata for the bloggers, including blog posts and tags. We construct the social identities of the bloggers using the metadata as mentioned in the section on Social Identity Construction. The DBLP dataset doesn’t have any details about the authors besides their venues. So we cluster the contacts of an author using the venue information.

Here we present the results of social identity construction of the nodes of the blogger network from BlogCatalog dataset. To avoid the high-dimensionality and synonymy and polysemy issues we use latent semantic analysis to transform the term space vector to concept space as mentioned before in the Social Identity Construction section. Since we use k-Means algorithm to construct the clusters, we need to find the optimal value of $k$ to compute the clusters. To determine the cluster number $k$, we try to maximize the following ratio:

$$
\frac{1}{k} \sum_{c_i} \left( \frac{2}{\|c_i\| \times (\|c_i\| - 1)} \sum_{v_m \in c_i, v_n \in c_i} \text{Cosine}(v_m, v_n) \right)
\frac{2}{k(k-1)} \sum_{c_i, c_j, i < j} \left( \frac{1}{\|c_i\| \times \|c_j\|} \sum_{v_m \in c_i} \sum_{v_n \in c_j} \text{Cosine}(v_m, v_n) \right)
$$

(4.3)

s.t. $2 \leq k \leq \|D\|

In the above formula, $c_i, c_j$ represent two different clusters $i$ and $j$. $v_m, v_n$ are two different vectors representing two different bloggers. $k$ varies from 2 to the number of contacts a node has, i.e. $\|D\|$. $\text{Cosine}(b_i, b_j)$ gives the cosine similarity between the two bloggers, $b_i, b_j$. Each blogger has two vectors: the content vector ($b_i^c, b_j^c$) and tag vector ($b_i^t, b_j^t$). We compute the cosine similarity between the two bloggers by linearly combining the cosine similarity of each of the two corresponding vectors by assigning 0.3 and 0.7 weight.
Within Similarity and Between Similarity by Different Clustering Methods.

<table>
<thead>
<tr>
<th></th>
<th>k-Means</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Similarity</td>
<td>0.71</td>
<td>0.52</td>
</tr>
<tr>
<td>Between Similarity</td>
<td>0.51</td>
<td>0.52</td>
</tr>
</tbody>
</table>

To evaluate the effectiveness of k-Means, we cluster the contacts by k-Means and random partition, by setting the $k$ to 30. For k-Means, we randomly choose the nodes to start clustering. For random partition, the contacts are distributed into 30 clusters randomly. The average Within Similarity and Between Similarity values are computed for the clusters obtained from both k-Means and random partition. Table 19 shows the average Within Similarity and Between Similarity values for k-Means and random partition method over 100 nodes in Figure 42. We fit a polynomial trendline to help visualize the trend of the increase in the ratio of Within Similarity and Between Similarity. It is evident from Figure 42 that after a certain value of $k$ (≈ 30), the increment in this ratio is small. This means that the ratio increases faster when $k$ is small, and the trend becomes flat for larger values of $k$. We simply set the number of clusters to 30.

To evaluate the effectiveness of k-Means, we cluster the contacts by k-Means and random partition, by setting the $k$ to 30. For k-Means, we randomly choose the nodes to start clustering. For random partition, the contacts are distributed into 30 clusters randomly. The average Within Similarity and Between Similarity values are computed for the clusters obtained from both k-Means and random partition. Table 19 shows the average Within Similarity and Between Similarity values for k-Means and random partition method over 100 nodes.

\[ \frac{d}{dk} \frac{Within\ Similarity}{Between\ Similarity} \]

These values of weights give the best result. Due to space constraint we do not present the results with different weight values.
runs. It is evident from Table 19 that k-Means clustering gives dense or cohesive and well-separated clusters as implied by higher Within Similarity and lower Between Similarity as compared to random partition.

4.7. Experiments - Searching Familiar Strangers

In this section we compare the proposed social identity based search approach with other alternatives, viz., Steiner tree approach, exhaustive search approach and random search approach. We compare these approaches in terms of accuracy and search space complexity as explained next.

