Two Steps for Self-Organized Social Network Pre-Construction

Wei-Lun Chang
Department of Business Administration
Tamkang University
New Taipei City, Taiwan
wlchang@mail.tku.edu.tw

***Abstract*—A social network is a relational connection between individuals, including any relationships related to the exchange of information such as those among friends or colleagues. The most significant problem associated with social networks is establishing the role of key individuals, who are charged with conveying important messages among everyone involved. However, the most appropriate representatives are difficult to identify. The aim of this study was to develop a method with which to facilitate the automatic pre-construction of a social network prior to any interaction and pre-identify representatives within the network. The goals of this research were: (1) construct a social network based on a self-organization maps and social network analysis, (2) verify differences between pre-constructed and actual social networks, and (3) identify key representatives and validate the efficiency of the social network.**

Keywords- SOM; SNA; Social Network Construction

I. INTRODUCTION

Since the late 1920's, Hawthrone study has been the basis of research into the motives behind human relationships and the resulting behavior, and research into group behavior gained considerable attention in the 1950's. Groups play a vital role in determining the attitudes and behavior of members within an organization. Schein [12] defined groups as a union of members who know each other and belong to the same cluster. Boone and Kurtz [2] indicated that groups comprise more than one member, and demonstrate interaction and a common purpose. The formation stage of group development is critical, due to the exchange of information or resources among members and subgroups emerge during the brainstorming stage of group development.

Investigating subgroups is important to help managers comprehend the behavior of the entire group. Existing approaches to the analysis of social networks also help firms to identify connections associated with subgroups. Social networks include specific individuals (e.g., group, organization or social entity) and a set of connections within the social structure (e.g., friendship) [6]. This research considers groups as an example of social networks, which have the potential to develop both in scope and coherence.

Weber [16] indicated that social interaction and the division of work are important factors within successful organizations. The research of Cross and Prusak [3] and Serrat [13] specified that efficiency improves if the top manager can identify key members within a group. The traditional approach to analyzing social networks includes interviews to collect data and focus groups to generate visual representations of social networks [15]. Recently, many researchers have incorporated a variety of data sources to generate social networks. AT&T examined communication records [1], IBM investigated web pages [10], Sahar and Jabeen [11] analyzed Bluetooth communication records, and Thelwall [14] evaluated social network website data. Nevertheless, these studies examined only the influence or relationships of groups or networks. There has been a lack of research investigating the effect prior to the formation of the group.

Krackhardt and Hanson [9] identified the significance of informal groups. This research considers subgroups to be informal groups and the relationships within the subgroups have an impact on the organization. The problems dealt with in the research are (1) determining how to automatically identify subgroups, (2) how to pre-construct a social network based on subgroups prior to the actual formation of the group, and (3) how to identify the key individuals in the pre-constructed network. To deal with these problems, we propose a two-step approach to identify clusters and build social networks. The first step it to use self-organization maps to generate homogeneous clusters (subgroups). The second step is to use social network analysis to construct social networks from each cluster and identify key individuals. This research aims to provide an automatic method to pre-construct potential social networks and identify major members in advance.

The remainder of this research is organized as follows. Section 2 presents the related literature, including data clustering methods, social network theory and social network analysis. Section 3 outlines our research framework. In Section 4, we analyze the collected data and discuss the implications. Finally, in Section 6 we provide our conclusions and the limitation of this research.

II. RESEARCH METHOD

This research combines two methods: self-organization maps and social network analysis. The first step is the transformation of raw data into numerical values as a multidimensional vector. Next, all vectors are input to SOM to generate clusters, which are considered sub-groups in the social

978-1-4577-0653-0/11/$26.00 ©2011 IEEE

network. The third step is to use SNA to analyze each cluster (subgroup) and pre-construct particular social networks. We not only pre-constructs the social networks but connects each subgroup. The last step is an interview of the group to verify the performance of our method based on specific indicators.

A. *Self-Organization Maps*

Kohonen proposed the concept of SOM in 1989 [7], as a means to analyze large quantities of data. The major feature of SOM is the ability to map multidimensional and non-linear data into low dimensions (generally two dimensions), based on a competitive learning approach (Figure 1). The result can be visualized in graphs.

