

# Blogosphere: Research Issues, Tools, and Applications

Nitin Agarwal      Huan Liu  
Computer Science and Engineering Department  
Arizona State University  
Tempe, AZ 85287

f Nitin.Agarwal.2, Huan.Liu@asu.edu

## ABSTRACT

Weblogs, or Blogs, have facilitated people to express their thoughts, voice their opinions, and share their experiences and ideas. Individuals experience a sense of community, a feeling of belonging, a bonding that members matter to one another and their niche needs will be met through online interactions. Its open standards and low barrier to publication have transformed information consumers to producers. This has created a plethora of open-source intelligence, or "collective wisdom" that acts as the storehouse of overwhelming amounts of knowledge about the members, their environment and the symbiosis between them. Nonetheless, vast amounts of this knowledge still remain to be discovered and exploited in its suitable way. In this paper, we introduce various state-of-the-art research issues, review some key elements of research such as tools and methodologies in Blogosphere, and present a case study of identifying the influential bloggers in a community to exemplify the integration of some major aspects discussed in this paper. Towards the end, we also compare and contrast the blogosphere and social networks and the research therein.

## 1. INTRODUCTION TO BLOGOSPHERE

Weblogs or Blogs are becoming one of the most popular media of communication and interaction among masses. A blog can be defined as a website that displays, in reverse chronological order, the entries by one or more individuals and usually has links to comments on specific postings. Each of these entries are called blog posts. A typical blog post can combine text, images, and links to other blogs, web pages, and other media related to its topic. Some blog posts provide a list of links to similar or related blog posts. Such a list of links is called blogroll. The ability for readers to leave comments in an interactive environment is an important part of blogging. People express their opinions, ideas, experiences, thoughts, wishes through these free-form writings. The individuals who author the blog posts are referred as bloggers. The websites that publish these blog posts are termed as blog sites or blogs. Blog sites often provide opinions, commentaries or news on a particular subject, such as food, politics, or local news; some function more like personal online diaries. The universe of all these blog sites is often referred as Blogosphere.

There has been a tremendous increase in user-generated content in the past couple of years via the phenomenon of blog-

ging. Acknowledging this fact, Times has named "You" as the person of the year 2006. This has created a considerable shift in the way information is assimilated by the individuals. This paradigm shift can be attributed to the low barrier to publication and open standards of content generation services like blogs, wikis, collaborative annotation, etc. These services have allowed the mass to contribute and edit articles publicly. Giving access to the mass to contribute or edit has also increased collaboration among the people unlike previously where there was no collaboration as the access to the content was limited to a chosen few. Increased collaboration has developed collective wisdom on the Internet. "We the media" [21], is a phenomenon named by Dan Gillmor: a world in which "the former audience", not a few people in the back room, now decides what is important. The "former" consumer of the information becomes the new producer, transforming the lecture style of information consumption to conversation-based assimilation.

Blogs have also made it easy for the content generators to author content independent of technical challenges of internet languages and scripts. Bloggers don't need to worry about the low level programming details, rather they focus only on the content. This simplifies the content generation process to a great extent and attracts novice or even computer illiterates to participate in blogging activities. Blogs provide a platform where anyone can express himself or herself freely without being even restrained by their limited computer knowledge yet being able to publish content on the Internet. Publishing on the Internet also facilitates the readers to comment instantly, giving bloggers a feeling of satisfaction.

For many years, psychologists, anthropologists and behavioral scientists have studied the societal capabilities of humans. They present studies and results that substantiate the fact that humans like engaging themselves in complex social relationships and yearn to be a part of social groups. People form communities and groups for the same reasons to quench the thirst for social interactions. Often these groups have like-minded members with similar interests who discuss various issues including politics, economics, technology, life style, entertainment, and what have you. These discussions could be between two members of the group or involve several members.

Internet has virtually reduced the distance between any two points on Earth to zero. It has made possible for people to connect with each other beyond all geographical barriers. Blogs, on the top of it, has tremendously affected social interactions between people and communities. People not only

participate in regional matters but also international issues. They can connect to people sitting on exactly the other side of globe and discuss whatever they like, i.e., a "flat world" [18]. Communities can be spread across several time zones. This humongous mesh of social interactions is termed as social networks. Blogs can be considered as a type of social networks that encompass interactions between different people, members of a community or members across different communities. Each person in a social network is represented as a node and the communications represent the links or edges among these nodes. Blogosphere comprises of several focused groups or communities that can be treated as sub-graphs. These communities are highly dynamic in nature that have fascinated researchers to study its structural and temporal characteristics.

There are myriad services offered under the umbrella of social networks along with Blogs. Other services include social friendship networks like Friendster<sup>1</sup>, Facebook<sup>2</sup>; collaborative annotation like del.icio.us<sup>3</sup>, StumbleUpon<sup>4</sup> that constitute "folksonomy"; media sharing services like Flickr<sup>5</sup>, YouTube<sup>6</sup>; and wikis<sup>7</sup>. All these services offer a fertile ground for research. In this paper we focus on the blogosphere.

