**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 5.258**

***IJCSMC, Vol. 5, Issue. 3, March 2016, pg.467 – 481***

# SURVEY ON SOCIAL NETWORK MINING

**Mohan.I**[1], **Pradeep Kumar.S**[2], **Gokula Krishnan.S**[3], **Aravindhan.S**[4]

Assistant Professor, Information Technology, Prathyusha Engineering College, India.[1]

Student, Information Technology, Prathyusha Engineering College, India.[2]

Student, Information Technology, Prathyusha Engineering College, India.[3]

Student, Information Technology, Prathyusha Engineering College, India.[4]

***Abstract: Social network has gained remarkable attention in the last decade. Accessing social network sites such as Twitter, Facebook LinkedIn and Google+ through the internet and the web 2.0 technologies has become more affordable. People are becoming more interested in and relying on social network for information, news and opinion of other users on diverse subject matters. The heavy reliance on social network sites causes them to generate massive data characterized by three computational issues namely; size, noise and dynamism. These issues often make social network data very complex to analyse manually, resulting in the pertinent use of computational means of analysing them. Data mining provides a wide range of techniques for detecting useful knowledge from massive datasets like trends, patterns and rules [44]. Data mining techniques are used for information retrieval, statistical modelling and machine learning. These techniques employ data pre-processing, data analysis, and data interpretation processes in the course of data analysis. This survey discusses different data mining techniques used in mining diverse aspects of the social network over decades going from the historical techniques to the up-to-date models, including our novel technique named TRCM. All the techniques covered in this survey are listed in the Table.1 including the tools employed as well as names of their authors.***

## 1. INTRODUCTION

A social network is a social structure made up of individuals called nodes, which are tied by one or more specific types of interdependency such as friendship, kinship, common interest, dislikes, beliefs. Social network analysis examines the structure of social relationships in a group to uncover the informal connection between people. Social network analysis is based on an assumption of the importance of relationships among interacting units. It indicates the way in which they are connected through different social familiarities ranging from casual acquaintances to close familiar bonds.[1] The social network perspective encompasses theories, models, and applications that are expressed in terms of relational concepts or processes. Along with growing interest and increased use of network analysis has come a consensus about the central principles underlying the network perspective.

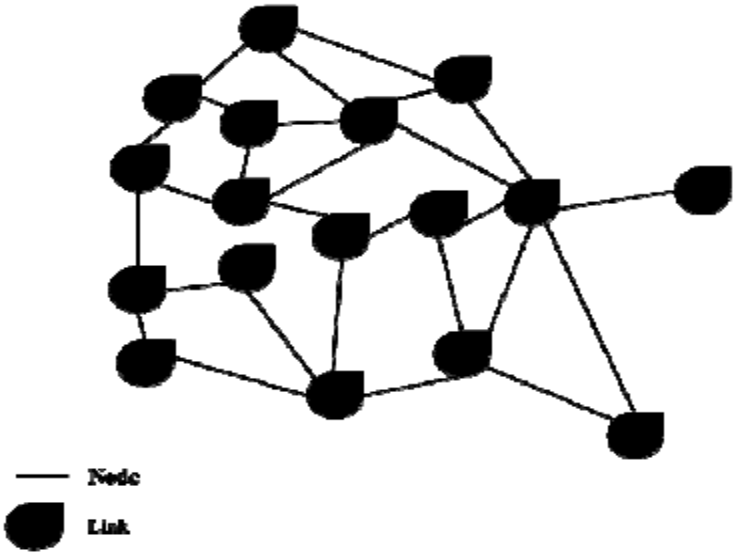



**Fig. 1. Social Network showing nodes and links**

In addition to the use of relational concepts, we note the following as being important: Actors and their actions are viewed as interdependent rather than independent, autonomous units  Relational ties (linkages) between actors are channels for transfer or "flow" of resources (either material or nonmaterial) Network models focusing on individuals view the network structural environment

as providing opportunities for or constraints on individual action Network models conceptualize structure (social, economic, political, and so forth) as lasting patterns of relations among actors. Most recently, SNA has become an important tool for organizational consultants seeking to understand connection between pattern and interactions and business outcomes such as job performance.