4.7.1. Evaluation Criteria

To compare the above-mentioned approaches we need to establish a ground truth. As mentioned in the Problem Formulation section, Steiner Tree based approach has the global view of the network $G$ with $V$ vertices, so we construct the ground truth using Steiner Tree based approach. For a given goal $\gamma$, Steiner Tree based approach extracts a subgraph $G'_\gamma$ from the original graph containing nodes that share a part or whole of the $\gamma$ (required vertices), $V'_{\gamma}$, as well as some nodes that do not share $\gamma$ at all (Steiner vertices or optional vertices), $V^{SV}_\gamma$. This subgraph could be used to identify the familiar strangers of any node which is a part of this subgraph. Basically, the required nodes that are not directly connected to a node $u$ in this subgraph are the familiar strangers of $u$ or $T_u$ and forms the ground truth, denoted by $V^{FS}_\gamma$ and is computed as $V'_{\gamma} - V^{SV}_\gamma$.

4.7.1.1. Accuracy

To evaluate an approach $E$ (where $E$ could be one of the social identity based search approach, random search approach, and exhaustive search approach), we pick a node $u$ from the given network such that the attribute values $A_u$ of $u$ and the goal $\gamma$ are similar. This constraint is realized by setting $\gamma \subseteq A_u$ as also defined in the Problem Formulation section. Recall that this is the same $\gamma$ that was used to generate the ground truth of familiar strangers using Steiner Tree based approach. Then we use the strategy $E$ to generate the familiar stranger nodes for $u$ denoted by $V^{E}_{u,\gamma}$. We repeat this process for all such possible nodes and aggregate the familiar strangers identified for each node, denoted by $\bigcup_{u \in V, \gamma \subseteq A_u} V^{E}_{u,\gamma}$. Then accuracy for approach $E$ is computed as the intersection between the ground truth computed by using Steiner Tree based approach and the familiar strangers identified by $E$ for the $\gamma$ normalized by the total number of the familiar
TABLE 20
Comparison of the Approaches in Terms of Accuracy and Search Space Complexity for BlogCatalog Dataset.

<table>
<thead>
<tr>
<th>Approach (E)</th>
<th>Accuracy (%)</th>
<th>Search Space Complexity (edge traversals)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steiner Tree</td>
<td>100%</td>
<td>3,565 ± 560</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>100%</td>
<td>4,531,967 ± 891,831</td>
</tr>
<tr>
<td>Random</td>
<td>1.0283% ± 0.862</td>
<td>1,823 ± 1,833</td>
</tr>
<tr>
<td>Social Identity</td>
<td>79.2908% ± 9.052</td>
<td>6,032 ± 2,117</td>
</tr>
</tbody>
</table>

strangers identified by the Steiner Tree based approach as the ground truth. Mathematically, we can represent accuracy of an approach E with respect to a goal \( \gamma \) as,

\[
\text{Acc}^E_\gamma = \frac{|V^F \cap (\bigcup_{u \in V, \gamma \subseteq A_u} V^E_u)|}{|V^F|} \tag{4.4}
\]

4.7.1.2. Search Space Complexity

We define the search space complexity of an approach \( E \) as the number of hops traversed to find the set of familiar stranger nodes with respect to a goal \( \gamma \) \((\bigcup_{u \in V, \gamma \subseteq A_u} V^E_u)\). Since Steiner Tree based approach finds the set of familiar stranger nodes with respect to a goal \( \gamma \) by traversing minimum number of edges. We exploit this property to establish the lower bound on the search space complexity for various approaches.

4.7.2. Results and Analysis

In our experiments we test for 1000 goal (\( \gamma \)) values. For each value of \( \gamma \) we generate the set of familiar stranger nodes using the approaches mentioned above. We compute the accuracy for each of the mentioned approaches as explained in the section on Accuracy and also compute the search space complexity in terms of the hops traversed as described in the section on Search Space Complexity. We average the accuracy values over all the goals, i.e., 1000 \( \gamma \) values. We report the average accuracy values along with the search space complexity for all the approaches in Tables 20 and 21.

From the Tables 20 and 21 it can be observed that, though exhaustive approach gives 100% accuracy it bears an overwhelming search cost to discover all the familiar stranger nodes. On the other hand social
TABLE 21

Comparison of the Approaches in Terms of Accuracy and Search Space Complexity for DBLP Dataset.

