

# Probabilistic Generative Model of Social Network Based on Web Features

Mahyuddin K. M. Nasution<sup>1,2</sup> and Shahrul Azman Noah<sup>2</sup>

<sup>1</sup> Department of Mathematics FMIPA USU, Medan, Indonesia  
mahyunst@yahoo.com, mahyuddin@usu.ac.id

<sup>2</sup> Knowledge Technology Group Research FTSM UKM, Bangi, Malaysia  
samm@ftsm.ukm.my

**Abstract.** In this paper, we develop a dynamic framework for the modeling and analysis of social networks to work with web documents. We illustrate the model with features of web, design a form to analyze relationships of attributes as a modality of social structure, and create the optimization of generative model based on Bayes Theorem.  
Keywords: conditional probability, graph, similarity, singleton, doubleton, framework.

## 1 Introduction

Social network describes a group of social entities and the pattern of inter-relationships among them. The concept of social networks is designed to map the relationship of entities among all of them that can be observed, to mark the patterns of ties between entities, to measure social capital: the values obtained by the entities individually or in groups, to present a variety of social structures according to the interests and its implementation, based on different domains or information sources [1]. Group discovery has many applications, such as understanding the social structure of organizations or native tribes. In law enforcement this is about organized crimes such as drugs and money laundering [2] or terrorism [3], knowing how the perpetrators are connected to one another would assist the effort to disrupt a criminal act or to identify additional suspects. In commerce, viral marketing exploits the relationship between existence of customers and potential customers to increase sales of products and services [4,5]. Members of a social network may also take advantage of their connections to get to know others, for instance through web sites facilitating networking or dating among their users [6].

Social networks explicitly exhibit relationships (called *ties* in social sciences) among individuals and groups (called *actors*). They have studied social sciences since the 1930s. Social scientists have conducted extensive research on group detection, especially in fields such as anthropology and political science. Recently, statisticians and computer scientists have begun to develop models that specifically discover group memberships [7,8,9]. There are two models use probability for characterizing information sources as well as image, has become a tool in machine learning research: probabilistic generative and relative models. Therefore, the approaches that addresses the issues of social network generally fall into two categories also, and this paper explores their relationship in order to obtain a framework to investigate social networks by engaging web features.

## 2 Related Work and Problem

Technically, an object called the network is a graph. A graph consist of a set of points along with a set of lines connecting pairs of points. Formally, a graph denoted by  $G(V, E)$ , where  $V \neq \emptyset$  is a set of vertices or nodes,  $V = \{v_i | i = 1, \dots, I\}$  and  $E$  is a set of edges connects between pair of vertices,  $E = \{e_j | j = 1, \dots, J\}$ . The nodes in a social network refer to actor names such as authors, recipients, researchers, artists, politicians, firms, organizations, or any entity, i.e.,

**Definition 1.** *A set of actors  $A = \{a_i | i = 1, \dots, I\}$  and there is a function  $\xi$  such that  $\xi : A \xrightarrow{1:1} V$ , or  $\forall a \in A \exists! v \in V$ .*

Each actor plays some role in the social interactions based on his/her background and then achievements obtained in each occasion, or kinds of: new articles and academic publications. In many situations, these are considered as attributes or characteristics, and the characteristic can be defined as follows.

**Definition 2.** *Let  $Z = \{z_k | k = 1, \dots, K\}$  is a set of attributes, and a pair of  $\langle A, Z \rangle$  is the instance of actors, where  $Z_i$  is subsets of  $Z$ ,  $Z_i$  are subsets of attributes of each actor  $a_i$ , i.e.,  $\langle a_i, Z_i \rangle$ ,  $i = 1, \dots, I$ , simply we denotes a set of attributes of actor  $a$  as  $Z_a$ .*

A social network is a network based on the relations between people in their society. Therefore, we can model an approach of other social network. When a computer network connects people, it is a social network [10]. Just as a computer network is a set of machines connected by a set of media (cables or airwaves), a social network also is a set of people connected by a set of social relationships such as friendship, co-working or information exchange such as Web. Information of people in web documents is very different from information of people in database. In any documents, the objects (actors and attributes) can be given literally, like the literal text of *Indonesia*, then all meaning of object based on words represented by the literal objects itself. To realize it, first we define that a word  $w$  is the basic unit of discrete data, defined to be an item from a vocabulary indexed by  $\{1, \dots, K\}$ , where  $w_k = 1$  if  $k \in K$ , and  $w_k = 0$  otherwise. Then, we define some instances related to words.