## 2. Social Media

Social media (Kaplan and Haenlein [28]) is defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchanges of user-generated content. Social media is conglomerate of different types of social media sites including traditional media such as newspaper, radio, and television and nontraditional media such as Facebook, Twitter, etc. Table 1 shows characteristics of different types of social media. Social media gives users an easy-to-use way to communicate and network with each other on an unprecedented scale and at rates unseen in traditional media. The popularity of social media continues to grow exponentially, resulting in an evolution of social networks, blogs, microblogs, location-based social networks (LBSNs), wikis, social bookmarking applications, social news, media (text, photo, audio, and video) sharing, product and business review sites, etc.

Facebook,1 the social networking site, recorded more than 845 million active users as of December 2011. This number suggests that China (approximately 1.3 billion) and India (approximately 1.1 billion) are the only two countries in the world that have larger populations than Facebook. Facebook and Twitter have accrued more than 1.2 billion users,2 more than thrice the population of the United States and more than the population of any continent except Asia.

TABLE 1. Characteristics of different types of social media.

| Type | Characteristics |
|---|---|
| Online social networking | Online social networks are Web-based services that allow individuals and communities to connect with real-world friends and acquaintances online. Users interact with each other through status updates, comments, media sharing, messages, etc. (e.g., Facebook, Myspace, LinkedIn). |
| Blogging | A blog is a journal-like website for users, aka bloggers, to contribute textual and multimedia content, arranged in reverse chronological order. Blogs are generally maintained by an individual or by a community (e.g., Huffington Post, Business Insider, Engadget). |
| Microblogging | Microblogs can be considered same a blogs but with limited content (e.g., Twitter, Tumblr, Plurk). |
| Wikis | A wiki is a collaborative editing environment that allow multiple users to develop Web pages (e.g., Wikipedia, Wikitravel, Wikihow). |
| Social news | Social news refers to the sharing and selection of news stories and articles by community of users (e.g., Digg, Slashdot, Reddit). |
| Social bookmarking | Social bookmarking sites allow users to bookmark Web content for storage, organization, and sharing (e.g., Delicious, StumbleUpon). |
| Media sharing | Media sharing is an umbrella term that refers to the sharing of variety of media on the Web including video, audio, and photo (e.g., YouTube, Flickr, UstreamTV). |
| Opinion, reviews, and ratings | The primary function of such sites is to collect and publish user-submitted content in the form of subjective commentary on existing products, services, entertainment, businesses, places, etc. Some of these sites also provide products reviews (e.g., Epinions, Yelp, Cnet). |
| Answers | These sites provide a platform for users seeking advice, guidance, or knowledge to ask questions. Other users from the community can answer these questions based on previous experiences, personal opinions, or relevent research. Answers are generally judged using ratings and comments (e.g., Yahoo! answers, WikiAnswers). |

## 3. Network Model

In May 2011, Facebook had 721 million users, represented by a graph of 721million nodes. A Facebook user at the time had an average of 190 friends; that is, all Facebook users, taken into account, had a total of 68.5 billion friendships (i.e., edges). What are the principal underlying processes that help initiate these friendships? More importantly, how can these seemingly independent friendships form this complex friendship network? In social media, many social networks contain millions of nodes and billions of edges. These complex networks have billions of friendships, the reasons for existence of most of which are obscure. Humbled by the complexity of these networks and the difficulty of independently analyzing each one of these friendships, we can design models that generate, on a smaller scale, graphs similar to real-world networks. On the assumption that these models simulate properties observed in real-world networks well, the analysis of real-world networks boils down to a cost-efficient measuring of different properties of simulated networks.

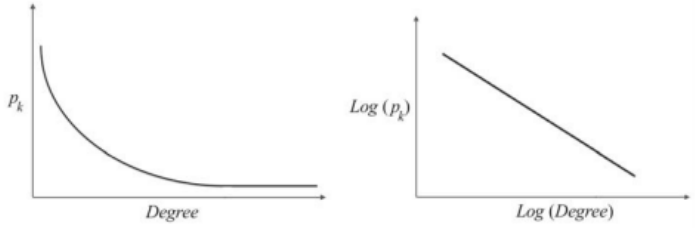
## 4. Properties of Real-World Networks

In addition, Real-world networks share common characteristics. When designing network models, we aim to devise models that can accurately describe these networks by mimicking these common characteristics. To determine these characteristics, a common practice is to identify their attributes and show that measurements for these attributes are consistent across networks. In particular, three network attributes exhibit consistent measurements across real-world networks: degree distribution, clustering coefficient, and average path length. As we recall, degree distribution denotes how node degrees are distributed across a network. The clustering coefficient measures transitivity of a network. Finally, average path length denotes the average distance (shortest path length) between pairs of nodes. We discuss how these three attributes behave in real-world networks next.

