EXPLORING INTERACTION BETWEEN ONLINE COMMUNITY USERS BY ANALYZING COMMENTS

Research-in-Progress

Bae, Soon-Han
Management of Information and Systems Dept., Hanyang University
#17, Haengdang 1-dong, Seongdong-gu, Seoul, Korea
ifsleeping@naver.com

Kim, Byoung-Suk
Management of Information and Systems Dept., Hanyang University
#17, Haengdang 1-dong, Seongdong-gu, Seoul, Korea
datam@hanmail.net

Kim, Young-Min
Management of Information and Systems Dept., Hanyang University
#17, Haengdang 1-dong, Seongdong-gu, Seoul, Korea
Sbaek@hanyang.ac.kr

Baek, Seung-Ik
Management of Information and Systems Dept., Hanyang University
#17, Haengdang 1-dong, Seongdong-gu, Seoul, Korea
Sbaek@hanyang.ac.kr

Abstract

Today's online community has shown profound impacts as a social network is expanding its sphere of influence across the whole society - not limited to politics, economy and culture. Online community goes beyond personal needs. It even includes political and economic issues. The traditional mass media such as newspapers and broadcasting stations did take the same role. Now online communities are a more active environment where users voluntarily communicate with each other, thus creating and sharing information and knowledge. It is the constant interplay among users that creates online communities and even makes them grow or disappear. The purpose of this study is to investigate how users interact with each other in online communities by analyzing the commenting platform in internet bulletin boards. Social Networking Analysis (SNA) was used for an analysis method. The whole study was conducted in two phases. First, we chose actual online communities to see the interactive patterns between users. In the second phase, simulation tests were conducted in many different conditions to see any changes in their patterns of interaction. The results are as following: in its early period, the communities are visited by a small number of people. Information is mainly created and shared by a few influential users; however, once visitors grow in number, more influential users appear. Online communities were run on a small scale by many important members, not by a handful of crucial ones. In the second phase of our study, Monte-Carlo's simulations were used to find out how commenters exhibit their interactive variations under diverse conditions; yet, most of their patterns were similar with those in the first phase. As a conclusion, this study interpreted and measured the interactive patterns among online community members from the perspective of SNA. At a first glimpse, online communities may seem chaotic and unsystematic due to their distinct technological nature and the anonymity of commenters. But numerical measures allowed us to see that certain structural patterns exist in the online communities and those patterns keep changing.

Keywords: *Social Network Analysis (SNA), Online Community, Mote-Carlro Simulation*

1. Introduction

Online community is no longer limited to a space where we can freely share information and knowledge. Its presence as a social network expands across political, economic, and cultural landscapes, impacting the society on the whole. In the past, communities on the web were merely an unofficial, passive media where people had small chats about their petty issues in life or gossips. But today's internet communities have evolved. Not just meeting the personal needs of communication, these web sites play a huge role in forming public opinion about various issues in politics, economy and etc. - which used to be promoted by conventional medias like the press and broadcast centers. The social impacts of online communities are obviously exemplified by the discussions over "mad cow disease" and "melamin-tainted milk powders". Online communities are not simply functioning as an open forum for exchanging opinions; web communities, have been elevated as a social media, triggering collective actions in offline reality. Sometimes, these powerful communities even keep in control the government policies and business practices. Thus, governments and businesses feel the necessity to keep a careful watch on the activities of online communities because these groups can turn into 'invisible friends or foes'. Previous studies had a narrow focus, only explaining what affects the performance of online communities. Dynamical changes within web communities have not been analyzed in an organized manner. Therefore, this study will shed light on the relationship and interactive patterns between participants in online communities and will investigate more scientifically how the structure of the communities alters as the interaction among participants play out. In this study, social network analysis (SNA) is adopted as a scientific method in order to explore the structural changes of interactions in online communities: we utilized the SNA method to analyze comments - which web community participants mostly rely on to communicate with each other; then, the interactive patterns among participants could be analyzed; and finally, via the analysis of these patterns, the interactive structure of online communities was identified. Also, this study inquired how the structure of online communities evolves when participants expand in number and more frequently interact with each other. Lastly, we attempted to see whether online communities go through any changes in structure depending on the type of topics discussed by members.