**Definition 3.** *A document is a sequence of  $n$  words denoted by  $D = \{w_1, \dots, w_N\}$ , where  $w_n$  is the  $n$ -th word in the sequence. Size of document is a cardinality of  $D$ , i.e.,  $|D| = N$ .  $\mathcal{D} = \{D_1, \dots, D_M\}$  is a collection of  $M$  documents that is called a corpus.*

**Definition 4.** *A term  $t_k$  consist of at least one word, or  $t_k = (w_1, \dots, w_l)$ ,  $l \leq k$ ,  $k$  is a number of parameters representing words.  $|t_k| = k$  is the number of word of  $t_k$ , and  $l$  is the number of vocabularies in  $t_k$ .*

Any information available on the web may be obtained with the help of search engines. The search engine is an important part of internet and is one of the easiest and useful tools to research information or to find websites. A search engine allows to categorize and make sense of the information that is available online. The search results generally presented in a list of results and called as *hits*, where information may consist of texts, images, video, hypertext, etc.

**Definition 5.** *Let the set of Web pages indexed by search engine be  $\Omega$ , i.e., a set contains ordered pair of the term  $t_{k_i}$  and the web page  $\omega_{k_j}$ ,  $(t_{k_i}, \omega_{k_j})$ ,  $i = 1, \dots, I$ ,  $j = 1, \dots, J$ . The relation table that consist of two columns  $t_k$  and  $\Omega_k$ , the table is a representation of  $(t_{k_i}, \omega_{k_j})$ , where  $\Omega_k = \{(t_k, \omega_k)_{ij}\} \subseteq \Omega$  or  $\Omega_k = \{\omega_{k_1}, \dots, \omega_{k_j}\}$ . The cardinality of  $\Omega$  is denoted by  $|\Omega|$ , and uniform mass probability function is  $P : \Omega \rightarrow [0, 1]$ .*

**Definition 6.** Let  $t_x$  is a search term, and  $t_x \in \mathcal{S}$  where  $\mathcal{S}$  is set of singleton search term of search engine. A vector space  $\mathbf{x} \subseteq \Omega$  is a singleton search engine event of Web pages that contain an occurrence of  $t_x \in \omega_x$ , and probability of an event  $\mathbf{x}$  is  $P(\mathbf{x}) = |\mathbf{x}|/|\Omega| \in [0, 1]$ .

**Proposition 1.** Let two singleton events  $\Omega_x$  and  $\Omega_y$  for search terms  $t_x$  and  $t_y$  respectively.  $\Omega_x \cap \Omega_y$  is a doubleton event of  $t_x$  and  $t_y$  such that  $P(\mathbf{x}, \mathbf{y}) = |\Omega_x \cap \Omega_y|/|\Omega|$ .

*Proof.* By using intersection operator of set to Definition 6, we have a direct conclusion, i.e.,  $P(\mathbf{x}) = |\Omega_x|/|\Omega|$  dan  $P(\mathbf{y}) = |\Omega_y|/|\Omega| \Rightarrow P(\mathbf{x}, \mathbf{y}) = |\Omega_x \cap \Omega_y|/|\Omega|$ .

A search for a particular index term, say  $t_x$ , it returns a certain number of hits  $n_x$ , i.e., number of web pages where this term occurred, we obtain  $p(t_x) = n_x/|\Omega|$ . So for  $(t_x, t_y)$  we have a doubleton  $n_{xy}$ , and  $n_x \geq n_{xy}$ .  $n_{xy}/|\Omega|$  means that  $p(t_x|t_y)$  or  $p(t_y|t_x)$  are the conditional probabilities. It is clear that a doubleton is a conditional probability of a term for other term.

We note that in the conditional probabilities the total number of web pages indexed by search engine,  $|\Omega|$  is divided out. Therefore, the conditional probabilities are independent of  $|\Omega|$ . The conditional probability  $P(\mathbf{x}|\mathbf{y}) > 0$  means that the search terms  $t_x$  and  $t_y$  occur together in some web pages or co-occurrence, but also parts of  $t_x$  or  $t_y$  occur together such that  $\mathbf{x}$  or  $\mathbf{y}$  are in bias.