<table>
<thead>
<tr>
<th>Approach (E)</th>
<th>Accuracy (%)</th>
<th>Search Space Complexity (edge traversals)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steiner Tree</td>
<td>100%</td>
<td>4,752 ± 907</td>
</tr>
<tr>
<td>Exhaustive</td>
<td>100%</td>
<td>909,543 ± 162,651</td>
</tr>
<tr>
<td>Random</td>
<td>2.304% ± 0.1264</td>
<td>58 ± 159</td>
</tr>
<tr>
<td>Social Identity</td>
<td>91.3495% ± 4.4398</td>
<td>12,182 ± 4,716</td>
</tr>
</tbody>
</table>

identity based search approach achieves 79.2908% accuracy for BlogCatalog and 91.3495% accuracy for DBLP dataset. However, the social identity based approach searches approximately 0.1331% of the space as searched by exhaustive search approach in BlogCatalog and 0.1136% for DBLP. This shows a phenomenal reduction in the search space using social identity of the nodes while searching for familiar strangers.

We present the results for Steiner Tree based approach as a lower bound for search space complexity in Tables 20 and 21. Since Steiner Tree based approach assumes global information of the network, it can discover all the familiar stranger nodes, hence it achieves an accuracy of 100% for both BlogCatalog and DBLP dataset in minimum search steps. However, social identity based search approach, which does not have global information about the network, searches only a couple of factors more of the search space (precisely, 1.69 and 2.56 for BlogCatalog and DBLP, respectively). Since the social identity based search approach has egocentric view of the network, it cannot achieve 100% accuracy, but it still performs reasonably well as compared to Steiner tree approach for both the datasets.

Deeper analysis explains the reason for such a drastic reduction in search space complexity. We computed the average number of contacts selected at each hop for the social identity based search, which comes out to be 0.030 ± 0.006 and 0.039 ± 0.011 for BlogCatalog and DBLP dataset, respectively. This means that as few as 3% and 4% of nodes are selected on average at each hop that propagates the search at the next hop, respectively for BlogCatalog and DBLP datasets. This extremely small fraction of nodes selected at each hop is the reason why social identity based search approach has such a small search space complexity.

To test the effectiveness of the social identity based approach we compare it with the random search
approach. Random search approach selects a percentage of nodes at each hop randomly and propagates the search to the next hop. This doesn’t involve any intelligent selection of the contacts. For a fair comparison we assigned $\sigma$ (the selectivity parameter for random search approach) as the selectivity for social identity based search approach, which was found to be $0.030 \pm 0.006$ and $0.039 \pm 0.011$ for BlogCatalog and DBLP dataset, respectively. A comparison of accuracy values between the random search approach and social identity based search approach (in Tables 20 and 21) clearly shows that intelligent selection of contacts based on social identity theory improves the accuracy phenomenally. Note that random search approach selects the contacts randomly at each hop so for each goal $\gamma$ value we run the random search 1000 times and report average accuracy and search space complexity results for a particular $\gamma$. Finally for all the 1000 goal $\gamma$ values we compute the average accuracy and search space complexity results.

Next we compare the various approaches at different accuracy values in terms of search complexity. This experiment is performed to observe the search space complexity behavior as we attempt to find increasingly larger number of familiar strangers. We report the results in Figure 43 for BlogCatalog dataset and in Figure 45 for DBLP dataset. Note that since the random search approach does not give reasonable accuracy ($< 10\%$) in both the datasets, we do not include it in these experiments. It is evident from the figures that exhaustive search based approach has an exponential behavior. The overwhelming search space complexity
of the exhaustive search approach overshadows the search space complexity behavior for social identity and Steiner Tree approach. To observe the search space complexities of social identity and Steiner Tree based search approach we plot accuracy vs. log of search steps in Figure 44 and 46 for BlogCatalog and DBLP datasets, respectively. It shows that social identity and Steiner Tree based search approach are comparable in terms of search space complexity. However, exhaustive search approach is almost 2-3 orders of magnitude higher than both the social identity and Steiner Tree based approach for both the datasets. This shows that social identity based search is closer to Steiner Tree based search approach in terms of search space complexity although social identity based search assumes only egocentric view unlike Steiner Tree based search approach that assumes global view of the network.