2. Theoretical Background

2.1 Social Network Analysis (SNA)

In this study, for the better understanding of the interactions between participants, SNS, social network analysis, was used in analyzing their comments. In 1930s, SNA came into use in the field of social and behavioral science, not noted by many researchers until the late 20th century. But around this time, many concepts such as socio-metric, graph theory, dyad, triad, subgroup and block-model were theorized. Current SNA is widely applied to various research fields like sociology, business administration, economics and etc. (No, 2009; Seo et al., 2010; Kelly, et al., 2005; Mutuswami & Winter, 2002). SNA is focused on the relationship formation of study participants through interaction. There were attempts to provide mathematical explanations about the structure of the relationship forming with the help of SNA. Generally, graph theory is employed in the process of SNA, producing a matrix about the links (interaction) between nodes (participants). The structure of social network can be explained by diverse graphical indicators of nodes and links (Wassernam & Faust, 1994).

2.2 Indicators for structure of social network

This paper studies how the number and frequent interplay, referred to as density, of participants affects the centralization and inclusiveness of online communities. The numerical characteristics and meaning of density, inclusiveness and centralization within the communities on the web is as follow.

2.2.1 Density

Density is a measure that shows the extent of interactions among participants within a network, describing how many relationships are formed among players in a network and how thoroughly a network is structured (Son, 2002).

Ideal online communities should be a platform where all users are connected enough to readily share every information and knowledge. Therefore, density can be translated as the ratio between total open links within a network and actually-used links: (Currently-used links divided by Total open links within a network). When the direction is excluded, the density of a network can be measured by the next formula.

$$Density = \frac{L}{\frac{n(n-1)}{2}} = \frac{2L}{n(n-1)}$$

- L : The links between nodes within a network
- n : The number of nodes within a network
- $\frac{n(n-1)}{2}$: Total open links within a network

2.2.2 Inclusiveness

Inclusiveness means the number of nodes that are connected with at least more than one different node within a network, focusing on the connection between nodes (*total nodes minus isolated nodes*). This indicator explains how broadly a network covers; thus, the more inclusive a network, the more diverse the interconnection among nodes. But the concept of absolute inclusiveness is influenced by the network scale which is the whole number of participants. Therefore, inclusiveness is calculated by a relative ratio in order for an easy comparison among networks. The specific formula is as follow:

$$Inclusiveness = \frac{N_c}{N}$$

- N_c : the number of connected nodes
- N : the totality of nodes in a network

2.2.3 Degree Centrality

The influence of a certain person in a social network, also known as a node, can be explained by centralization - a degree of how centralized a whole network is toward its center; in other words, it is the impact of a node on the entire network. For example, if a small number of people dominate the interaction, the network is showing a high centralization. On the other hand, the interaction among diverse participants renders the network lowly-centralized. As mentioned above, the density is defined as an indicator about the connection between participants. High density equals the presence of various interactions; higher density leads to less concentration on a handful of participants, (low centralization), while lower density means higher concentration on them, (high centralization). This proves that density and centralization have a reverse relation. Also, while density is a mean or average calculated from the numbers about the relationship between participants, centralization is a concept of variation. Therefore, centralization clarifies how evenly communication and interaction occur or how they are concentrated on a minority of people; hence, it is possible to estimate the influence and ripple effect of a few participants in an online community. Centralization is classified degree centralization, closeness centralization, and betweenness centralization (Freeman, 1979; Son, 2002). First, degree centralization is used to measure the degree of connection between nodes. Second, closeness centralization is based on the distance between nodes. Third, betweenness centralization focuses on how many other nodes are mediated by one node. In fact, degree centralization is the fittest indicator for the purpose of this study, which is to grasp the structure of online communities, since it explains the degree of influence within a network based on the changes in the number of links. The mathematical explanation of degree centralization is as follow:

$$C_D = \frac{\sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]}{Max\sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]}$$

$$= \frac{\sum_{i=1}^{g}[C_D(n^*) - C_D(n_i)]}{[(n-1)(n-2)]}$$

- C_D : degree centralization
- $C_D(n^*)$: the biggest value of centralization in a network
- $C_D(n_i)$: the centralization of each node in a network
- n : the number of nodes in a network

3. Empirical Study

3.1 Research design and method

The communication structure within online communities keeps evolving, not being fixated, which driven by the constant interaction among participants. Online communication in the communities primarily takes the form of commenting on an internet bulletin board. An analysis of the commenting process offers a better grasp about the interactive characteristics of the community users. Therefore, this study looks closely into the commenting patterns among web community participants, exploring how the centrality and data-sharing range of online communities alter depending on the increasing number and interactive frequency of participants.