The popularity and widespread use of blogs can be attributed to the changes brought by Web 2.0 in the way users interact with the web. Blogs have been around for quite some time but it became unprecedentedly popular with the advent of Web 2.0. Although Web 2.0 may not be a technological shift, it changed the way now people interact through the Internet. People could not only consume information on the Internet but also contribute to it. Easier, more intuitive interfaces with desktop-like experience enticed users to stay connected and contribute their knowledge in terms of blog posts, wiki articles, developing folksonomies, etc. Wikis is an excellent example of Web 2.0 that slowly takes over online encyclopedias due to its sheer breadth of knowledge made possible by mass editing. Since more and more people are trying to be a part of Web 2.0, it has generated enormous amounts of information on the web which is also known as collective wisdom or open source intelligence. The basic differences between Web 1.0 (or, the way Web was accessed previously) and Web 2.0 can be listed as follows:

- 2 Former information consumers are now also producers. Web 2.0 has allowed the mass to contribute and edit articles through wikis and blogs.
- 2 Giving access to the mass to contribute or edit has also increased collaboration among the people unlike Web 1.0 where there was no collaboration as the access to the content was limited to a chosen few.
- 2 Increased collaboration has generated enormous open source intelligence or collective wisdom on the internet which was not there in Web 1.0.

<sup>1</sup> <http://www.friendster.com/>

<sup>2</sup> <http://www.facebook.com/>

<sup>3</sup> <http://del.icio.us/>

<sup>4</sup> <http://www.stumbleupon.com/>

<sup>5</sup> <http://www.flickr.com/>

<sup>6</sup> <http://www.youtube.com/>

<sup>7</sup> <http://www.wikipedia.org/>

Table 1: Comparing Individual and Community Blog Sites.

Blog sites can be categorized into individual blog sites or single-authored blog sites and community blog sites or multi-authored blog sites. Individual blog sites are the ones owned and maintained by an individual. Examples of individual blogs could be Sifry's Alerts: David Sifry's musings<sup>8</sup> (Founder & CEO, Technorati), Ratcli@e Blog{Mitch's Open Notebook<sup>9</sup>, The Webquarters<sup>10</sup> etc. On the other hand, community blog sites are owned and maintained by a group of like-minded users. Examples of community blogs could be Google's Official Blog site<sup>11</sup>, The Unofficial Apple Weblog<sup>12</sup>, Engadget<sup>13</sup>, Boing Boing: A Directory of Wonderful Things<sup>14</sup> etc. We summarize the differences between individual and community blogs in Table 1.

Such an interactive information delivery medium like blogs hosts a conducive ground for the virtual communities or communities that originate over the Internet. There has been a lot of ongoing research to mine knowledge in Blogosphere. This survey is organized as follows: Section 2 introduces various issues pertinent to the blogosphere. Section 3 reviews tools, general methodologies, datasets, and performance metrics that are useful for conducting research in Blogosphere. Section 4 presents a case study. Section 5 discusses the connection between Blogosphere and state of the art social networks. Section 6 concludes the paper with some possible future directions for research in the blogosphere.

## 2. RESEARCH ISSUES

Here we study various research issues and challenges with potential applications. We discuss the research issues in terms of modeling, clustering, mining, community discovery and factorization, influence and propagation, trust and reputation, and spam blog filtering.

### 2.1 Modeling the Blogosphere

The first and foremost challenge lies in developing an appropriate model for the blogosphere. Often researchers and practitioners ask, which is the model that best describes the structure and properties of the blogosphere. Such a model can help in gaining deeper insights into the relationships between bloggers, commenters, blog posts, comments, viewers/readers, and different blog sites in the blogosphere. This

<sup>8</sup> <http://www.sifry.com/alerts/>

<sup>9</sup> <http://www.ratcli@eblog.com/>

<sup>10</sup> <http://webquarters.blogspot.com/>

<sup>11</sup> <http://googleblog.blogspot.com/>

<sup>12</sup> <http://www.tuaw.com/>

<sup>13</sup> <http://www.engadget.com/>

<sup>14</sup> <http://boingboing.net/>

can help us in understanding and defining various concepts of the blogosphere at an abstract level. These type of models would also help in tackling several other challenges of the blogosphere. A model for the blogosphere would be useful in generating an artificial dataset, tuning the parameters to simulate a special scenario and compare different algorithms and studies. Such a model will also help in studying peculiarities in the blogosphere and infer latent patterns and structures that could explain certain phenomena like community discovery, spam blogs, information diffusion and influence, etc., to be discussed later in this section.

Modeling the blogosphere is often associated with modeling the web. Researchers represent the web as a webgraph, where each webpage forms a node and hyperlinks between them as edges. This kind of representation results in a directed cyclic graph. Weights can be associated with these edges. Such a model that converts the web into a graphic model is extensively exploited. One prominent example is the search engine domain which relies on this graph based model of the web to rank webpages [10; 32]. Although the web models seem to be an appropriate choice for modeling the blogosphere but certain key differences prevent reusing the web models in the blogosphere domain. First, models developed for the web assumes a dense graph structure due to a large number of interconnecting hyperlinks within webpages. This assumption does not hold true in the blogosphere, since the hyperlink structure in the blogosphere is very sparse, as shown in [35]. Second, the level of interaction in terms of comments and replies to a blog post makes the blogosphere different from the web. Third, the highly dynamic and "short-lived" nature of the blog posts could not be simulated by the web models. Web models do not consider this dynamicity in the web pages. They assume web pages accumulate links over time. However, in a blog network, where blog posts are the nodes, it is impractical to construct a static graph like the one for the web. These differences necessitate the need for a model more towards the characteristics of the blogosphere.