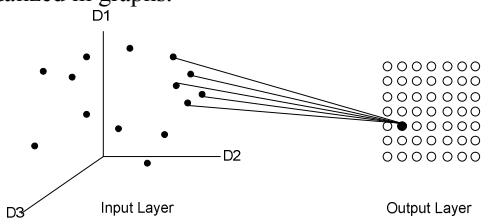


Figure 1. Concept of Self-Organization Maps

SOM has two major functions: training and mapping. The training process can be considered a competitive learning process similar to neurons in the human brain, capable of learning and processing various tasks. The neurons are connected and near neurons deal with similar tasks. The competitive learning process enables near neurons to move close to one another through an iterative learning loop, representing an unsupervised learning concept. The output of SOM is represented in two dimensional graphs. The process of SOM is shown as follows [8]. First, we assume an input vector x and a weight vector m_i in which $x = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \cdots, \varepsilon_n)$, $x \in \Re^n$ and $m_i \in \Re^n$. The first step is the estimation of the length of the vector in terms of normalization. ε'_k is the normalized vector and ε_k is the original vector of a data (**Eq.(1)**).

$$\varepsilon'_k = \frac{\varepsilon_k}{length} \quad \text{and} \quad length = \sqrt{\sum_{k=1}^{n} \varepsilon_k^2} \quad \ldots\ldots(1)$$

In the second step, we utilize the activation function to calculate the distance between input vector and weight vector. This research uses the concept of Euclidean distance as activation function (**Eq.(2)**).

$$\eta_i(t) = \|m_i(t) - x(t)\| = \sqrt{\sum_{k=1}^{n} (\mu_{ik}(t) - \varepsilon_k(t))^2} \quad \ldots\ldots(2)$$

The third step is to discover the winner among neurons following the competitive learning approach. The neuron with the shortest distance between data is considered the winner; in other words, the shortest distance between x and m_i. In SOM, the winner is also called the best matching unit (BMU). For instance, $c(t)$ is the BMU in **Eq.(3)**.

$$c(t): \eta_c(t) = \min_i(\eta_i(t)) = \min_i(\|m_i(t) - x(t)\|) \quad \ldots(3)$$

The fourth step is to adjust the distance between the BMU and near neurons. In **Eq.(4)**, $h_{ci}(t)$ is the neighborhood function, which is also a decreasing function. The purpose of the decreasing function is to ensure that the distance between BMU and near neuron is closer. That is, $h_{ci}(t)$ allows BMU c and near neuron i to adjust the distance.

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \quad \ldots\ldots(4)$$

In **Eq.(4)**, □(t) indicates the learning rate, thereby ensuring that the learning process is terminated within a limited time period. In addition, this research utilizes the most popular Gaussian function (**Eq.(5)**). The learning process enables near neurons to move closer and generates data clusters of high homogeneity (Figure 2).

$$h_{ci}(t) = \varepsilon(t) \cdot e^{-\frac{\|c-i\|}{2 \cdot v(t)^2}} \quad \ldots\ldots(5)$$

This research uses the concept of SOM as the first step for the pre-construction of social networks. We assumed that highly homogeneous data can be clustered efficiently. In the meantime, unsupervised learning allows an unsupervised number of clusters. If we assume that the entire data set represents a social network, the generated clusters are considered subgroups in the social network.

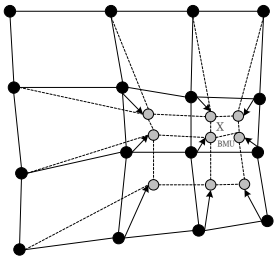


Figure 2. Concept of Best Matching Unit in learning process

B. *Social Network Analysis*

Social network analysis (SNA) is a method of examining the structure of a social relationship for a group and investigating the informal connections and relationships among individuals [5]. The basic assumption of SNA is that each individual is interdependent. This research utilizes the concept of SNA, as derived from the research of Tichy in 1979. We first require the size and density for SNA. Size

indicates the scale of a social network, as presented by the number of nodes in the social network. Density indicates the degree of closeness among members in the social network. The concept of density is shown in **Eq.(6)**; specifically, R is the number of relationships in the network and n is the number of nodes.

$$Density = \frac{R}{n(n-1)} \quad \text{(6)}$$

We also used three indicators from the research of Freeman [4], which are degree centrality, betweenness centrality and closeness centrality. The concept of centrality is considered an indicator to verify the efficiency of solving problems or delivering information to the group. Degree centrality indicates the number of adjacency individuals of a specific individual, used to interpret the degree of control in the movement of information or resources. The higher the degree centrality is, the closer an individual is to the center of the social network. In this research, we took into account indirectional relationships among nodes. We assumed a specific node P_k and estimated the number of adjacency nodes for P_k and n is the number of nodes; that is, $\sum_{i=1}^{n} a(P_i, P_k)$. $c_D(P_k)$ is the estimated degree centrality of P_k (**Eq.(7)**).