The edges in a social network refer to ties. A tie relates two actors. Ties could be directed or undirected, and they could be dichotomous (present or absent) or valued (weighted). There may be many types of ties (e.g., kinship, friendship) and the collection of all ties of the same type is a relation. Relations, sometimes called strands, are characterized by content, direction and strength. Let  $R$  is a set of relations, the relations among actors formed by sharing attributes, ideas, concepts, etc, between them, which can be depicted as the intersection between their attributes [11] as follows

$$r_k(a, b) = Z_a \cap Z_b, r_k \in R. \quad (1)$$

The content of a relation refers to the source that is exchanged, such as communication about administrative, personal, work-related or

social matters. Communication, defined generally as transfer of information or resources, is common among socially related people whereby the electronic trails of communication can be traced include emails [12], newsgroups [13], and instant messaging [14]. In the content, sometime we find self-report, links reported by individual actors. Such links are directed and naturally subjective, such as in classical tools like questionnaires and interviews are based on this principle [15], homepages or profile pages in community-centric sites: LiveJournal weblogs [16] or Facebook [6] commonly display a self-professed list of friends within the community. Therefore, a relation can be directed or undirected: one person may give social support to a second person, or there are two relations here: giving support and receiving support. The relations also differ in strength. Such strength can be operationalized in a number of ways [17,18]. Therefore, the types of relations important in social network research, it have included the exchange of complex or difficult information [19], emotional support [19,20,21], uncertain or equivocal communication [22,23], and communication to generate ideas, created consensus [24,25,26,27,28,29,30], support work, forter sociable relations [31,32], or support virtual community [33]. In general, the strength relation generated by similarity measures, and some of them are dice, overlap, dan Jaccard [34,35,36] as follows

$$sim(x, y) = \frac{2n_{\mathbf{x} \wedge \mathbf{y}}}{n_{\mathbf{x}} + n_{\mathbf{y}}}, \quad (2)$$

$$sim(x, y) = \frac{n_{\mathbf{x} \wedge \mathbf{y}}}{\min(n_{\mathbf{x}}, n_{\mathbf{y}})}, \quad (3)$$

and

$$sim(x, y) = \frac{n_{\mathbf{x} \wedge \mathbf{y}}}{n_{\mathbf{x} \vee \mathbf{y}}}. \quad (4)$$

Similarity has its foundation on the sociological idea that friends tend to be alike [37]. Other forms of similarity include having the same communication partners [12] and sharing the same opinions or areas of interest [4]. Each similarity measure relative to the other, and also against the probability, therefore we call these measures as probabilistic relative model. We generally define it as

**Definition 7.** *Probabilistic relative model (PRM) utilizes the Cartesian product for clustering the nodes in  $A$ , i.e.,*

$$\gamma : A \times A \rightarrow R$$

*such that  $\gamma(a, b) \in R$ ,  $a, b \in A$ .*

However, this model not only adrift with a bias in the measurement but also is difficult to produce descriptions of the relationship. In the real world and its application, a social network requires the labels with a weight as a modality as well as explaining the roles of each actor in the social. We define an approach to generate labels as follows

**Definition 8.** *Probabilistic generative model (PGM) employs a function  $\lambda$  for classifying  $Z$ , i.e.,*

$$\lambda : Z \rightarrow C$$

*such that  $\lambda(z) = c$ ,  $z \in Z$ , and  $c \in C$  is a class of labels, where  $C = \{c_1, c_2, \dots, c_{|C|}\}$  is a data set as special target attributes,  $|C| \geq 2$  is the number of classes, and  $Z \cap C = \emptyset$ .*

Statistical natural language processing, an analysis that capture the richness of the language contents of the interactions: the words, the topics, and other high-dimensional specifics of the interactions between actors. Statistically, Bayes theorem has paid dividends in the computing world, especially for artificial intelligence and learning, in which the PGM has played a role in particular. Some PGMs are a direct offspring of Latent Dirichlet Allocation (LDA) [38], the Multi-label Mixture Model [39], and the Author-Topic Model [41,40], with the distinction that ART is specifically designed to capture language used in a directed network of correspondents. However, the PGMs concern with extraction of network based on predefined labels only, and thus cannot be adapted to the another description of relation.