In the previous experiments random search approach did not perform well due to the low selectivity ($\sigma = 0.03$ and 0.039 for BlogCatalog and DBLP, respectively). We used these selectivity values to ensure fair comparison between random search approach and social identity based search approach. Next we perform experiments to explore the characteristics of the random search approach by varying the selectivity parameter.

First, we study how accuracy behaves as we vary selectivity ($\sigma$) from 0.01 to 0.1 in increments of 0.01 and 0.1 to 1.0 in increments of 0.1. This is done to zoom-in to the behavior of random search approach during the interval of $\sigma$ between 0.01 and 0.1. We report the results in Figure 47 and Figure 48 for BlogCatalog and
Fig. 47. Selectivity vs. Accuracy for BlogCatalog Dataset.

Fig. 48. Selectivity vs. Accuracy for DBLP Dataset.

DBLP datasets respectively. As $\sigma$ is varied accuracy increases and then stabilizes at 100% at a quite high value of $\sigma$. To reach the accuracy value reported by social identity based search approach $\sigma$ has to be much higher for random search approach for both BlogCatalog (0.3) and DBLP (0.5) datasets. However, increasing the $\sigma$ for the random search approach to this extent proves costly in terms of search space complexity. As shown in Figure 49 and Figure 50 search space complexity increases linearly with the $\sigma$, for both the datasets. Note that $\sigma = 1.0$ is a special case of random search approach which is equivalent to exhaustive search approach.

Next we study the search space complexity as we attempt to find increasingly larger number of familiar strangers for random search approach for varying $\sigma$. We report the results for both BlogCatalog and DBLP in Figure 51 and Figure 52, respectively. Each curve in the figures represent the search steps vs. accuracy plot for a particular $\sigma$ value. Both the figures show almost similar behavior. As $\sigma$ increases and approaches 1.0 the search space complexity gets more exponential with respect to accuracy. This is due to the fact that as $\sigma$ increases random search approach has to search more contacts at every hop, thus becoming more exponential.

4.8. Related Work

To the best of our knowledge no work uses the social identity theory to search for familiar strangers, so we review extant literature in identifying latent friends and clustering nodes of a social network.

**Identifying Latent Friends** Authors in [82] use Social Network Analysis (SNA) to discover groups of
individuals sharing the same connectivity properties of networks. Since this does not consider the textual information of the entities, it limits the applications of SNA. Authors in [79] [83] use LDA and its variations to mine relationships between people based on the content. These approaches develop topic models on the documents submitted by the authors. Authors may produce several documents often with coauthors, making it unclear how the topics generated for these documents might be used to describe the interests of the authors. Moreover, it is challenging to learn the parameters in these approaches even though well-established approximation techniques exist. Considering the limitations of author-topic model based approaches to identify latent relations, authors in [84] train an SVM to predict the topics for bloggers from external topic taxonomies. Based on the topic similarity, further refined by the cosine similarity of actual blog content, similar bloggers can be recommended. As topic taxonomies keep evolving, it requires re-training the classifier that adds complexity to the solution. Moreover, detecting bloggers true interests in some of their writings could be a big challenge at times. Unlike the familiars strangers, the latent bloggers identified by [84] could possibly know each other. Other key differences are, the constraint of egocentric network view and use of social identity theory in searching familiar strangers.

Clustering in Social Networks Since we perform clustering of members of social network, we briefly review existing approaches in this domain and compare and contrast them with the work presented here.
A more comprehensive literature survey on clustering in Blogosphere can be found in Chapter 3. Girvan and Newman [6] proposed a divisive algorithm by measuring “edge betweenness” based on the observation that the inter cluster edges have a large “edge betweenness” value if the communities are loosely interconnected. [85] improves the former work by considering the “edge-clustering coefficient” as the number of triangles to which a given edge belongs, divided by the number of triangles that might potentially include it, which is similar to the definition of “clustering coefficient” first introduced by [80]. Another measure to detect the community is modularity [86] which estimates the fraction of in-links in a community minus the expected value of in-links in a network with the same community structure but random connections between the nodes. Unlike above methods which search for the non-overlapping communities, [87] explores overlapping communities based on the idea that a community consists of several complete subgraphs that share several nodes.