There are several models for the web like random graph [49], preferential attachment graph [6], hybrid graph [44], and random walk on graph [9]. A random graph constructs edges between each pair of nodes with some probability which fails to exhibit the power law degree distribution or scale-free graph structure. For this reason random graph models cannot be used to model the blogosphere. Preferential attachment graph models follow the phenomenon of "the rich gets richer", where the probability of a new edge to a node to be added is based on its degree. The more the degree of a node the better the chances are for a different node to be connected with this node. These models exhibit the power law distribution. Hybrid graph models are basically a mixture of both random graphs and preferential attachment models, so as to give a "lucky" node a chance to get "rich". Blogosphere can be modeled using this model with some modifications. To solve the problem of irreducibility (strong connectedness with few isolated subgraphs), random walk on a graph model proposes a random jump with a fixed probability between 0.8 and 0.9 in addition to the preferential attachment model. The above models have been used to model the blogosphere with modifications, but these models could not explain the blogosphere precisely. This has motivated researchers to come up with models specific to the blogosphere. Leskovec et al. [38] studied the temporal patterns of the blogosphere

like how often people create blog posts, burstiness and popularity, how these blog posts are linked, and what is the link density. They reported that these phenomena follow power law distributions. Based on their findings, they developed a cascade model similar to the SIS (susceptible-infected-susceptible) model from the epidemiology. This way any randomly picked blog can infect its uninfected immediate neighbors probabilistically, which repeats the same process until no node remains uninfected. In the end, this gives a blog network. Kumar et al. [37] use the blogrolls given on a blog post to create a network of connected posts with the underlying assumption that blogrolls have links to related or similar blog posts. A lot of research has been conducted that posits a known network structure of the blogosphere to model the problem domain. Such models are specific to problem domains and are discussed next in reference to problem domains.

## 2.2 Blog Clustering

Blogosphere is a storehouse of several publicly regulated media. Technorati<sup>15</sup> reported that 175,000 blog posts were created daily which is 2 blog posts per second. This explosive growth makes it beyond human capabilities to look for interesting and relevant blog posts. Therefore a lot of research is going on to automatically cluster different blogs into meaningful groups such that readers can focus on interesting categories, rather than filtering out relevant blogs from the jungle. Often blog sites allow their users to provide tags to the blog posts. The human labeled tag information forms the so-called "folksonomy". Brooks and Montanez [11] presented a study where the human labeled tags are good for classifying the blog posts into broad categories while they were less effective in indicating the particular content of a blog post. They used the tf-idf measure to pick the top three most famous words in every blog post and computed the pairwise similarity among all the blog posts and clustered them. They compared the results with the clustering obtained using the human labeled tags and reported significant improvement. In another research [39], authors tried to cluster blog posts by assigning different weights to title, body and comments of a blog post. However, these approaches rely on the keyword-based clustering which suffers from high-dimensionality and sparsity. Agarwal et al. [2] proposed WisClus that uses the collective wisdom of the bloggers to cluster the blogs. They have used the blog categories and construct the category relation graph to merge different categories and cluster the blogs that belong to these categories. Edges in the category relation graph represent the similarity between different categories which are the nodes in this graph. The similarity between two categories is computed using the number of blogs that simultaneously uses these categories as their blog labels. Experiments show that the collective wisdom based clustering performs better than keyword based clustering even after reducing the dimensionality and sparsity to the concept space using Latent Semantic Indexing (LSI) [14]. Clustering different blog posts would also help blog search engines like Technorati to narrow down the search space once the query context is clear. Websites like Blogcatalog<sup>16</sup> organize blogs into a taxonomy that helps in focussed browsing of blogs.

<sup>15</sup> <http://www.technorati.com/>

<sup>16</sup> <http://www.blogcatalog.com/directory>

## 2.3 Blog Mining

Blog mining as a technique is evolving and taking the form of qualitative research. Companies are using blogs as qualitative research tools. Historically, the interaction between marketers and consumers has been a closed loop. Marketers used to send out messages to consumers and sought their feedback through traditional research. Now, consumers can not only speak their mind but also broadcast their opinions. The surge of marketing messages combined with low consumer trust, has led to people relying on one another's opinions to make informed decisions, prompting conversations between them. These interactions are found on the blogs and have attracted the attention of several companies. Blogs are immensely valuable resources to track consumers' beliefs and opinions, initial reaction to a launch, understand consumer language, track trends and buzzwords, tune information needs. Blog conversations leave behind the trails of links, useful for understanding how information flows and how opinions are shaped and influenced. Tracking blogs also help in gaining deeper insights as bloggers share their views from various perspectives hence giving a 'context' to the information collected.

Mining sentiments from free text forms poses several challenges as compared to the historic feedback and surveys. A prototype system called Pulse [19] uses a Naïve Bayes classifier trained on manually annotated sentences with positive/negative sentiments and iterates until all unlabeled data is adequately classified. Another system presented in [5] improves the blog retrieval by using opinionated words acquired from WordNet in the query proximity. Some well-known opinion mining and sentiment analysis techniques [41] could also be borrowed from text mining domain due to high textual nature of blogs.