3.1.1 Research Procedures

The research object was 'Agora' page of Daum (www.daum.net), the most popular online debate forum in South Korea; this study selected two specific social issues - rat's head in a Korean snack and mad cow's disease - and inspected every online community that opened up debates on them. Web forums in Agora are organized in the following structure (Picture 1): a participant raises an issue by posting on the bulletin board, triggering others to interact by writing on the topic or replying to other writings. This research is based on a premise that an online community is created when a certain debate topic appears in the board and that the articles and following postings are defined as relies. The interactive patterns among debaters were analyzed through the SNA method by scrutinizing the structure of replies.

3.1.2 Data collection methods and procedures

The Agora consists of numerous writings and subsequent replies - relatively described as 'issues' and as 'debate articles' and 'replies'. The target samples for analysis are the issue articles searched by the two key words - 'rat's head in snacks' and 'mad cow disease'; 74 issue articles were collected for the former topic and 72 for the latter; also, the number of hits and replies were counted (Table 1). Most of all, an adjacency matrix should be made to analyze the interactive patterns among online community participants.

Table 1 The characteristics of samples in online communities						
	Total		Rat's head in snack		mad cow disease	
	Average	S.D.	Average	S.D	Average	Standard deviation	Note
Hits	6781.6	5943.3	8138.2	4358.4	5387.3	6979.8	The number of networks (Issues) Rat's head in snack: 74 Mad-Cow Disease: 72
Replies	109.3	77.9	132.6	81.4	85.3	66.5	
Participants	84.0	59.7	108.6	59.0	58.7	49.0	
Centralization	0.11	0.14	0.06	0.04	0.15	0.17	
Inclusiveness	0.33	0.22	0.29	0.13	0.38	0.27	
Density	0.018	0.415	0.003	0.003	0.032	0.055	

A network is defined as the structure of 'debate articles' and 'replies' in terms of a topic: An article followed by another reply is assumed as an interaction between two participants, marked as '1' in the adjacency matrix; the absence of replies is considered non-communication, thus marked as '0' (Figure 1). For instance, the intersection between the row of the nickname '두찌맘(a)' and the column of '누룽지(b)' is valued as '1', indicating the presence of interaction where 두찌맘(a) posted a debate article and 누룽지(b) replied to it. With the help of the adjacency matrix, the two indicators of interactive patterns in a network - degree centralization and inclusiveness - were measured by

using UCINet. Also, SPSS 17.0 was employed for comparing the correlation between the social network indicators and others. Monte-Calro simulations were implemented through Excel program.

Picture 1. The analyzed Internet bulletin board and adjacency matrix
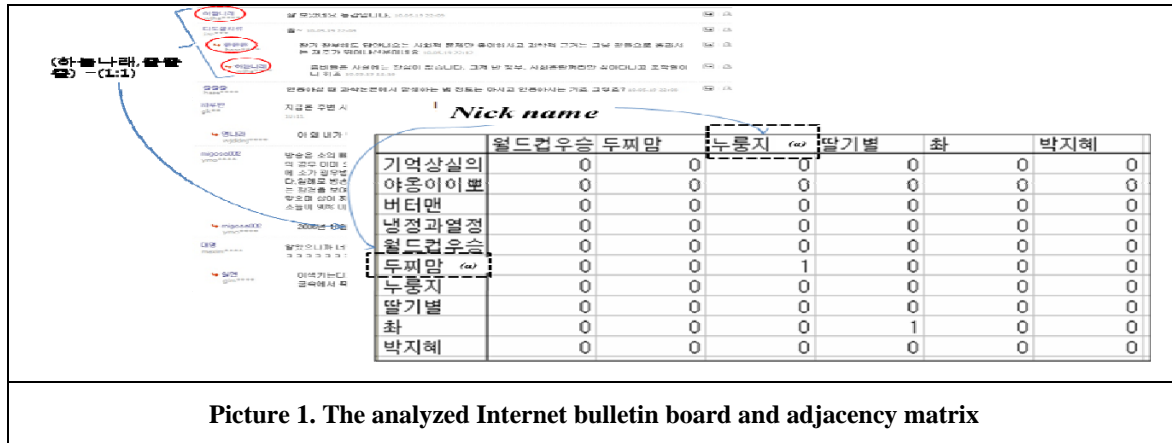
5. Analysis