- **Degree Distribution**

Consider the distribution of wealth among individuals. Most individuals have an average amount of capital, whereas a few are considered extremely wealthy. In fact, we observe exponentially more individuals with an average amount of capital than wealthier ones. Similarly, consider the population of cities. A few metropolitan areas are densely populated, whereas other cities have an average population size. In social media, we observe the same phenomenon regularly when measuring popularity or interestingness for entities. For instance,

- Many sites are visited less than a thousand times a month, whereas a few are visited more than a million times daily.
- Most social media users are active on a few sites, whereas a few individuals are active on hundreds of sites.
- There are exponentially more modestly priced products for sale compared to expensive ones.



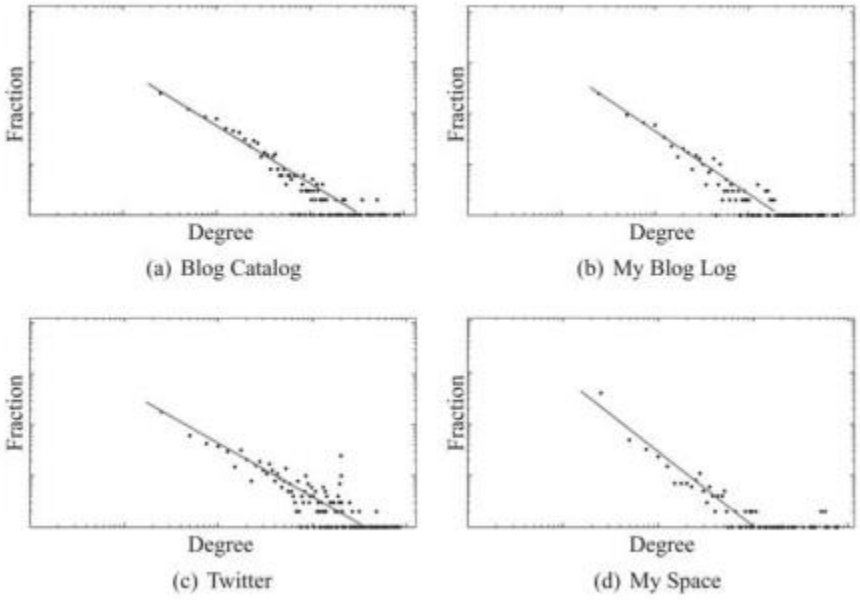

(a) Power-Law Degree Distribution (b) Log-Log Plot of Power-Law Degree Distribution

Figure 4.1: Power-Law Degree Distribution and Its Log-Log Plot.

- There exist many individuals with a few friends and a handful of users with thousands of friends.

The last observation is directly related to node degrees in social media. The degree of a node in social media often denotes the number of friends an individual has. Thus, the distribution of the number of friends denotes the degree distribution of the network. It turns out that in all provided observations, the distribution of values follows a power-law distribution.




- Pick a popularity measure and compute it for the whole network.

For instance, we can take the number of friends in a social network as a measure. We denote the measured value as k.

- Compute pk, the fraction of individuals having popularity k.
- Plot a log-log graph, where the x-axis represents ln k and the y-axis represents ln pk.
- If a power-law distribution exists, we should observe a straight line in the plot.

| Web | Facebook | Flickr | LiveJournal | Orkut | YouTube |
|---|---|---|---|---|---|
| 0.081 | 0.14 (with 100 friends) | 0.31 | 0.33 | 0.17 | 0.13 |

Networks exhibiting power-law degree distribution are often called scale-free networks. Since the majority of social networks are scale-free, we are interested in models that can generate synthetic networks with power-law degree distribution.

- **Clustering Coefficient**

In real-world social networks, friendships are highly transitive. In other words, friends of an individual are often friends with one another. These friendships form triads of friendships that are frequently observed in social networks. These triads result in networks with high average [local] clustering coefficients. In May 2011, Facebook had an average clustering coefficient of 0.5 for individuals who had two friends; their degree was 2 [284]. This indicates that for 50% of all users with two friends, their two friends were also friends with each other. Table 4.1 provides the average clustering coefficient for several real-world social networks and the web.