4.9. Summary

In this chapter, we studied the familiar strangers in online social networks and identify the numerous research opportunities and business advantages of identifying and aggregating the familiar strangers. We formulate the problem and propose a social identity theory based solution with other alternatives. We also show that under certain circumstances, the problem of identifying familiar strangers can be reduced to a well-known
np-complete Steiner tree problem and study its 2-approximation solution to estimate the lower bound on the search space. The Steiner tree solution is also used to generate the ground truth. We performed extensive experiments on a real world blogger social network dataset, BlogCatalog and citation network dataset, DBLP to show that the proposed social identity based approach outperforms the other alternative approaches and is quite close to the Steiner tree based search approach in terms of search space complexity.
In earlier chapters we looked at various challenges in the blogosphere, such as, phenomenal growth, dynamism, and the long tail phenomenon. Specifically in Chapter 2, we discussed the challenges faced by the phenomenal growth of the blogosphere. We proposed a model to identify influential bloggers based on their social gestures, who can be considered as representatives of the community. We studied different characteristics of influential bloggers based on their activity and temporal patterns. We also observed evaluation challenges and proposed an avant-garde evaluation framework leveraging social media to verify the accuracy of the proposed model. In Chapter 3, we studied the challenges faced by the dynamism of the blogosphere environment. Specifically we observed how people, unaware of the dynamic category structure of the blogosphere, assign inappropriate category labels and affect their visibility and unintentionally pollute the data. We proposed an approach that leverages wisdom of the crowds, or conventionally known as the collective wisdom, to cluster their blogs under appropriate categories on one hand and is dynamic on the other hand. Chapter 4 studied the long tail distribution in the blogosphere and its affects on search, personalization services such as customization and recommendation, and business opportunities. The power law distribution causes several bloggers in the long tail with similar niche interests to be largely unknown (or formally, non-adjacent). However, aggregating such familiar strangers allows to collect more data and hence better personalized services and business opportunities. We proposed a social identity based approach to search and aggregate familiar strangers amidst issues like egocentric view, constructing social identity, and evaluation.

Another significant challenge in the blogosphere is the link sparsity, due to the casual nature of the blogosphere. Bloggers often do not cite the source which inspired them to write the blog post, creating a lot of missing links in the blog network. This creates a lot of challenges to the existing link analysis approaches for a variety of tasks, such as, search, community extraction, etc. Besides, typical link analysis based approach quickly runs into exponential search space. Observing these issue in the blogosphere, we perform an empirical study of community interaction beyond link analysis [11]. The empirical study circumvents the challenges with link analysis based approach, to observe interaction within community blogs via an observed event and community reaction to that by studying the opinion and sentiments of the members towards that
event. We perform a case study on ethnic community blogs exploiting the proposed model and report our preliminary findings and observations.

Interaction between individuals or communities could be studied using link analysis. The underlying assumption is that like-minded people interact more often. This can be used to extract communities and further explore the interaction. However, as mentioned earlier, link analysis based approaches have inherent problems like exhaustive search space and link sparsity, we explored a new type of interaction between individuals or communities, based on an observed event. Based on the reactions that individuals/communities have on an event/issue one can identify whether two individuals/communities are similar or not. Intuitively, if individuals/communities consistently express similar feelings on an issue or an event then they tend be similar. This concept forms the bottom-line of our proposed approach for identifying similar individuals/communities via interaction through observation. Hereafter individual and communities refer to individual blog and community blog, respectively. We will explain the approach in more detail after a small necessary background information on types of sentiments that is important to understand the approach.

Reactions of an individual or community blog to an event could be viewed as either like, dislike or indifferent. Since an indifferent reaction of two individuals or communities towards an event does not imply their similarity, we restrict our focus to like and dislike reactions. These reactions are illustrated in the Figure 53. The event here is ‘the new Macbook’ and the blogs on the left show dislike reaction (because of the ‘excessive heat generated by these laptops’) while the blogs on the right show like reaction (because of its ‘great usability’ and ‘cool features’).