## 2.4 Community Discovery and Factorization

Another important research which branched out from the blog-site clustering is determining and inferring communities. Several studies looked into identifying communities in Blogosphere. One method that researchers commonly use is content analysis and text analysis of the blog posts to identify communities in the blogosphere [7], [16], [37]. Kleinberg [32] used an alternative approach in identifying communities in web using a hub and authority based approach, clustering all the expert communities together by identifying them as authorities. Kumar et al. [36] extended the idea of hubs and authorities and included co-citations as a way to extract all communities on the web and used graph theoretic algorithms to identify all instances of graph structures that reflect community characteristics. While Chin and Chignell [12] proposed a model for finding communities taking the blogging behavior of bloggers into account, they aligned behavioral approaches in studying community with the network and link analysis approaches. They used a case study to first calibrate the measure to evaluate a community based on behavioral aspects using a behavioral survey which could be generalized later on, pruning the need of such surveys.

Several researchers have also studied community extraction and social network formation using newsgroups and discussion boards. Although different from the blogosphere we include these here because discussion boards and newsgroups are also very similar to blogs in the sense that they do not have an explicit link structure, and the communication is

not \person-to-person", rather it is more \person-to-group". Blanchard and Markus [8] studied \Virtual Settlement" - a Multiple Sport Newsgroup and analyzed the possibility of emerging virtual communities in it. They studied the characteristics of the newsgroup by conducting interviews with three different kinds of members: leaders (active and well respected), participants (active occasionally to events like triathlons) and lurkers (readers only). They reported that different virtual communities emerge between athletes and those who join the community to keep themselves informed of the latest developments.

## 2.5 Influence in Blogs and Propagation

As communities evolve over time, so do the bellwethers or leaders of the communities who possess the power to influence the mainstream. According to the studies in [30], 83% people prefer consulting family, friends or an expert over traditional advertising before trying a new restaurant, 71% people prefer to do so before buying a prescription drug or visiting a place, 61% of people prefer to do so before watching a movie. This style of marketing is known as \word-of-mouth". \Word-of-mouth" has been found to be more effective than the traditional advertising in physical communities. Studies from [30] show that before people buy, they talk, and they listen. Experts can influence decisions of people. For this reason these experts are aptly termed as the Influentials. Influential bloggers tend to submit influential blog posts that affect other members' decisions and opinions. They accrue respect in the community over time. Other members tend to listen to what the influentials say before making decisions.

Identification of these influential bloggers [4] could lead to several interesting applications. The influentials are potential market-movers. Since they can influence buying decisions of mainstream, companies can promote them as latent brand ambassadors for their products. Being such a highly interactive medium, blogs tend to host several vivid discussions on various issues including new products, services, marketing strategies and their comparative studies. Often this discussion also acts as \word-of-mouth" advertising of several products and services. A lot of advertising companies, approximately 64% [17] have acknowledged this fact and are shifting their focus towards blog advertising and identifying these influentials.

The influentials could sway opinions in political campaigns, elections and reactions to government policies [15]. Because they know many people and soak up a large amount of information, the influentials stand out as knowledgeable, informed sources of advice and insight. Approximately, 84% of the influentials in physical communities are interested in politics and are sought out by others for their perspectives on politics and government, 55% on a regular basis.

The influentials could help in customer support and troubleshooting. A lot of companies these days host their own customer blogs, where people could discuss issues related to a product. Often the influentials on these blogs troubleshoot the problems peer consumers are having, which could be trusted because of the sense of authority these influentials possess. Often the influentials offer suggestions to improve their products. These invaluable comments could be really helpful for companies and customers. Instead of going through each member's blog posts, companies can focus

on the influential's blog posts. For instance, Macromedia<sup>17</sup> aggregates, categorizes and searches the blog posts of 500 people who write about Macromedia's technology.

Some recent numbers from Technorati show a 100% increase in the size of the blogosphere every six months. It has grown over 60 times during the past three years. Approximately 2 new blog posts appear every second<sup>18</sup>. New blog posts being generated with such a blazing fast rate, it is impossible to keep track of what is going on in the blogosphere. Many blog readers/subscribers just want to know the most insightful and authoritative stories before delving into the discussions. Blog posts from influential bloggers would exactly serve this purpose by standing out as representative articles of a blog site. The influential can be the showcases of a group on the blogosphere.

These interesting applications have attracted a surge of research in identifying influential blog sites as well as influential bloggers. Some try to find influential blog sites, in the entire blogosphere and study how they influence the external world and within the blogosphere [20]. The problem of ranking blog sites or bloggers differs from that of finding authoritative webpages. As pointed out in [35], blog sites in the blogosphere are very sparsely linked and it is not suitable to rank blog sites using Web ranking algorithms like PageRank [43] and HITS [32]. The Random Surfer model of webpage ranking algorithms [43] does not work well for sparsely linked structures. The temporal aspect is most significant in the blog domain. While a webpage may acquire authority over time (its adjacency matrix gets denser), a blog post or a blogger's influence diminishes over time. Consequently, the adjacency matrix of blogs (considered as a graph) will get sparser as thousands of new sparsely-linked blog posts appear every day.

Some recent work [35] suggests to add implicit links to increase the density of link information based on topics. If two blogs are talking about the same topic, an edge can be added between these two blogs based on the topic similarity or information epidemics. However, constructing links based on the topic models still remains an area of research. A similar strategy adopted by Adar et al. [1] is to consider the implicit link structure of blog posts. In their iRank algorithm, a classifier is built to predict whether or not two blogs should be linked. The objective in this work is to find out the path of infection (how one piece of information is propagated). iRank tries to find the blogs which initiates the epidemics. Note that an initiator might not be an influential as they might affect only limited blogs. Influentials should be those which play a key role in the information epidemics.