$$c_D(P_k) = \frac{\sum_{i=1}^{n} a(P_i, P_k)}{n-1} \quad \text{(7)}$$

Betweenness centrality is the importance of an individual between any other two nodes. The higher the degree of betweenness centrality is, the greater the ability of the in-between node to transmit information or resource. In **Eq.(8)**, $c_B(P_k)$ indicates the concept of betweenness centrality of P_k. b_{ij} is the shortest path between i and j through P_k. After standardization based on the concept of indirect relationships, $c_B'(P_k)$ is the estimated betweenness centrality in **Eq.(9)**.

$$c_B(P_k) = \sum_{i}^{n} \sum_{j}^{n} b_{ij}(P_k) \quad \text{(8)}$$

$$c_B'(P_k) = \frac{c_B(P_k)}{(n-1)(n-2)/2} \quad \text{(9)}$$

Closeness centrality indicates how close a node is to other nodes, representing how fast a node can connect to other nodes in the social network. In **Eq.(10)**, $c_a(P_k)$ is the estimated closeness centrality and $\sum_{i=1}^{n} d(P_i, P_k)$ is the sum of all shortest distances between P_k and other nodes.

$$c_a(P_k) = \frac{n-1}{\sum_{i=1}^{n} d(P_i, P_k)} \quad \text{(10)}$$

This research uses the concept of SNA as the second step to analyze all clusters based on three indicators. This study attempts to identify key individuals in each pre-constructed social network. Once the key actor is identified, the organization can deal with potential problems in advance.

III. DATA ANALYSIS

A. Data Description

This research sampled four classes from the first to the fourth year in the department of business administration of Tamkang University, Taiwan. The class representative was considered the leader of the class, similar to what would encounter in a small organization. In addition, the informal groups were formed automatically without manipulation. The process of allowing members to interact and generate more sub-groups is an unsupervised learning process. As shown in Figure 3, only the number of males (31) for first year was greater than females (28). In the second, third, and fourth year classes, females outnumbered males; however, the ratio was approximately 50 %. As shown in Figure 4, the number of participants with O-type blood was the highest for all four classes, at nearly 50 %. The number of participants with B type and A type were nearly equally for all four classes, approaching 50 %. AB type was the least common blood type for four classes.

B. Performance Indicator

In this section, we evaluate the accuracy of pre-constructed social networks for four classes (Figure 7). Traditional SNA research uses interviews and surveys to construct social networks; however, all people need to be interviewed, and this can be extremely time consuming. We interviewed only the specific individuals identified by the proposed method proactively. This study also invited representatives of the four classes to interviews. We provided students with analysis results to validate the situation in the real world. Because the scale of each student was different, we used the mid-point to determine whether the identified individuals with high or low degrees of centrality matched. Participants needed only to recall whether these individuals actually had high or low values based on a simple judgment of dichotomy. We believe that this method easily reaches a consensus in the interview process. Thus, the accuracy of our method can be measured as: the identified high/low value of a specific centrality for an actor / real situation by the interviewed participants. For example, if A has high value of degree centrality and participants believe this to be so, then the accuracy for predicting degree centrality will be 100 %.

For the first year class, we invited six students, including 2 class representatives, for interviews. We also showed the identified individuals from the pre-constructed social network and requested their feedback. In our evaluation of degree centrality, 5 individuals were believed to match seven identified individuals from pre-constructed social networks.

The accuracy of degree centrality was 71 % (i.e., 5/7). In the evaluation of betweenness centrality, 8 individuals were believed to match more than eight identified individuals from pre-constructed social networks. The accuracy of degree centrality was 100 % (i.e., 8/8). In the evaluation of closeness centrality, four individuals were believed to match more than seven identified indivduals from pre-constructed social networks. The accuracy of degree centrality was 57 % (i.e., 4/7). The average accuracy of the pre-constructed social network was 76 % (i.e., (71 % + 100 % + 57 %)/3).

For the second year class, we invited 6 students for interviews. In the evaluation of degree centrality, seven individuals were believed to match more than ten identified individuals from pre-constructed social network. The accuracy of degree centrality was 70 % (i.e., 7/10). In the evaluation of betweenness centrality, eight individuals were believed to match more than nine identified individuals from pre-constructed social networks. The accuracy of degree centrality was 89 % (i.e., 8/9). In the evaluation of closeness centrality, eight individuals were believed to match more than eleven identified individuals from pre-constructed social networks. The accuracy of degree centrality was 73 % (i.e., 8/11). The average accuracy of the pre-constructed social networks was 77 %.