This paper utilized the adjacency matrix of online communities to explore the shifting patterns of degree centralization and inclusiveness due to the changing density and the different number of participants - density and degree centralization; the number of participants and degree centralization; density and inclusiveness; and the number of participants and inclusiveness. Especially, this study applied a curve fitting method in the analysis of the relations among variables, producing the optimal model from the perspective of R^2. The estimated model is shown in <Table 2>

	Model		X-axis	Y-axis	interpretation
Table 2. Estimated model					
	Model		X-axis	Y-axis	interpretation
(1)		Value	Density	Degree Centralization	· The speed of information diffusion (The facilitation of information search)
		Meaning	the interactive degree among participants	influence, the presence of a hub	
(2)		Value	the number of participants	Degree Centralization	
		Meaning	the participants within a network	influence, the presence of a hub	
(3)		Value	density	inclusiveness	· The range of information diffusion
		Meaning	the interactive degree among participants	The range of information exchange within a network	
(4)		Value	the number of participants	inclusiveness	
		Meaning	the participants within a network	The range of information exchange within a network	

5.1 Model (1) & (2): Degree centralization in patterns with regard to the changing density and participants
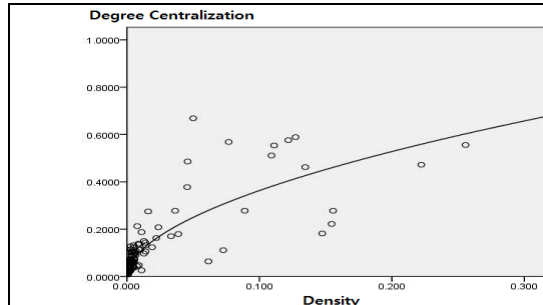
Density is a ratio between the entire links and the actual connection between participants. In other words, density indicates the proportion of participants who are actually engaged in intercommunication and information sharing. The <Table 3> elucidates the correlation between: (1) density and degree centralization; (2) the number of

participants and degree centralization. Both correlations are most highly-rated in the power model respectively with 76% and 57% in - visualized by the picture 2 and 3. In the picture 2, the issues with higher interactions (or density) among participants show a gradual increase in degree centralization; yet, the opposite is true in the picture 3. Thus, it can be inferred that amid intense interactions among participants, a few influential leaders emerge, strengthening the discussions about a topic and encouraging others to share and spread their views. In addition, a high density brings about the smooth exchange of information among online community participants; in such a state, new visitors will know the presence and location of leaders and attempt to communicate with them. However, if the density is low, new comers will be engaged in limited interactions with other participants due to lacking information about the leaders. Therefore, in terms of efficiency of information diffusion within a community, when participants are lower in number and density is higher, the spread of information in online communities will be faster. If the degree centralization is 0, there exists no crucial leaders, ending up with the inefficient dissemination of information and hence the obsolescence as a community. In order to liven up a community, participants should more frequently interact with each other - high density - so that other participants can recognize the presence and location of leaders. This is why online communities are swifter in diffusing and sharing information than offline equivalents. In offline communities, approaching and sharing information from leaders require many resources like time, materials, costs and efforts. However, the relying structure of online communities makes it far easier to spot leading participants and to communicate with them. Therefore, it is not just the internet speed that accelerates the spread of information within online communities; yet, the key factor is the users who readily identify leaders and thus reduce the cost of finding information.
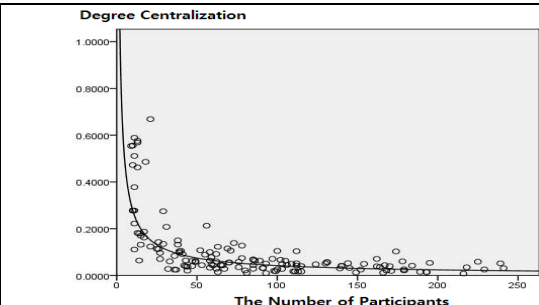
Table 3. The relevancy of model for density, the number of Participants and Degree centralization(R^2)					
	Estimated Model	Linear	Quadratic	Cubic	Power
(1)	Density - Degree centralization	59.6%(.000)	68.8%(.000)	74.9%(.000)	**76.8%(.000)****
(2)	The number of participant - Degree centralization	27.1%(.000)	46.1%(.000)	55.7%(.000)	**57.0%(.000)****
					** : $p<0.05$, ***: $p<0.01$