Gruhl et al [24] study information diffusion of various topics in the blogosphere between different blog sites, drawing on the theory of infectious diseases. A general cascade model [23] is adopted. They derived their model from independent cascade model and generalized it to the general cascade model by relaxing the independence assumption. They associate 'read' probability and 'copy' probability with each edge of the blog graph indicating the tendency of a blog to be read and copied, respectively. They also parameterize the stickiness of a topic which is analogous to the virulence of a disease. An interesting problem related to viral market-

ing [46; 31] is how to maximize the total influence among the nodes (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) [28].

Finding influential blog sites is perpendicular to the problem of identifying influential bloggers. Given the nature of the blogosphere, influential blog sites are few. A large number of non-influential sites belong to the long tail [3] where abundant new business, marketing, and development opportunities can be explored. Agarwal et al. [4] studied and modeled the influence of a blogger on a community blog site regardless of the site being influential or not. They modeled the blog site as a graph using inherent link structure, including inlinks and outlinks, as edges and treating different bloggers as nodes. Using the link structure the influence flow across different bloggers is observed, recursively. Other blog post level statistics like blog post quality and comments' information were also used to achieve better results. The model used different weights to regulate the contribution of different statistics. These weights could be tuned to obtain different breeds of influential bloggers. Influential bloggers are not necessarily active bloggers at a blog site [4]. Many blog websites list top 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) [20]. With the speedy growth of the blogosphere, it is increasingly difficult, if at all possible, to manually track the development and happenings in the blogosphere, in particular, at many blog sites where many bloggers enthusiastically participate in discussions, getting information, inquiring and seeking answers, and voicing their complaints and needs.

## 2.6 Trust and Reputation

Open standards and low barrier to publishing has allowed anyone to submit blog posts and contribute to the participatory journalism. On one hand, it has created an overwhelming amount of collective wisdom; on the other hand, it has made difficult for readers to decide whom to trust or believe. This has been a great challenge since the inception of the World Wide Web which created the problem of authoritative webpages. Kleinberg [32] and Page et al [43] tried to give a solution for this problem by exploiting the link structure of webpages. But social networking sites and especially Blogosphere allow mass to create and edit content compromising (risking) the sanctity of the original content. Researchers anticipated this problem in social networking and recommender systems and conducted research in those areas. However, the potential of this research is still underestimated for the blogosphere domain and not much research is reported. Here we briefly point out the work already done in social networks to provide an insight to this problem and mention the current state of trust related research in Blogosphere.

In social networks it is important not only to detect the influential members or experts in case of knowledge sharing in communities but also to assess to what extent some of the members are recognized as experts by their colleagues in the community. This leads to the estimation of trust and reputation of these experts. Some social friendship networks

<sup>17</sup><http://weblogs.macromedia.com/>

<sup>18</sup><http://www.sifry.com/alerts/archives/000436.html>

like Orkut<sup>19</sup> allow users to assign trust ratings implying a more explicit notion of trust. Whereas some websites have an implicit notion of trust where creating a link to a person on a webpage implies some amount of business trust for the person. In other cases, trust and reputation of experts could be typically assessed as a function of the quality of their response to other members' knowledge solicitations. Pujol et al [45] proposed a NodeMatching algorithm to compute the authority or reputation of a node based on its location in the social friendship network. A node's authority depends upon the authority of the nodes that refer to this node and also on the authority of other nodes that this node refers to. The basic idea is to propagate the reputation of nodes in the social friendship network.

While Pujol et al. [45] proposed an approach to establish reputation based on the position of each member in the social friendship network, the authors of [50] developed a model for reputation management based on the Dempster-Shafer theory of evidence in the wake of spurious testimonies provided by malicious members of the social friendship network. Each member of a social friendship network is an agent. Each agent has a set of acquaintances a subset of which forms its neighbors. Each agent builds a model for its acquaintances to quantify their expertise and sociability. These models are dynamic and change based on the agent's direct interactions with the given acquaintance, interactions with agents referred to by the acquaintance, and on the ratings this acquaintance received from other agents. The authors point out a significant problem with this approach which arises if some acquaintances or other agents generate spurious ratings or exaggerate positive or negative ratings, or offer testimonies that are outright false.

Sabater and Sierra [47] propose a combination of reputation scores on three different dimensions. They combined reputation scores not only through social relations governed by a social friendship network (termed as social dimension) but also past experiences based on individual interactions (termed as individual dimension) and reputation scores based on other dimensions (termed as ontological dimension). For large social networks it is not always possible to get reputation scores based on just the individual dimension, so they can use the social dimension and ontological dimension that would enhance the reputation estimation by considering different contexts. The ontological dimension is very similar to the work proposed in [48], where the authors recommend collaboration in social networks based on several factors. They explain the importance of context in recommending a member of social network for collaboration.