Our approach initially requires identifying an event that incites reactions of sufficient posts to analyze. From the blog posts we summarize the text using the tool Subject Search Summarizer (SSSummarizer) [88]. SSSummarizer generates and displays summaries as a list of key sentences the product extracts from documents. By presenting and translating sentences it reflect the subject of a given document thus providing the key information. The SSSummarizer allows you to choose the number of sentences to display in the summary. We tried 10, 20 and 30 sentences for different posts and compared the performance of the tool
by manual analysis. We were convinced that when the number of sentences is set to 20, the tool provides relevant information that is good enough for further analysis.

The stop words are eliminated from the summarized text and it is fed into a tag cloud generator\(^1\) that spits out the representative words from the summarized text. The tag cloud generator is an online tool from artviper that generates tag clouds for the given text. The tag cloud generator works well when it is used after the SSSummarizer. If used without the tool and raw posts are given directly as input, the identified tags are extremely noisy and irrelevant and appear in larger font increasing its significance. Moreover the number of tags generated is huge and if in case we restrict the number then the quality of results is compromised.

We then compare the identified tags with the list of sentiment key words given by WeFeelFine\(^2\), augmented by Thesaurus\(^3\), and label the matching words as sentimental words. WeFeelFine provides an API that contains a list of sentiments that have been identified from the blogosphere. Each word has a number associated with

\(^{1}\)http://www.artviper.net/texttagcloud/
\(^{2}\)http://www.wefeelfine.org/
\(^{3}\)http://www.thesaurus.com/
that indicates the number of times it has been identified as sentiment in the blogs. We consider a word to be a Sentiment word only if this number is greater than 10. Thesaurus.com has an online searchable collection of words and is grouped together with antonyms. These groups of words are used to augment the sentiment words obtained from WeFeelFine.

From this point by manually analyzing each of these words we tag them as either a positive (like) sentiment or a negative (dislike) sentiment. From this collection of positive and negative sentiment words we will be able to decide the type of reaction of the individuals/communities for that event. If the reactions of these individuals/communities remain consistently similar then these individuals/communities are more likely to be similar. Thus the approach based on interaction through observation, enables us to easily break the barriers due to link analysis and thus find two similar individuals/communities that are disconnected. A summarized flow chart of the whole approach is illustrated in Figure 54. Next we present a case study on a real world community blog and report the interesting observations. We also present the challenges that we encountered with the proposed model.

A Case Study: As an example consider an event such as Saddam Husseins Verdict. The Sunnis opposed the event stating it to be ridiculous. At the same time the Shiites felt it was a good decision and they were supporting the event. Such interactions could be found in the blogs but there is no direct way to identify how each group reacted to such events without reading the full post. Though there are several tools available to
summarize, identify concepts, themes there is no such tool to find this directly. By identifying sentiments from these blog posts, we can observe their feelings and reactions.

Identifying similar blogs through observing an event involves the following steps.

1. We obtain the posts from the three sites - Iraq The Model, Baghdad Burning and East Kurd in the month of the event, i.e., Saddam Hussein’s Verdict.

2. We use the SSSummarizer to obtain the summary of the posts for each site.

3. The Stop words are eliminated from the summarized text of the blog posts obtained from these blogs.

Fig. 55. Blog Reactions to Saddam Hussein’s Verdict.
4. The summarized text after stop word elimination is given as input to the Tag Cloud Generator to identify the tags.

5. These tags are then checked with the API provided by WeFeelFine (augmented by the Thesaurus) and the matching words are tagged as Sentiment words.

6. The words identified as sentiment words are then tagged manually as either positive sentiments or negative sentiments. From these words that have been identified for the three blogs we are able to observe that they have different feelings towards the event.

The sentiment words clearly revealed that one website i.e. Baghdad Burning opposed the event while the other two i.e., Iraq The Model and East Kurd, were in favor of the event. We also aligned our findings with the ground truth obtained from the news site and came to a conclusion that Iraq The Model aligns well with the Shia, Baghdad Burning aligns well with the Sunnis and East Kurd aligns well with the Kurds. Figure 55 shows the reactions of Iraq The Model and Baghdad Burning to Saddam Husseins verdict. Iraq The Model has a very positive reaction to the event as evident by the tags like ‘accept’, ‘agree’, ‘building’, and ‘patriotic’ highlighted in green color whereas Baghdad Burning has a very negative reaction to the event as evident by the tags like ‘bad’, ‘dead’, ‘demonstration’, ‘shut’, and ‘stupidity’ highlighted in red color.