Picture 2. Density - Degree Centralization

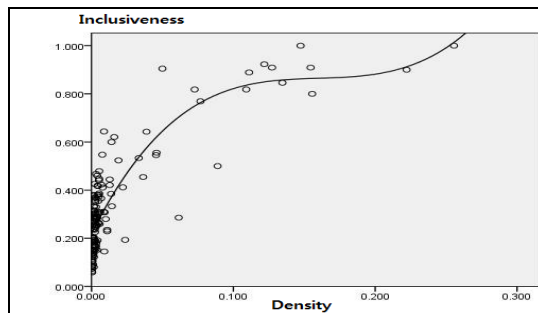**Picture 3. The number of participant-Degree centralization**

As density increases, however, degree centralization slows down. Likewise, according to the number of participants, degree centralization sharply drops from some point. Thus, when the number of participants and their interaction reach a certain level, the original leading node will see new ones with similar influence. In this point of view, there is a high possibility of information diffusion and its distortion.

5.2 Model (3) & (4): Degree centralization in patterns with regard to the changing density and participants
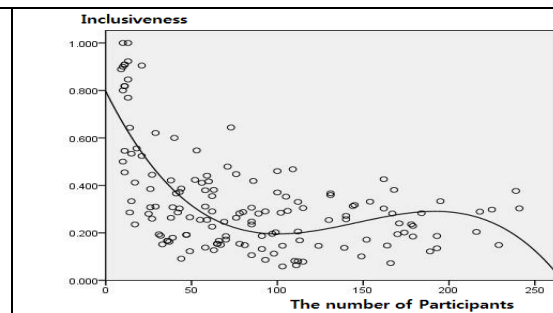
Inclusiveness is the number of participants who have interacted within the community with at least one person through replies. High inclusiveness means a huge number of people have communicated with each other in their community. As aforementioned, the range of information diffusion in a specific community can be estimated by inclusiveness. The <Table 4> presents the relation between: (3) density and inclusiveness; (4) the number of participants and inclusiveness. The inclusiveness patterns according to changing density are shown by a cubic model

(=73.7%); so was the degree centralization according to the number of participants (=48.6%). Such changes in patterns are reflected in the graphs of the picture 4 and 5. Like in the case of degree centralization, inclusiveness becomes higher when density increases, while it's reduced by the expansion of participants. Unlike degree centralization, however, inclusiveness is in the relation of cubic function with density and the number of participant.

Table 4. The relevancy of model for the number of Participants and Inclusiveness(R^2)					
	Estimated Model	Linear	Quadratic	Cubic	Power
(3)	Density - Inclusiveness	64.6%(.000)	72.2%(.000)	**73.7%(.000)**	68.5%(.000)
(4)	The number of participant-Inclusiveness	21.5%(.000)	40.8%(.000)	**48.6%(.000)**	32.3%(.000)
				** : $p<0.05$, *** : $p<0.01$	