In [22], authors consider those social networking sites where users explicitly provide trust ratings to other members. However, for large social networks it is infeasible to assign trust ratings to each and every member so they propose an inferring mechanism which would assign binary trust ratings (trustworthy/non-trustworthy) to those who have not been assigned one. They demonstrate the use of these trust values in an email filtering application and report encouraging results. Authors also assume three crucial properties of trust for their approach to work: transitivity, asymmetry, and personalization. These trust scores are often transitive, meaning, if Alice trusts Bob and Bob trusts Charles then Alice can trust Charles. Asymmetry says that for two people

involved in a relationship, trust is not necessarily identical in both directions. This is contrary to what was proposed in [50], who assume symmetric trust values in the social network between two members. Also, consolidating the trust scores for a member and computing a global trust score for each member might not give a reasonable estimation. Trust of a member is absolutely a personal opinion. Therefore, authors propose personalization of trust which means that a member could have different trust values with respect to different members. Guha et al [25] proposed another trust propagation scheme in social friendship networks based on a series of matrix operations, including the element of distrust along with the trust scores.

Although there has been a lot of work that deals with trust in social networks and recommender systems, not many have considered trust in the blogosphere. Researchers have tried to transform the blogosphere domain to the problem domain considered in trust in social networks. Authors in [29] consider a window of words around the links in a blog post to mine the sentiments about the cited blog post. Using VoteLinks, these sentiments can be classified as positive, negative or neutral sentiments. These bags of sentiments can then be used to compute the link polarity between a pair of blog posts. Using Gruhl's et al [25] trust propagation model, they compute the trust in the network of blog sites in the blogosphere. There is still a lot of information unexploited in this approach like comments from the readers on the blog post that can also be used to judge a blogger's or a blog post's trust.

## 2.7 Filtering Spam Blogs

Spam blogs, often called splogs, is one of the major concerns in the blogosphere. Besides degrading search quality results it also wastes the network resources. So researchers are looking into this aspect of the blogosphere. Although it is a relatively new phenomenon, researchers have compared it with the existing work on web (link) spam detection. For web spam detection, authors in [42] distinguish between normal web pages and spam webpages based on the statistical properties like, number of words, average length of words, anchor text, title keyword frequency, tokenized URL. Some works [26; 27] also use PageRank to compute the spam score of a webpage. Some researchers consider splogs as a special case web spam. Authors in [33; 34] consider each blog post as a static webpage and use both content and hyperlinks to classify a blog post as spam using a SVM based classifier. However, there are some critical differences between web spam detection and splog detection. The content on blog sites is very dynamic as compared to that of web pages, so content based spam filters are ineffective. Moreover, spammers can copy the content from some regular blog posts to evade content based spam filters. Link based spam filters can easily be beaten by creating links pointing to the splogs. Authors in [40] consider the temporal dynamics of blog posts and propose a self similarity based splog detection algorithm based on characteristic patterns found in splogs like, regularities or patterns in posting times of splogs, content similarity in splogs, and similar links in splogs.

## 3. TOOLS AND OTHER RELATED ISSUES

Having presented the status quo of the ongoing research in Blogosphere, we now discuss available tools to analyze the domain, methodologies across different disciplines, data

<sup>19</sup><http://www.orkut.com>

collection and pre-processing, and performance metrics.

### 3.1 Tools and APIs

Several modeling tools are available to simulate the social networks that help study various characteristics of these networks and conduct experiments, including:

- <sup>20</sup> NetLogo<sup>20</sup>: A multi-agent programming language and modeling environment designed in Logo programming language. Modelers can give instructions to hundreds or thousands of concurrently operating autonomous "agents". This helps in exploring the connection between the individuals (micro-level) and the patterns that emerge from the interaction of many individuals (macro-level).
- <sup>21</sup> StarLogo<sup>21</sup>: An extension of Logo programming language. It is used to model the behavior of decentralized systems like social networks.
- <sup>22</sup> Repast<sup>22</sup>: Recursive Porous Agent Simulation Toolkit is an agent-based social network modeling toolkit. It has libraries for genetic algorithms, neural networks, etc. and allows users to dynamically access and modify agents at run time.
- <sup>23</sup> Swarm<sup>23</sup>: A multi-agent simulation package to simulate the social or biological interaction of agents and their emergent collective behavior.
- <sup>24</sup> UCINET<sup>24</sup>: A comprehensive package for the analysis of social network data including centrality measures, subgroup identification, role analysis, elementary graph theory, and permutation-based statistical analysis. In addition, the package has strong matrix analysis routines, such as matrix algebra and multivariate statistics.
- <sup>25</sup> Pajek<sup>25</sup>: (Slovenian: spider) A software for analyzing and visualizing large networks like social networks.
- <sup>26</sup> Network package in R<sup>26</sup>: The network class can represent a range of relational data types, and support arbitrary vertex/edge/graph attributes. This is used to create and/or modify the network objects and is used for social network analysis (SNA).
- <sup>27</sup> InFlow<sup>27</sup>: Another integrated product for network analysis and visualization. It has been used in the SNA domain.
- <sup>28</sup> NetMiner<sup>28</sup>: A tool for exploratory network data analysis and visualization. NetMiner allows to explore network data visually and interactively, and helps in detecting underlying patterns and structures of the network.

<sup>20</sup> <http://ccl.northwestern.edu/netlogo/>

<sup>21</sup> <http://education.mit.edu/starlogo/>

<sup>22</sup> <http://repast.sourceforge.net/>

<sup>23</sup> [http://www.swarm.org/wiki/Main\\_Page](http://www.swarm.org/wiki/Main_Page)

<sup>24</sup> <http://www.analytictech.com/>

<sup>25</sup> <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>

<sup>26</sup> <http://cran.r-project.org/src/contrib/Descriptions/network.html>

<sup>27</sup> <http://www.orgnet.com/inflow3.html>

<sup>28</sup> <http://www.netminer.com/>

- <sup>29</sup> SocNetV<sup>29</sup>: A Linux based SNA and visualizing utility. SocNetV can compute network and actor properties, such as distances, centralities, diameter etc. Furthermore, it can create simple random networks (lattice, same degree, etc.).