For the third year class, we invited seven students for interviews. In the evaluation of degree centrality, eight individuals were believed to match over 10 identified individuals from pre-constructed social networks. The accuracy of degree centrality was 80 % (i.e., 8/10). In the evaluation of betweenness centrality, seven individuals were believed to match over 9 identified individuals from pre-constructed social network. The accuracy of degree centrality was 78 % (i.e., 7/9). In the evaluation of closeness centrality, eight individuals were believed to match more than ten identified individuals from pre-constructed social networks. The accuracy of degree centrality was 80 % (i.e., 8/10). The average accuracy of the pre-constructed social networks was 79 %.

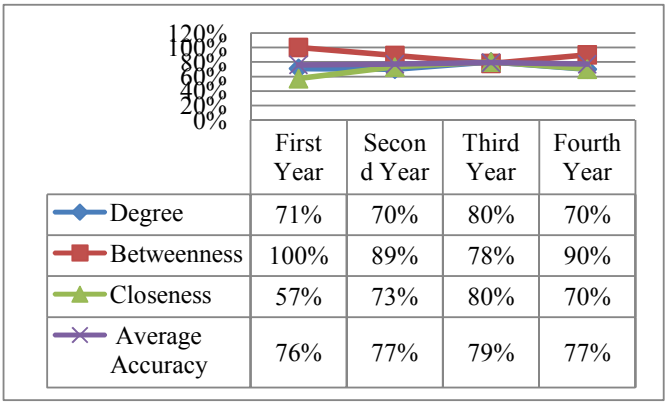
For the fourth year class, we invited five students for interviews. In the evaluation of degree centrality, seven individuals were believed to match more than ten identified individuals from pre-constructed social networks. The accuracy of degree centrality was 70 % (i.e., 7/10). In the evaluation of betweenness centrality, nine individuals were believed to match more than ten identified individuals from pre-constructed social networks. The accuracy of degree centrality was 90 % (i.e., 9/10). In the evaluation of closeness centrality, seven individuals were believed to match more than ten identified individuals from pre-constructed social networks. The accuracy of degree centrality was 70 % (i.e., 7/10). The average accuracy of the pre-constructed social networks was 77 %.

C. *Discussion*

The result reveals the accuracy of betweenness centrality is the highest. Betweenness centrality was used to measure the influence of leadership on actors in the network. The first year class had the highest degree of accuracy for betweenness centrality. According to the theory of Tuckman (1965), the first year class could be considered the first step to form the group. Because individuals avoid conflicts in this stage, leadership is easy to build for actors, and the degree of betweenness centrality is therefore high. The lowest degree of accuracy for betweenness centrality was the third year class because the third year is in a brain-storming stage forming groups and many actors have conflicts in attempting to become leaders. Thus, the accuracy of betweenness centrality is low in this stage.

The second highest degree of accuracy is degree centrality, at 73 %. Specifically, the third year class had the highest accuracy for degree centrality (80 %). Degree centrality focuses on the connections of actors. Because we interviewed only select actors, the actual connections of each actor were difficult to observe among few individuals. Thus, we infer that the accuracy of degree centrality is low. In addition, the difference of accuracy among the four classes was insignificant.

The lowest accuracy was demonstrated for betweenness centrality and the variance among four classes was significant. Betweenness centrality measures the degree of tightness of actors. All participants considered betweenness centrality difficult to measure in the social network. We inferred that first year students had many required courses that may result in superior understanding of each student. Conversely, students of fourth year had fewer required courses and more electives. This could result in a decrease in tightness among actors in the social network. Time is another factor reducing the degree of tightness among actors; therefore, the accuracy in the fourth year is less than the first year.



	First Year	Second Year	Third Year	Fourth Year
Degree	71%	70%	80%	70%
Betweenness	100%	89%	78%	90%
Closeness	57%	73%	80%	70%
Average Accuracy	76%	77%	79%	77%

Figure 3. Comparison of accuracy of centralities

The average accuracy of three centralities was between 76 % and 79 %, indicating that the influence of time does not influence the accuracy of our approach. Time has an insignificant influence on the degree centrality among the four classes with accuracy of between 70 % and 80 % (i.e., 10 % of variance). Time significantly influences closeness centrality among four classes with accuracy between 78 % and 100 % (i.e., 22 % of variance) and the closeness centrality among four classes with accuracy between 57 % and 80 % (i.e., 23 % of variance). In summary, we infer that the result of accuracy is influenced by the fundamental measurement of three

centralities. Betweenness centrality is easy to measure from a third party perspective. Our results also confirm that the accuracy is the highest. Degree centrality poses the number of connection for a particular actor which is mainly subjective and not easy to measure from third party perspective. Finally, closeness centrality is the most difficult measure among the three centralities. Our results also reveal that the variance of accuracy is significant for closeness centrality.