### 3 The Concept of Probability Use

The Bayes approach defines the classification problem in terms of probabilities. There are three main concepts required are conditional

probability, Bayes theorem, and the bayes decision rule. The conditional probability  $(a|D)$ , which is used to define independent events of an actor to a document in a corpus, is defined by  $P(a|D) = P(a \cup D)/P(D)$ , where  $P(a|D)$  is the probability that event  $a \in A$  happens, given that  $D$  is observed. Similarly,  $P(D|a) = P(a \cup D)/P(a)$ , where  $P(D|a)$  is the probability that event  $D$  happens, given that  $a \in A$  is observed. It then follows (by substitution) that  $P(a \cap D) = P(a)P(D|a)$ . The premise of Bayes Theorem starts with an initial degree of belief that an event will occur, and then with new information about a degree of belief. These two degrees are reresented, respectively, by the prior probability  $P(a|D)$  and the posterior probability  $P(D|a)$ , which are related by

$$P(a|D) = \frac{P(a)P(D|a)}{P(D)} \quad (5)$$

The Bayes decision rule states that based on the posterior probabilities, it is possible to assign an element  $w$  to a class with the largest probability. For example, let  $w$  be a data sample (vector of features) and  $z_i$  one of the possible classes, then  $P(w|z_i)$  is prior probability, because it can be obtained based on prior knowledge.

**Proposition 2.** *Let  $P : D \rightarrow [0, 1]$  is a mass probability function whereby the probability of  $w_i$  is  $p(w_i) = 1/|D| \in [0, 1]$ . If  $w_{v_i}$  are the vocabularies in document, then  $p(w_{v_i}) = |w_{v_i}|/|D|$ , where  $|w_{v_i}|$  is number of  $w_i$  in  $D$ .*

When working with general web documents in general. We explore the features of documents in a corpus. Thus it can be stated that the probability  $P(w)$  is the conditional probability  $P(w|D)$  of event  $w$  to the document  $D$  in the corpus  $\mathcal{D}$ . Here are the events of  $w$  in the corpus.

**Lemma 1.** *Let probability of  $w_{v_i}$  in  $D$  is  $p(w_{v_i}) = |w_{v_i}|/|D|$ , then probability of  $w_{v_i}$  in  $\mathcal{D}$  is*

$$p(w_{v_i}) = \sum_{j=1}^M \frac{|w_{v_i}|}{M|D|} = \sum_{j=1}^M \sum_{k=1}^{|w_{v_i}|} \frac{1}{MN_j}.$$

*Set of  $p(w_{v_i})$ ,  $V = [p(w_{v_i})|i = 1, \dots, n]$  is a vector space where  $p(w_{v_i})$  is a weight of  $w_i$ .*

*Proof.* The direct consequence of Definition 3 and Proposition 2.

Let  $t_a$  is a search term, each search engine produces  $n_a$  as hit counts, there are as many  $n_a$   $ta \in \Omega$ , or for all  $t_a \in \omega_a$ ,  $|\Omega_a| = n_a$ ,  $t_a \in \Omega_a \subseteq \Omega$ . Let us define that  $\Omega = \sum_{j=1}^M N_j$ , where  $M$  is a number of web pages in  $\Omega$  and  $N_j$  are the number of words/terms in  $\omega_j$ . Assume that a set of web pages  $\Omega$  as corpus  $\mathcal{D}$ , then we obtain

$$P(t_a) = P(a) = |\Omega_a|/|\Omega| = \sum_{j=1}^M \sum_{k=1}^{|t_a|} \frac{1}{MN_j} \quad (6)$$

Therefore, we can define that probability of  $D \in \mathcal{D}$  is equivalent to probability of  $\omega \in \Omega$ .