- **Average Path Length**

In real-world networks, any two members of the network are usually connected via short paths. In other words, the average path length is small. This is known as the small-world phenomenon. In the well-known small-world experiment conducted in the 1960s by Stanley Milgram, Milgram conjectured that people around the world are connected to one another via a path of at most six individuals (i.e., the six degrees of separation). Similarly, we observe small average path lengths in social networks. For example, in May 2011, the average path length between individuals in the Facebook graph was 4.7. This average was 4.3 for individuals in the United States at the same time [284]. Table 4.2 provides the average path length for real-world social networks and the web.

- **Random Graphs**

We start with the most basic assumption on how friendships can be formed: Edges (i.e., friendships) between nodes (i.e., individuals) are formed randomly. Small-world and Six The random graph model follows this basic assumption. In reality Degrees of Separation friendships in real-world networks are far from random. By assuming random friendships, we simplify the process of friendship formation in real-world networks, hoping that these random friendships ultimately create networks that exhibit common characteristics observed in real-world networks. Formally, we can assume that for a graph with a fixed number of nodes n, any of the n 2 edges can be formed independently, with probability p. G(n, p) This graph is called a random graph and we denote it as the G(n, p) model. This model was first proposed independently by Edgar Gilbert [100] and Solomonoff and Rapoport [262]. Another way of randomly generating graphs is to assume that both the number of nodes n and the number of edges m are fixed. However, we need to

determine which m edges are selected from the set of n 2 possible edges. Let Ω denote the set of graphs with n nodes and m edges. To generate a random graph, we can uniformly select one of the graphs in Ω. The number of graphs with n nodes and m

$$|\Omega| = \binom{\binom{n}{2}}{m}.$$

The uniform random graph selection probability is 1 |Ω| . One can think of the probability of uniformly selecting a graph as an analog to p, the probability of selecting an edge in G(n, p). The second model was introduced by Paul Erdos and Alfred R ″ enyi [83] ′ and is denoted as the G(n, m) model. In the limit, both models act similarly. G(n, m) The expected number of edges in G(n, p) is n 2 p. Now, if we set n 2 p = m, in the limit, both models act the same because they contain the same number of edges. Note that the G(n, m) model contains a fixed number of edges; however, the second model G(n, p) is likely to contain none or all possible edges. Mathematically, the G(n, p) model is almost always simpler to analyze; hence the rest of this section deals with properties of this model. Note that there exist many graphs with n nodes and m edges (i.e., generated by G(n, m)). The same argument holds for G(n, p), and many graphs can be generated by the model. Therefore, when measuring properties in random graphs, the measures are calculated over all graphs that can be generated by the model and then averaged. This is particularly useful when we are interested in the average, and not specific, behavior of large graphs. In G(n, p), the number of edges is not fixed; therefore, we first examine some mathematical properties regarding the expected number of edges that are connected to a node, the expected number of edges observed in the graph, and the likelihood of observing m edges in a random graph generated by the G(n, p) process.

- **Small-World Model**

The assumption behind the random graph model is that connections in real-world networks are formed at random. Although unrealistic, random graphs can model average path lengths in real-world networks properly, but underestimate the clustering coefficient. To mitigate this problem, Duncan J. Watts and Steven Strogatz in 1997 proposed the small-world model. In real-world interactions, many individuals have a limited and often at least, a fixed number of connections. Individuals connect with their parents, brothers, sisters, grandparents, and teachers, among others. Thus, instead of assuming random connections, as we did in random

graph models, one can assume an egalitarian model in real-world networks, where people have the same number of neighbors (friends). This again is unrealistic; however, it models more accurately the clustering coefficient of real-world networks. In graph theory terms, this assumption is equiva- Regular Ring Lattice lent to embedding individuals in a regular network. A regular (ring) lattice is a special case of regular networks where there exists a certain pattern for how ordered nodes are connected to one another. In particular, in a regular lattice of degree c, nodes are connected to their previous c/2 and following c/2 neighbors. Formally, for node set V = {v1, v2, v3, . . . , vn}