IV. CONCLUSION

This research proposes a novel approach to automatically pre-construct social network within an unknown group. The major difference between our method and the traditional method is that the known relationships and connections among actors are unnecessary. This work relies on personal data to generate clusters (step 1) and pre-construct the social network (step 2). This research uses SOM to break down data into clusters, in accordance with the concept of Cross and Prusak (2002). SOM transforms data into vectors and uses concept of distance between data to represent connections. In the second step, we use traditional social network analysis to construct the social network based on the clusters. We also use degree centrality, betweenness centrality, and closeness centrality as indicators to measure the accuracy of the proposed approach.

We recruited students from four classes at Tamkang University in Taiwan as our sample. Our results reveal an average accuracy of 77 %. The accuracy of betweenness centrality was the highest, averaged to 89 %. The reason is that measuring betweenness centrality is easier than for the other two centralities. Moreover, the accuracy of the first year class was 100 % (highest) and fourth year of class was 78 % (lowest). These results also confirm to the findings of Tuckman (1965). The average accuracy of degree centrality was 73 %, and the influence of time was low for degree centrality. Finally the average accuracy of closeness centrality was the lowest, at 70 %. The influence of time was significant for closeness centrality. External factors also influenced closeness centrality. Hence, the observation from third party may influence the accuracy of our approach.

There are several limitations to this research. First, personal information may generate bias for pre-constructing social networks. Personal data cannot reflect interactions, relationships, and connections in a social network. The generated social networks may not explain the phenomenon completely. Traditional social network analysis requires interviews with all actors in the social network; however, this research uses a focus group to verify the pre-constructed social network, thereby saving time but lacking the completeness provided by interviews. Third, changes within organizations occur all the time (Serrat, 2009). This research did not consider the adjustment of the pre-constructed social networks. In summary, the proposed method not only provides a different perspective from which to analyze social networks, but also helps managers to preview the network and identify key individuals in advance. The results of this research could furnish a roadmap for related research in the future.

REFERENCES

[1] Abello, J., Pardalos, P. M. and Resende, M. G. C. (1998), *On Maximum Clique Problems in Very Large Graphs*, Research technical report: TR 98. 32. 1, AT&T Labs.

[2] Boone, L. E. and Kurtz, D. L. (1987), *Original: Management, 4th ed*, New York : McGraw-Hill.

[3] Cross, R. and Prusak, L. (2002, June), 'The People Who Make Organizations Go-or Stop', *Harvard Business Review,* 80(6), 104 – 112.

[4] Freeman L.C. (1979), 'Centrality in Social Networks : Conceptual Clarification', *Social Networks*, 1(3), 215-239.

[5] Hanneman, R., and Riddle, M. (2005), *Introduction to Social Network Methods*, Riverside, CA: University of California, Riverside.

[6] Jamali, M., and Abolhassani, H. (2006), 'Different Aspects of Social Network Analysis', In WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Washington, DC, USA.

[7] Kohonen, T. (1990), 'Self-Organizing Maps', *Proceedings of the IEEE*, 78(9) 1464 – 1480.

[8] Kohonen, T. (1995), *Self-Organizing Maps*, Berlin: Springer.

[9] Krackhardt, D. and Hanson, J. (2000, July), 'Informal Networks the Company Behind the Charts', *Harvard Business Review*, 78(4), 104 – 111.

[10] Kumar, R., Raghavan P., Rajagopalan S., and Tomkins A. (1999), 'Trawling the Web for Emerging Cyber-Communities', *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 31(11-16), 1481-1493.

[11] Sahar, G. and Jabee, G. (2009), 'Group Formation in Social Network: An Overview', *MASAUM Journal of Reviews and Survey*, 1(1), 105-109

[12] Schein, E. H. (1980), *Organization Psychology, Englewood Cliffs*, New Jersy : Prentice-Hall

[13] Serrat, O. (2009, February), 'Social Network Analysis', *Knowledge Solution, ADB*.

[14] Thelwall, M. (2008), 'Social Networks, Genders, and Friending: An Analysis of MySpace Member Profiles', *Journal of the American Society for Information Science and Technology*, 59(8), 1321-1330.

[15] van Duijn, M. A. J. and Vermunt, J. K. (2006), 'What is Special About Social Network Analysis', *Methodology*, 2(1), 2-6.

[16] Weber, M. (1947), The theory of Social and Economic Organization, New York: Free Press.