**Proposition 3.** *Let  $p$  is a document  $D$  from the corpus  $\mathcal{D}$  with probability  $P(d)$ , then probability  $P(\omega) = \sum_{j=1}^M N_j/|\Omega|$*

**Definition 9.** *Let  $t_a$  is a term search.  $\mathcal{S}_a = \{S_i | i = 1, \dots, n\}$  is a list of snippets that are returned by a search engine for  $t_a$ , where a snippet  $S = \{w_j | j = 1, \dots, |S|\}$ , and  $|S| = \pm 50$  words.*

Each word in document or snippet can take the meaning by giving a vector space. One of methods is to use the  $tf \cdot idf$ , i.e.,

$$tf \cdot idf = tf(w) \cdot idf(w) = \left( \sum_{j=1}^M \sum_{i=1}^{|w_{v_i}|} \frac{1}{N_j} \right) \left( \log \frac{M}{df(w)} \right) \quad (7)$$

$df(w)$  is the number of document where the word  $w$  appearance. Normalization of  $tf \cdot idf$  is  $tf \cdot idf/h$ ,  $h$  is a highest score of  $tf \cdot idf$ .

**Proposition 4.** *If  $z_k$  is a latent class, the  $z_k$  relates to the document  $D$ .*

*Proof.* Each class of  $z_k$  contains a set of words  $\mathbf{w}_k$  as characteristics of class, and each  $\mathbf{w}$  with the values is a vector space whereby if there are  $k$  classes of words then intersection of classes is emptyset,  $\cap_{i \in I} \mathbf{w} = \emptyset$ . In  $D$ , based on Lemma 1  $\forall w \in D \exists! p(w) \in [0, 1]$ , so if  $\forall w \in \cup_{k=1}^K \mathbf{w}_k$  then  $\exists! p(w) \in [0, 1]$ . Thus, there is a possibility that  $z_k$  associated with  $D$ .



Specifically, a document  $D$  is potentially related to several topics  $\mathcal{Z}$  with different probabilities, i.e., a set of latent variables  $\mathcal{Z} = \{z_1, \dots, z_K\}$ . Therefore, by using same reason we obtain the following lemma, where latent variables consequently generate a set of words  $\mathbf{w}$ .

**Lemma 2.** *If  $z_k$  is a latent class, the word  $w$  can be generated with probability  $P(w|z_k)$ .*

**Proposition 5.** *If  $z_k$  is a latent class, the  $z_k$  relates to the actor  $a \in A$ .*

*Proof.* Each actor will use words to communicate with other actors in the social, some of communication be recorded in the document, so it will relate to the intended actor. When that word also be present in every  $z_k$ , then we obtain a relation between  $a$  and  $z_k$ .

The latent variables can generate name appearances  $A$  that are closely related to a specific topic.

**Lemma 3.** *If  $z_k$  is a latent class, the actor  $w$  can generated with probability  $P(a|z_k)$ .*

A concept treats document-name-word as a triplet  $\langle D, a, w \rangle$  is to represent an instance that a name  $a$  appears in document  $D$ , which contains the word  $w$ . The relationship inherent in this concept is associated by a set of topics  $\mathcal{Z}$ , where a set of latent variables  $z$  can break the direct relationships between documents, words, and names.

**Theorem 1.** *Let  $\mathcal{Z}$  is a set of latent variables,  $\mathcal{Z} = \{z_l | l = 1, \dots, L\}$ , with size  $L$ , each of which represents a latent topic, if and only if the relationships between  $D \in \mathcal{D}$ ,  $a \in A$ , and  $w \in D$  are connected by  $\mathcal{Z}$ .*

*Proof.* A direct result of some of previous lemmas.

## 4 The Framework

There are five steps of Bayesian methodology [42] to classify something based on Definition 8: (i) Collect data, and estimate parameters such as mean and coveriance for each class. In this case, assume

that all the probability density functions have a Gaussian behavior. (ii) Choose a set of features. (iii) Choose a mode and derive a decision rule with these parameters. (iv) Train the classifier and apply the decision rule by using a discriminant function, and apply it to a test data set to classify each sample. (v) Evaluate the decision rule. Measure the accuracy/error rate in order to improve the choice of features and the overall design of the classifier.

In the exploration of social networks that involves web pages, the ties is not just acquired from the hit count, but other features. There are features such as web-snippets that are not only composed a collection of words, a part of them refers to the person names, but there are also the addresses of the web page, URL. For each search term  $t_a$ , represents a name  $a \in A$ , the list of snippets  $\mathcal{S}$  will be obtained where  $a \in \mathcal{S} \subseteq \mathcal{S}$ . Therefore, we can consider the results returned by search engines based on the name as a probability  $p(\omega|a)$ . Furthermore, each list of snippets can be modeled as a bag of words to generate a vector space in which their weight is obtained by Equation (7), and this can be interpreted as the conditional probability, i.e.,  $P(w|a)$ ,  $P(w|S)$ , and  $P(w|\mathcal{S})$  which is meaningful as the current context [36] of each person name.