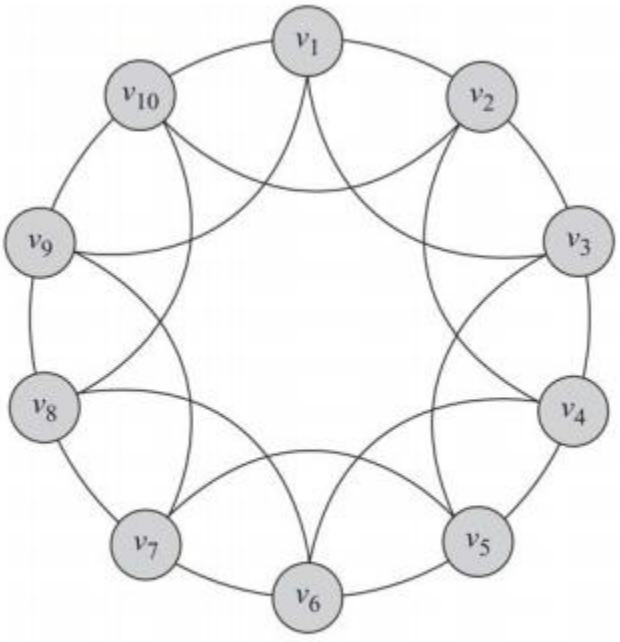



Figure 4.4: Regular Lattice of Degree 4.

**These models**

- **Matrix to represent social relation:**

The most common form of matrix in social network analysis is very simple one composed of as many rows and columns as there are actors in the data set. The simplest and the most matrix is binary. That is if a tie is present, a one is entered in a cell but if there is no tie ,a zero is entered. This type of matrix is the starting of network analysis and is known as adjacency matrix.

- **Statistical model for analysis:**

This type of models spansover70years.since,1970,one of the major directions in the field was to model probabilities between relational ties between interacting units such as actors. Extensive introduction to earlier methods is provided by Wassermann and f Faust [3].

- **Using graphs to represent social relations:**

Network analysis uses one kind of graphical display that consists of nodes to represent actors and lines or edges to represent relations or ties. When sociologists borrowed this concept of graphing they renamed the graph as "sociograms"

## 5. Social Network Sites: A Definition

We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site. While we use the term ''social network site'' to describe this phenomenon, the term ''social networking sites'' also appears in public discourse, and the two terms are often used interchangeably. We chose not to employ the term ''networking'' for two reasons: emphasis and scope. ''Networking'' emphasizes relationship initiation, often between strangers. While networking is possible on these sites, it is not the primary practice on many of them, nor is it what differentiates them from other forms of computer-mediated communication (CMC). What makes social network sites unique is not that they allow individuals to meet strangers, but rather that they enable users to articulate and make visible their social networks. This can result in connections between individuals that would not otherwise be made, but that is often not the goal, and these meetings are frequently between ''latent ties'' (Haythornthwaite, 2005) who share some offline connection. On many of the large SNSs, participants are not necessarily ''networking'' or looking to meet new people; instead, they are primarily communicating with people who are already a part of their extended social network. To emphasize this articulated social network as a critical organizing feature of these sites, we label them ''social network sites.'' While SNSs have implemented a wide variety of technical features, their backbone consists of visible profiles that display an articulated list of Friends1 who are also users of the system. Profiles are unique pages where one can ''type oneself into being'' (Sunde´n, 2003, p. 3). After joining an SNS, an individual is asked to fill

out forms containing a series of questions. The profile is generated using the answers to these questions, which typically include descriptors such as age, location, interests, and an ‘‘about me’’ section. Most sites also encourage users to upload a profile photo. Some sites allow users to enhance their profiles by adding multimedia content or modifying their profile’s look and feel.