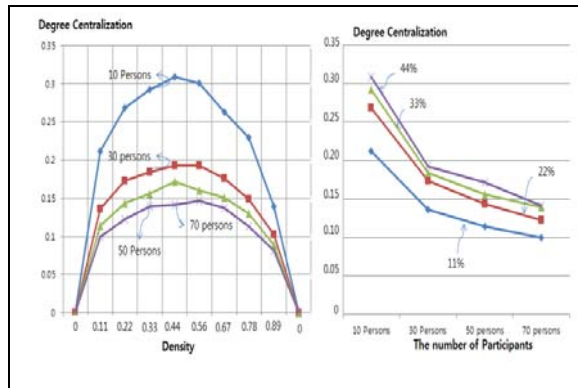
Picture 4. Density - Inclusiveness	Picture 5. Participant - Inclusiveness

According to above pictures, more connections among people, which is higher density, broaden and then slow down inclusiveness; in particular, once there are 100 participants, inclusiveness drops to the lowest point - meaning that the isolated nodes without any connections are maximized in number. When participants increase to 200, inclusiveness gets bigger, (the reduction of unconnected nodes), and shrinks after that, (the increase of unconnected nodes) - which tells that the connection with the number of participant is not linear but cubic and that online communities do not guarantee unconditional information diffusion and handy communication.
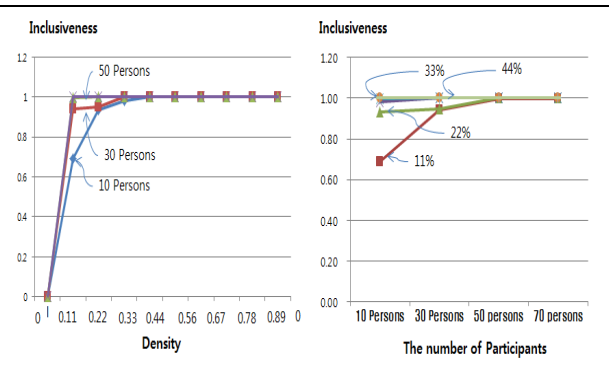
5.3 Monte-Calro simulation test

In order to find the variations undetected in the previous research procedures, this study conducted a Monte-Calro simulation test to further explore how the online community structure changes under diverse conditions. Monte-Calro technique is a simulation experiment often used in an uncertain environment via stochastical methods so as to produce both incidental and stochastical results. In this research, online communities were investigated through the random selection of participant number, density and subsequent connections. Relevant conditions are specified in the following <Table 5>:

Table 5. Linkage in relation to participant number and density								
Density / Participant Number	0.11%	0.22%	0.33 %	0.44%	0.56%	0.67 %	0.78 %	0.89 %
10	5	10	15	20	25	30	35	40
30	48	97	145	193	242	290	338	387
50	136	272	408	544	681	817	953	1089
70	268	537	805	1073	1342	1610	1878	2147

Picture 6. Relation between density and inclusiveness according to participant number	Picture 7. Relation between density and inclusiveness according to participant number

5.3.1 Simulation test on degree centralization in relation to density and participant number

Above all, the simulation results of density, participant number and degree centralization presents that under the same number of participants, degree centralization is on the rise when density is around 0-50%, reaches its peak when the density is between 40-50%, and declines after the density section of 50-100%. Also, when density is controlled, the overall linkage shrinks as the number of participant increase from 10 to 70.(See Picture 6). These pictures demonstrate the limited power of the most influential node despite the linkage of numerous online community participants. Our simulation tests proved that communities with density over 50% have an interactive structure where individual nodes are involved in a variety of separate information exchanges, not the system dominated by a handful of highly-influential members.

5.3.2 Simulation test on density, participant number and inclusiveness

According to the results of simulation test on density shown in above <picture 7>, participant number and inclusiveness, inclusiveness rapidly approaches the value of 1 when the participant number is identical and density progresses from 0%, 11%, 22%, 33% to 44%, thus concluding that inclusiveness is huge in relation to the same density and increasing participants. In this simulation experiment, every participant was linked and isolated dots no longer exist after the density of 20%. This simulation test concludes that inclusiveness drastically rises according to the increase in density and participant number and that all people are placed in the state of perfect linkage when density and participant number reach a certain lever. Our empirical analysis and the simulation test results are epitomized in the next <Table 6>:

Table 6. The summary of results				
X	Y	Empirical analysis		Simulation test
D	DCZ	• Higher possibility to connect with influential people • Appearance of influential people at an early period of interaction • Degree centralization rapidly expands at the initial period of interaction; the spread of influence slows down as interaction increases.		• Higher chances of communication with influential participant in case of stronger interaction among people • Influence reaching its peak in case that interaction rises to a certain degree • Reduction of density at around 0.5
	IC	• Communication with diverse people encouraged by increasing interaction • Active communication centered around a specific participant in case that interaction rises to a certain level		• Sharp reduction in the ratio of isolated people due to stronger interaction
NP	DCZ	• New participants communicate with people		• Growing interaction of new participants

		around the hub.	with the people on their periphery
	IC	• Difficulty to find people to communicate with in case of the increase in participant number • Occurrence of conflicts among participants with disparate interest • Emergence of participants with different interests	• Tendency to interact with surrounding people owing to actual difficulty to find the most influential person in case of the surge in participants
D : Density / DCZ : Degree Centralization/ IC : inclusiveness / NP : the number of people			

6. Conclusion

6.1 Findings and implications

  The structure of online communities is far from standardized or fixated space, but a product of constant communication among people. Nonetheless, consistent patterns of structural changes exist within seemingly chaotic online communities. For the detection of such patterns, replies on actual web communities were analyzed and Monte-Carlo simulations were conducted. As a result, following patterns were found. Firstly, in an initial stage of online community development, the information of a few influential participants is the center of attention, attracting a variety of people to their periphery and igniting an active flow of knowledge and experience - a far cry from the conclusion suggested with a concept called 'scale free network' by Barabasi (2001) that new nodes prefer other nodes with high connections beforehand. In fact, the linkage patterns of replying structure in online bulletin boards displays that new information channels are created while communicating with surrounding people rather than specific people with power and influence. These conclusions were proven by the changes in degree centralization and inclusiveness in relation to density and participant number. Secondly, this study estimated in a numerical manner the interaction among online community participants via the analysis of replying structure and clarified the structural pattern of digital communities, thus defining how the information sharing and diffusion takes place within online communities and predicting its preventability. This research will provide rich implications for diverse activities of management and administration closely related with online communities.

5.2. Acknowledgements and limitations of study

  This study contains following limitations: the structure of online network is confined to internet bulletin boards of web communities; merely two issues were adopted. Therefore, we should avoid hasty generalizations about the network structure of online communities on the basis of the patterns explored in this paper. However, as the results of Monte-Carlo simulation tests mostly corresponded with the findings of this research, it is safe to say that we have accurately depicted a partial picture of the changes in structural patterns of online communities. Additionally, for the inspection of social networks, posted articles and following replies were collected and organized in an adjacent matrix; yet, during the process, writings and replies may include informal languages and swear words which barely reflect individual views. Of course, unnecessary parts were removed during the data collection; there is a lack of scientific standards for the individual behavior of replying. This is because the vast volume of data was an extremely daunting task as the data collection for the social network analysis was processed at first hand. In the later works, date collection and analysis method should be enhanced so as to measure more precisely individual replying behaviors and to better comprehend the structure of online network.

Reference

Seo, J.G., Bae, S.H., Baek, S. I.(2010), Exploring Centralities of an Online Community", Journal of Knowledge Management", Vol.11, No.2. pp17-35"

Freeman, L.C.(1979), "Centrality in social networks-Conceptual clarification", Social Networks, Vol.1, pp.215-239

Kelly, J., D. Fisher., and M. Smith(2005), "Debate, Division and diversity : Political Discourse Networks in USENET Newsgroups Pater prepared for the "Online Deliberation Conference 2005", Standford University.

Mutuswami, S. and Winter, E. (2002), "Subscription Mechanisms for Network Formation", Journal of Economic Theory, 106, 242-264.

No, K.Y.(2008), "A Study on and Communication by New media and Its interactivity", KISDI (Korea Information society development institute)Son, D. W.(2002), social network, Kyoung Moon Press, Seoul

Wasserman, S. and Faust, K.(1994), Social Network Analysis Methods and Applications, Cambridge University Press, NY.