Besides simulation and modeling toolkits there are APIs from Facebook, StumbleUpon, Technorati, del.icio.us, Digg, etc. These APIs could be used to download real-world data and study properties of social networks and concepts such as small worlds, random networks, scale-free networks, laws and distributions (normal distribution, Zipf's law, power law), search in networks, computation/propagation of influence and trust, diffusion (epidemics), robustness in networks, collective wisdom, collaborative filtering, social decision making, social criminals, individual profiling and privacy, story construction, provenance, and the unique characteristics of Long Tail blogs/blog sites and Short Head blogs/blog sites.

### 3.2 Methodologies

We now discuss the broad technical concepts that form the necessary background in conducting research in social network domain through centrality measures, network models, content analysis, link analysis, supervised learning, decision theoretic approach, and agent-based modeling.

Network centrality measures form an essential part of social network analysis. Social network analysis is used to identify leaders, mavens, brokers, groups, connectors (bridges between groups), mavericks, etc. Researchers have used several centrality measures to gauge the information flow across a social network which could help in identifying different roles of nodes mentioned above. Centrality measures help in studying the structural attributes of nodes in a network. They help in studying the structural location of a node in the network which could decide the importance, influence or prominence of a node in the network. Centrality measures help in estimating the extent to which the network revolves around a node. Different centrality measures include Degree centrality, Closeness centrality, Betweenness centrality, and Eigenvector centrality. Degree centrality refers to the total number of connections or ties a node has in the network. This could be imagined as a "hubness" value of that node. Rows or column sums of an adjacency matrix would give the degree centrality for that node. Closeness centrality is defined by the sum of all the geodesic distance of a node with all other nodes in the network. This could be imagined as the "nearest" node to the other nodes in the network. Betweenness centrality refers to the extent a node is directly connected to nodes that are not directly connected, or the number of geodesic paths that pass through this node. This evaluates how well a node can act as a "bridge" or intermediary between different subgraphs. A high betweenness centrality node can become a "broker" between different subgraphs. Eigenvector centrality defines a node to be central if it is connected to those who are central. This could be gauged as the "authoritativeness" of a node. It is the principal eigenvector of the adjacency matrix of the network. Other SNA measures used for analyzing social networks are clustering coefficient (the likelihood that associates of a nodes are associates among themselves to ensure greater cliquishness), cohesion (extent to which the actors are con-

<sup>29</sup> <http://socnetv.sourceforge.net/>

nected directly to each other), density (proportion of ties of a node to the total number of ties this node's friends have), radiality (extent to which an individual's network reaches out into the network and provides novel information), and reach (extent to which any member of a network can reach other members of the network).

Useful concepts in modeling social networks demand a lucid understanding of various network models such as scale-free, random [49], preferential attachment [6], hybrid [44], cascade models, etc. The link structure could be modeled using the scale free power law distribution ( $P(k) \propto k^{-\alpha}$ ). There are two generic aspects of real networks (e.g., Social networks, Blog networks, World Wide Web, biological networks, etc.) that make scale-free power law models an appropriate choice as compared to random models. First, the number of nodes ( $N$ ) in the real networks is not static.  $N$  increases throughout the lifetime of the network and the new nodes attach to the vertices already present in the network. Second, the random network models assume that the probability that two vertices are connected is random and uniform. However, most real networks exhibit preferential connectivity. For example, a newly created webpage will be more likely to include links to well-known popular documents with already high connectivity. Thus the probability with which a new vertex connects to the existing vertices is not uniform; there is a higher probability that it will be linked to a vertex that already has a large number of connections. This property of scale free power law models is also known as preferential attachment models. Some works [44] have shown the relative importance of hybrid models in simulating social networks by determining the appropriate proportion of random and scale free networks. Information flow across the network could be studied with the help of cascade models. Information diffusion could be considered analogous to the spread of a viral disease. Models from epidemiology have been borrowed and studied to model diffusion aspect in social networks. The key is to exploit different properties of scale-free, random, preferential attachment, hybrid models, cascade models to efficiently and effectively model the social networks.

Blogs have rich textual content. Not only people create new content, they also enrich the existing content by providing meta data such as labels and tags. These human-generated tags are also called "folksonomies". State-of-the-art content analysis techniques could be used for basic clustering, classification of the blog posts/blog sites. Traditional text analysis approaches like tf-idf could be used for indexing the blog entries. Folksonomies could be considered as class labels and supervised machine learning could be performed, classification models could be learned on labeled dataset, and learned models could be used to predict the tags of unlabeled corpus. This forms an essential concept for semi-automatically generating "tag-clouds" with least human intervention.

Link analysis helps in understanding several interesting phenomena of social networks. Text around the links give us knowledge about the linked blog posts. Based on the links, hubs and authorities could be discovered. This could be achieved exactly the same way as it is done for webpages. This approach could lead to the identification of expert communities. Several researchers have also pointed out the sparsity in the link structure of social networks which makes it different from the World Wide Web model. Many of them

like Blogosphere assume implicit link information among bloggers. Links could be constructed using the topic analysis. For example, blog posts talking about same topic could be connected. Supervised learning algorithms could be used to predict topics of unlabeled blog posts, which helps achieve link construction.