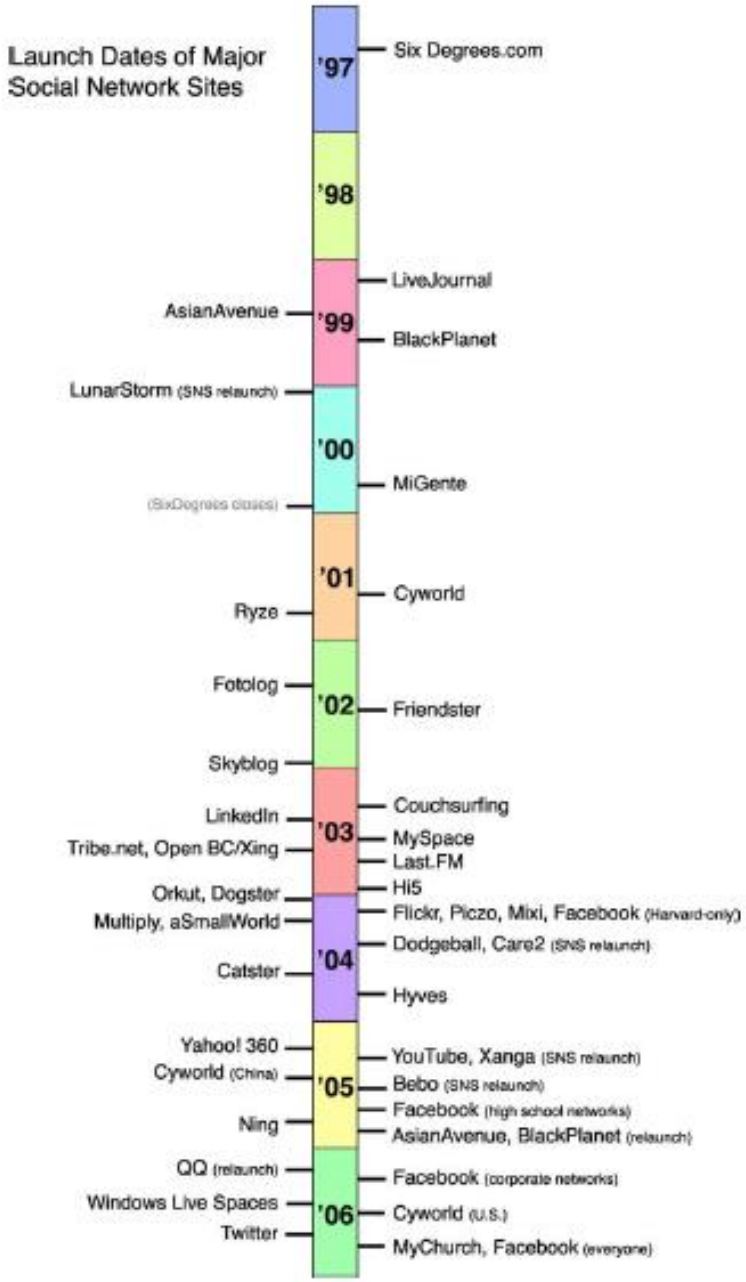



**Figure 1** Timeline of the launch dates of many major SNSs and dates when community sites re-launched with SNS features

Others, such as Facebook, allow users to add modules (‘‘Applications’’) that enhance their profile. The visibility of a profile varies by site and according to user discretion. By default, profiles on Friendster and Tribe.net are crawled by search engines, making them visible to anyone, regardless of whether or not the viewer has an account. Alternatively, LinkedIn controls what a viewer may see based on whether she or he has a paid account. Sites like MySpace allow users to choose whether they want their profile to be public or ‘‘Friends only.’’ Facebook takes a different approach—by default, users who are part of the same ‘‘network’’ can view each other’s profiles, unless a profile owner has decided to deny permission to those in their network. Structural variations around visibility and access are one of the primary ways that SNSs differentiate themselves from each other. After joining a social network site, users are prompted to identify others in the system with whom they have a relationship. The label for these relationships differs depending on the site—popular terms include ‘‘Friends,’’ ‘‘Contacts,’’ and ‘‘Fans.’’ Most SNSs require bi-directional confirmation for Friendship, but some do not. These one-directional ties are sometimes labeled as ‘‘Fans’’ or ‘‘Followers,’’ but many sites call these Friends as well. The term ‘‘Friends’’ can be misleading, because the connection does not necessarily mean friendship in the everyday vernacular sense, and the reasons people connect are varied (boyd, 2006a). The public display of connections is a crucial component of SNSs. The Friends list contains links to each Friend’s profile, enabling viewers to traverse the network graph by clicking through the Friends lists. On most sites, the list of Friends is visible to anyone who is permitted to view the profile, although there are exceptions. For instance, some MySpace users have hacked their profiles to hide the Friends display, and LinkedIn allows users to opt out of displaying their network. Most SNSs also provide a mechanism for users to leave messages on their Friends’ profiles. This feature typically involves leaving ‘‘comments,’’ although sites employ various labels for this feature. In addition, SNSs often have a private messaging feature similar to webmail. While both private messages and comments are popular on most of the major SNSs, they are not universally available. Not all social network sites began as such. QQ started as a Chinese instant messaging service, LunarStorm as a community site, Cyworld as a Korean discussion forum tool, and Skyrock (formerly Skyblog) was a French blogging service before adding SNS features. Classmates.com, a directory of school affiliates launched in 1995, began supporting articulated lists of Friends after SNSs became popular. AsianAvenue, MiGente, and BlackPlanet were early popular ethnic community sites with limited Friends functionality before re-launching in 2005–2006 with SNS features and structure.