We can provide a latent class  $z$  by using the clusters of words from the snippets. For every  $a \in A$ ,  $\mathcal{S}_a$  is potentially a description of the actors and the relations between the actors by using Definition 6, Proposition 1 and Equation (4) so that the current context is in the trees as the optimal form of relationship between words and person name  $a \in A$ . In this case, once provided the topics for example based on the existing research group, the trees of words can be selected based on their proximity to the topics by involving singleton and doubleton. The selected tree of words will be a description of the latent variable  $z$ . Under this scenario, we can derive the aspect model that involves joint probability over  $D \times a \times w$  is expressed as mixture

$$P(S, a, w) = P(S)P(a, w|S) \quad (8)$$

$$P(a, w|S) = \sum_{z \in \mathcal{Z}} P(a, w|z)P(z|S) \quad (9)$$

The latent variables  $z$  connected each instance of which can be obtained from the list of snippets such as names, words and docu-

ments, but every instance that was also having relations with one another. Therefore, based on Theorem 1, the optional form of this relationship is only connected by a latent variables with disconnecting from each relation between instances, and an symmetric model can be parameterized by

$$P(S, a, w) = \sum_{z \in \mathcal{Z}} P(z)P(S|z)P(w|z)P(a|z). \quad (10)$$

## 5 Conclusion and Future Work

We have proposed a novel framework for acquiring the social networks from snippets. We will demonstrated this after getting the Expectation-Maximization (EM) of the model.

## References

1. Nasution, Mahyuddin K. M. and Noah, S. A. 2010. Extracting social networks from web documents. *The 1st National Doctoral Seminar on Artificial Intelligence Technology*. CAIT, Universiti Kebangsaan Malaysia, Bangi: 278-281.
2. Xu, J. and Chen, H. 2004. Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. *Decision Support Systems*, 28: 473-487.
3. Krebs, V. E. 2002. Mapping networks of terrorist cells. *Connections*, 24: 43-52.
4. Richardson, M. and Domingo, P. 2002. Mining knowledge-sharing sites for viral marketing. *KDD*: 61-70.
5. Kempe, D., Kleinberg, J. and Tardos, E. 2003. Maximizing the spread of influence through a social network. *KDD*: 137-146.
6. Boyd, D. M. 2004. Friendster and publicly articulated social networks. *ACM Conference on Human Factors in Computing Systems ACM Press*, Vienna, 2004: 1279-1282.
7. Nowicki, K. and Snijders, T.A.B. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077-1087.
8. Kubica, J., Moore, M., Schneider, J., and Yang, Y. 2002. Stochastic link and group detection. In *The 17th National Conference on Artificial Intelligence (AAAI)*.
9. Bhattacharya, I. and Getoor, L. 2004. Deduplication and group detection using links. In *The 10th SIGKDD Conference Workshop on Link Analysis and Group Detection (LinkKDD)*.
10. Garton, L., Haythornthwaite, C., and Wellman, B. 1997. Studying online social networks. *Journal of Computer Mediated Communication*, 3(1).
11. Nasution, Mahyuddin K. M. and Noah, S. A. 2011. Extraction of academic social network from online database. *2011 International Conference on Semantic Technology and Information Retrieval (STAIR'11)*, IEEE Computer Society: 64-69.