Several studies have been conducted to study decision theoretic approach for group-individual interaction and the effect of decision on an individual and/or a community as a whole. Decision theory studies what is the best possible decision to take given a fully informed decision maker. In the context of social networks this could be applied in finding the node in the network that is the best to make decisions with least possible side-effects and maximum possible gains for the rest of the nodes. This is a classic subject of study in microeconomics. Some decisions are difficult to make because of the need to reach a consensus among other members and the uncertainty in the response of different individuals. The analysis of such social decisions is dealt through game theory.

Social networks are also studied from the perspective of agent-based modeling. Basically each node in a social network can be treated as an agent. This agent could be a blogger in the blogosphere domain. Then assuming the network follows some distribution, usually a scale-free model, we can model the decision making ability of the agent probabilistically. This can help us in studying the factors that affect his/her blogging behavior, what and how (s)he makes decisions, etc. Neural networks or genetic algorithms could also be used to train the model of these agents to closely simulate some real-world scenario, which means, iteratively tuning the model parameters and keep improving the model.

### 3.3 Data Collection

Data collection is an essential part of studying and evaluating concepts in empirical research. Since social networking is a socio-psychological phenomenon and is more prevalent in the real world than the theoretical study, so enormous amounts of data exist on actual social networking websites. Moreover, since this involves user information, it is sensitive when the data is used in open research. Much of such data is unavailable due to privacy concerns. A few available datasets are:

- <sup>2</sup> Nielsen Buzzmetrics dataset: The dataset consists of about 14M blog posts from 3M blog sites collected by Nielsen BuzzMetrics<sup>30</sup> in May 2006. The data is annotated with 1.7M blog-blog links. However, up to a half of the blog outlinks are missing. Only 51% of the total blog posts are in English.
- <sup>2</sup> Enron Email dataset: It contains data from about 150 users, mostly senior management of Enron. The corpus<sup>31</sup> contains a total of about 0.5M messages. People have studied the social networks between users based on link construction. Links are constructed based on email senders and recipients.
- <sup>2</sup> APIs: APIs provided by Facebook, Digg, StumbleUpon, Technorati, del.icio.us etc. can be used to download data from a corresponding social networking website. Nevertheless, the API usage is often restricted

<sup>30</sup> <http://www.nielsenbuzzmetrics.com/>

<sup>31</sup> <http://www.cs.cmu.edu/~enron/>

to either last 30 days or top 100 results or in case of friendship networks like Facebook, one can only download data of his/her network of friends.

There is another more challenging yet appropriate option to obtain datasets. People can write crawlers and parsers to download data from blog sites. These custom datasets can be downloaded and pre-processed to serve more specific needs. We discuss more in Section 4.

### 3.4 Experiments and Performance Metrics

The fact that many concepts like, influence, trust, information propagation, identification of information routers, brokers, etc. in social network domain like Blogosphere are socio-psychological and highly subjective in nature, setting up experiments and evaluating the results is non-trivial. The absence of ground truth makes it even harder to compare different approaches available in the spectrum. Lack of ground truth makes an option to use search engines' ranking algorithms as the baseline for most of the existing works, even though theoretically it has been proven that current search engines are not suited for social network data and link structure. Recent work like [4] have used another Web 2.0 application, i.e., Digg, to evaluate the influence in the blogosphere. Results and discussion are included in Section 4.

## 4. A CASE STUDY

Here we present a study of identifying influential bloggers in a community [4]. We discuss model development, data collection and model tuning and verification through experiments.

### 4.1 Model Development

Assuming the domain to be community or multi-authored blogs, the influential bloggers are defined as: A blogger can be influential if s/he has more than one influential blog post. The model assigns an influence score to each blog post of the blogger. These blog post level influence scores are used to calculate the influence of the blogger.

An initial set of intuitive properties is proposed in [4] to approximately represent influential blog posts.

- <sup>2</sup> 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, or its number of inlinks ( $\eta$ ) is large. 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.
- <sup>2</sup> 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 ( $\epsilon$ ) 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 [33; 40].
- <sup>2</sup> Novelty - Novel ideas exert more influence as suggested in [30]. Hence, the number of outlinks is an indicator

of a post's novelty. A large number of outlinks ( $\mu$ ) may suggest that a post refers to many other blog posts or articles, indicating that it is less likely to be novel. Correlation experiments in [4] have reported that the number of outlinks is negatively correlated with the number of comments which means more outlinks reduces people's attention.

- <sup>2</sup> Eloquence - An influential is often eloquent [30]. 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. Correlation experiments in [4] have reported that the blog post length is positively correlated with number of comments which means longer blog posts attract people's attention.

### 4.2 Data Collection

Data collection is one of the critical tasks in this work. Since 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<sup>32</sup> 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<sup>33</sup>. 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 so far<sup>34</sup>. 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<sup>35</sup>.

### 4.3 Verification

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. Using the number of posts of a blogger

<sup>32</sup> <http://www.nielsenbuzzmetrics.com/cgm.asp>

<sup>33</sup> <http://technorati.com/developers/api/cosmos.html>

<sup>34</sup> January 31, 2007.

<sup>35</sup> This dataset will be made available upon request for research purposes.