## 6. A History of Social Network Sites

- **The Early Years**

According to the definition above, the first recognizable social network site launched in 1997. SixDegrees.com allowed users to create profiles, list their Friends and, beginning in 1998, surf the Friends lists. Each of these features existed in some form before SixDegrees, of course. Profiles existed on most major dating sites and many community sites. AIM and ICQ buddy lists supported lists of Friends, although those Friends were not visible to others. Classmates.com allowed people to affiliate with their high school or college and surf the network for others who were also affiliated, but users could not create profiles or list Friends until years later. SixDegrees was the first to combine these features. SixDegrees promoted itself as a tool to help people connect with and send messages to others. While SixDegrees attracted millions of users, it failed to become a sustainable business and, in 2000, the service closed. Looking back, its founder believes that SixDegrees was simply ahead of its time (A. Weinreich, personal communication, July 11, 2007). While people were already flocking to the Internet, most did not have extended networks of friends who were online. Early adopters complained that there was little to do after accepting Friend requests, and most users were not interested in meeting strangers. From 1997 to 2001, a number of community tools began supporting various combinations of profiles and publicly articulated Friends. AsianAvenue, BlackPlanet, and MiGente allowed users to create personal, professional, and dating profiles users could identify Friends on their personal profiles without seeking approval for those connections (O. Wasow, personal communication, August 16, 2007).

- **Future Research**

The work described above and included in this special theme section contributes to an on-going dialogue about the importance of social network sites, both for practitioners and researchers. Vast, uncharted waters still remain to be explored. Methodologically, SNS researchers' ability to make causal claims is limited by a lack of experimental or longitudinal studies. Although the situation is rapidly changing, scholars still have a limited understanding of who is and who is not using these sites, why, and for what purposes, especially outside the U.S. Such questions will require large-scale quantitative and qualitative research. Richer, ethnographic research on populations more difficult to access (including non-users) would further aid scholars' ability to understand the long-term implications of these tools. We hope that the work described here and included in this collection will help build a foundation for future investigations of these and other important issues surrounding social network sites.

## 7. CONCLUSION

This paper provides a more current evaluation and update of social network analysis research available. Literatures have been reviewed based on different aspects of social network analysis. Survey on recent works in the field of social network analysis depicts that different research exposures are there in the field of social network analysis. Recent trends on research are in area of link analysis, dark web analysis, and spam behavior detection.

## References

[1] M. A. Abbasi, S. Kumar, J. A. Andrade Filho, and H. Liu. Lessons learned in using social media for disaster relief—ASU Crisis Response Game. Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer-Verlag, Berlin, 282–289, 2012.

[2] N. Agarwal and H. Liu. Modeling and Data Mining in Blogosphere. Morgan & Claypool Publishers, San Rafael, CA, 2009.

[3] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. Proceedings of the International Conference on Web Search and Web Data Mining. Association for Computing Machinery, New York, 207–218, 2008.

[4] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Modeling blogger influence in a community. Social Network Analysis and Mining 2(2):139–162, 2012.

[5] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences of the United States of America 106(51):21544, 2009.

[6] D. Artz and Y. Gil. A survey of trust in computer science and the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 5(2):58–71, 2007.

[7] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. Proceedings of the Fourth ACM

International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, 635–644, 2011.

[8] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, 44–54, 2006.

[9] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin. Persona: An online social network with user-defined privacy. ACM SIGCOMM Computer Communication Review 39(4):135–146, 2009.

[10] N. T. J. Bailey. The Mathematical Theory of Infectious Diseases and Its Applications. Charles Griffin, High Wycombe, UK, 1975.

[11] E. Berger. Dynamic monopolies of constant size. Journal of Combinatorial Theory, Series B 83(2):191–200, 2001.

[12] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on Twitter: Human, bot, or cyborg? Proceedings of the 26th Annual Computer Security Applications Conference. Association for Computing Machinery, New York, 21–30, 2010.