12. Schwartz, M F. and Wood, D. C. M. 1993. Discovering shared interests using graph analysis. *CACM*, 36: 78-89.
13. Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. 2003. Mining newsgroups using networks arising from social behavior. *WWW*: 688-703.
14. Resig, J, Dawara, S., Homan, C. M., and eredessai, A. 2004. Extracting social networks from instant messaging populations. In *Workshop on Link Analysis and Group Detection* (In conj. with KDD).
15. Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press: Cambridge.
16. Kumar, R., Novak, J., Raghavan, P., and Romkins, A. 2004. Structure and evolution of blogspace. *Communications of the ACM*, 47(12): 35-39.
17. Marsde, P., and Campbell, K. E. 1984. Measuring the strength. *Social Forces*, 63: 482-501.
18. Wellman, B., and Wortley, S. 1990. Different strokes from different folks: Community ties and social support. *American Journal of Sociology*, 96: 558-588.
19. Fish, R. S., Kraut, R. E., Root, R. W., and Rice, R. E. 1992. Evaluating video as a technology for informal communication. Paper presented at the Striking a Balance: *CHI'92 Conference Proceedings*, New York, NY.
20. Haythornthwaite, C., Wellman, B., and Mantei, M. 1995. Work relationships and media use: A social network analysis. *Group Decision and Negotiation*, 4(3): 193-211.
21. Rice, R. and Love, G. 1987. Electronic emotion: Socioemotional content in a computer-mediated communication network. *Communication Research*, 14(1): 85-108.
22. Daft, R. L., and Lengel, R. H. 1986. Organizational information requirements, media richness and structural design. *Management Science*, 32: 554-571.
23. Van de ven, A., Walker, G., and Liston, J. 1979. Coordination patterns within an interorganizational network. *Human Relations*, 32: 19-36.
24. Kiesler, S., and Sprauull, L. 1992. Group decision making and communication technology. *Organizational Behaviour and Human Deceision Processes*, 52: 96-123.
25. McCallum, A., Corrada-Emmanuel, A., and Wang, X. 2004. The Author-Recipient Topic model for topic and role discovery in social networks: Experiemnts with Enron and Academic Email. *Technical Report, Univeristy of Massachusetts*, Amherst, UM-CS-2004-096.
26. McCallum, A., Corrada-Emmanuel, A., and Wang, X. 2005. Topic and role discovery in socialnetworks. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*: 786-791.
27. McCallum, A., Wang, X., and Corrada-Emmanuel, A. 2007. Topic and role discovery in social networks with experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research*, Vol. 30: 249-272.
28. McGrath, J. E. 1983. *Groups: Interaction and Performance*. Englewood Cliffs. NJ Prentice-Hall.
29. McGrath, J. E. 1990. Time matters in groups. In J. Galegher, R. E. Kraut, and C. Egidio (Eds.), *Intellectual Teamwork: Social and Technological Foundation of Cooperative Work*, Hillsdale, NJ: Laurence Erlbaum Associate: 23-61.
30. McGrath, J. E. 1991. Time, interaction and performance (TIP): A theory of groups. *Small Groups Research*, 22: 147-174.
31. Garton, L., and Wellman, B. 1995. Social impacts of electronic mail in organizations: A review of the research literature. *Communication Yearbook*, 18: 434-453.

32. Haythornthwaite, C. 1996. Which kinds of network members communicate by email or face-to-face for what kinds of work?. University of Toronto: *Centre for Urban and Community Studies*, 45.
33. Wellman, B., and Gulia, M. 1997. Net surfers don't ride alone: Virtual community as community. In P. Kollack & M. Smith (Eds.), *Communities in Cyberspace*, Berkeley: University of California Press.
34. Matsuo, Y., Mori, J., and Hamasaki, M. 2006. POLYPHONET: An advanced social network extraction system from the web. *WWW 2006*, Edinburgh, Scotland: ACM.
35. Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K. and Ishizuka, M. 2007. POLYPHONET: An advanced social network extraction system from the web. *Journal of Web Semantics*, Web Semantics: Science, Services and Agents on the World Wide Web, 5: 262-278.
36. Nasution, Mahyuddin K. M. and Noah, S. A. 2010. Superficial method for extracting social network for academic using web snippets. In Yu, J. et al., (eds.), *Rough Set and Knowledge Technology*, LNAI 6401, Heidelberg: Springer: 483-490.
37. Carley, K. 1991. A theory of group stability. *American Sociological Review*, 56: 331-354.
38. Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3: 993-1022.
39. McCallum, A. 1999. Multi-label text classification with a mixture model trained by EM. *AAAI Workshop on Text Learning*.
40. Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. 2004. The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Banff, Alberta, Canada.
41. Steyvers, M., Smyth, P., Rosen-Zvi, M. and Griffiths, T. 2004. Probabilistic author-topic models for information discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington.
42. Duda, O., and Hart, P. 1973. *Pattern Recognition and Scene Analysis*. New York: Wiley.