We also considered another event that was not very famous as the Saddams verdict; the series of suicide bombings in Iraq during the month of August 2006. This event did not have as many posts compared to that of the Saddam Hussein’s verdict. We considered posts from the same three blog sites to identify their reaction to this event. We identified that all the three sites had posts that indicated a negative sentiment. This clearly indicates that the three blog sites strongly oppose the event. This was also manually verified by reading the blog posts. The blogs revealed their grief in the events and people expressed how much they were affected by these bombings in their place. Based on these findings we can observe that East Kurd and Iraq The Model are very similar in terms of their reaction to these events. This analysis could be highly useful in case of a future event. The results for both the events are summarized in Table 22.

Based on the sentiment polarity of blogs as shown in Table 22, correlation can be computed and hence we
TABLE 22
Summary of the Results of the Reaction of Three Different Blogs to the Events.

<table>
<thead>
<tr>
<th>Blogs</th>
<th>Events</th>
<th>Iraq The Model</th>
<th>Baghdad Burning</th>
<th>East Kurd</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1: Nov 2006 - Saddam Verdict - death sentence</td>
<td>Accept and support the verdict</td>
<td>Oppose the verdict, feels its lynching</td>
<td>Accept and support the verdict</td>
<td></td>
</tr>
<tr>
<td>E2: August 2006 - Series of Suicide bomb explosions</td>
<td>Feel bad for it and oppose</td>
<td>Feel bad and oppose the event</td>
<td>Feels bad and opposes</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E_m</td>
<td></td>
<td></td>
<td></td>
<td>??</td>
</tr>
</tbody>
</table>

can obtain similarity between blogs. If two blogs consistently exhibit similar sentiment polarity towards a set of events then they can be considered similar. Correlation between two blogs $b_i$ and $b_j$ can be computed as:

$$Corr(b_i, b_j) = \frac{P_i \times P_j}{\|P_i\| \|P_j\|}$$  \hspace{1cm} (5.1)

where $P_i$ and $P_j$ are the sentiment polarity vectors for blogs $b_i$ and $b_j$, respectively, for a set of events. Also from Table 22, we can predict the behavior of a blog towards an event $E_m$ by finding other blogs with which it has the maximum correlation and observing their sentiment polarity towards the event. This can be used to identify the affiliations of emerging blogs.

**Challenges with the Model:** The current tools available provide a lot of blog posts for a given search query. It is still difficult to affiliate a community into one of the three categories Shia, Sunni or Kurd. This is primarily due to lack of ground truth that hinders in classifying these posts. Manually establishing this truth is a challenge which will enable the categorization of other posts based on this. Moreover detecting the positive or negative (bi-polar feelings) is also a challenge. Currently the list of events is derived from BBC’s
events database. Automatically detecting events is a challenging task which we intend to explore as a future direction.

The API provided by WeFeelFine restricts the sentimental words and does not contain few colloquial words that people use to express their feelings. For instance the word “for” or “pro” that is often used in the place of “support” was not found in the list provided by WeFeelFine. Moreover it would also be eliminated as a stop word. To start with a solution to this we have refined the word list with the help of Thesaurus that contains a list of sentiment words that have been categorized into several groups. This list can be used to identify the sentiment words and also the group to which they belong to. Though we have identified the sentiments from the WeFeelFine, it is pretty tough to tag them as either positive or negative without human intervention. The sentiment list provided by Thesaurus has groups of sentiment words along with the antonyms. This list can be used as a base to color the word as either a positive or negative sentiment. Opinion mining and sentiment analysis is an altogether different research area which we would like to explore further as a future direction to automate the sentiment analysis as much as possible.
REFERENCES


[31] X. Yin, J. Han, and P. S. Yu, “Truth Discovery with Multiple Conflicting Information Providers on the Web,” *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2007.


