

Information Navigation and Knowledge Discovery in Virtual Communities

Wenyuan Li

Nanyang Technological University, Singapore

Wee-Keong Ng

Nanyang Technological University, Singapore

Kok-Leong Ong

Deakin University, Australia

INTRODUCTION

Over the past decade, advances in the Internet and media technology have literally brought people closer than ever before. It is interesting to note that traditional sociological definitions of a community have been outmoded, for community has extended far beyond the geographical boundaries that were held by traditional definitions (Wellman & Gulia, 1999). Virtual or online community was defined in such a context to describe various forms of computer-mediated communication (CMC). Although virtual communities do not necessarily arise from the Internet, the overwhelming popularity of the Internet is one of the main reasons that virtual communities receive so much attention (Rheingold, 1999). The beginning of virtual communities is attributed to scientists who exchanged information and cooperatively conduct research during the 1970s. There are four needs of participants in a virtual community: member interest, social interaction, imagination, and transaction (Hagel & Armstrong, 1997). The first two focus more on the information exchange and knowledge discovery; the imagination is for entertainment; and the transaction is for commerce strategy.

In this article, we investigate the function of information exchange and knowledge discovery in virtual communities. There are two important inherent properties embedded in virtual communities (Wellman, 2001):

- **Social Networks:** When people interact with one another in virtual communities, they are inherently social. The typical case is CMC. As stated by Wellman (2001), computer networks principally support social networks, not groups. As a result, CMC has become part of people's lives rather than being a separate set of relationships.
- **Information and Knowledge Bearing:** If we view people in virtual community as information and knowledge bearers, they become an informa-

tion and knowledge base. These pieces of information and knowledge possessed by community members are the important assets of the virtual community.

These two properties indicate different views of the virtual community. The first reveals the networked nature of virtual communities and provides an interesting and novel perspective to explore and understand them. The second, our main target of research, represents an important function in virtual communities. In this paper, we discuss virtual community from these two perspectives, by taking social network property as the target to be explored, and by setting information navigation and knowledge discovery as the function to pursue and implement. For this purpose, the representation and phenomena of social networks shall be introduced. Thereafter, approaches and analysis for information navigation and knowledge discovery based on these phenomena and representation shall be investigated.

BACKGROUND

The Internet increases people's social capital, increasing contact with friends and relatives. Wellman (2001) proposed that computer networks are inherently social networks, linking people, organizations, and knowledge. For social networks to characterize the intrinsic property of virtual communities (Wellman & Gulia, 1999), techniques and theories from social network analysis are helpful to visualize, analyze, and understand virtual communities. However, there is much research on ethnographic studies and CMC that tends toward experimental contexts, while there are few studies on relational data (the ties and patterns of interaction amongst the participants in virtual communities) that is a direct indicator of the phenomena they are researching (Chen, 2002). These relational data embedded in virtual com-

munities naturally generates social networks. Any interaction among participants in virtual communities may be viewed as a kind of relationship.

In mathematics, the graph is the most important abstract model to characterize social networks. Vertices and edges in the graph can be instantiated by different contexts. There has been much research on observing, modeling, and analyzing graphs, not only from pure mathematics but also from physical phenomena in the real world. Therefore, graph representation of social networks in virtual communities provides a concrete theoretical base for the observation and analysis of the function of information navigation and knowledge discovery. In essence, graph representation of the virtual community encodes information of how information and knowledge flows and distributes in a virtual community.

Mathematically, a graph G is represented as a triplet $G=(V, E, f)$, where V is a set of vertices, E is a set of edges connecting some vertex pairs in V , and f is a mapping $f:E \rightarrow V$. In a virtual community, vertices represent unique participants who stand for information or knowledge nodes, and edges may represent their interactions that are dynamic and that may have different related values to stand for their roles in transferring information or knowledge. A graph is measured by a series of metrics from many related research areas, which shows insights into how a graph looks like and provides the necessary quantitative analysis. The important phenomena and approaches of graphs in the real world are represented by these metrics. The following shows a summary of some popular graph metrics:

- **Degree:** The number of edges connected to a vertex.
- **Path:** A sequence of distinct vertices (v_0, v_1, \dots, v_n) with edges connecting v_{i-1} and v_i for all $1 \leq i \leq n$.
- **Distance:** The number of edges in the shortest path between two vertices. For the entire graph, the average distance is computed over all pairs of vertices.
- **Clustering Coefficient:** the clustering coefficient of the vertex v whose degree is k edges is the ratio between the number E_v of edges that actually exist between these k vertices and the total number of $k(k-1)/2$.
- **Betweenness:** The number of paths between two other vertices that connect through this vertex.
- **PageRank:** It measures the authority of a vertex by summing the “votes” of vertices connected to it. It has become infamous through its prominent use by Google (Lawrence, Sergey, Rajeev, & Terry, 1999).

MAIN THRUST

Information Navigation and Knowledge Discovery in Virtual Communities

Virtual communities assemble information and knowledge from their participants and thus, actually become informal knowledge bases. Therefore, it is a crucial entity functioning as knowledge and information sharing (Burk, 2000). However, it is different from typical man-made knowledge bases by predefined and formulated rules and formats. In some sense, information and knowledge are stored in social networks, where nodes carry information and knowledge, and relationships indicate how and what information flows (interactions among community members). This knowledge and information organization format has the characteristics of mobility and indetermination—for its relationships and knowledge bearers often change and shift. This situation is different from that addressed by traditional organizational theory, which comprehends densely knit workgroups neatly structured in bureaucratic, hierarchical organizational tree (Contractor, 1999; Wellman, 1997). Besides, in such networks, finding knowledge and information becomes more important, as it does not provide the search function. In general, we often use the term search function to find and retrieve what we need in virtual communities. However, in many cases, the issue is to find out who knows what—a more complex task in virtual communities (Cross & Borgatti, 2000). In particular, how do people in virtual communities obtain knowledge from others when they do not know whom to ask? This question is of immediate practical importance to virtual communities. Normally, one attempts to examine the documentation or other help sources (these functions are mainly focused on searches based on the attributes of document contents) and then asks for familiar friends. The problem becomes acute, however, when the knowledge base is built on social networks. The commonly used retrieval of and access to the knowledge base are implemented by a typical search function. Clues for linking knowledge requesters and holders are not provided by such search functions.

Therefore, to address these issues, new techniques are needed to help people navigate and find knowledge in complex, fragmented, networked societies (Wellman, 2001). Finding knowledge in virtual communities can be redefined as information navigation and knowledge discovery in virtual communities. A formal model of information seeking was proposed (Borgatti, 2003) in which the probability of seeking information from another person is a function of (1) knowing what that person knows; (2) valuing what that person knows; (3)

being able to gain timely access to that person's thinking; and (4) perceiving that seeking information from that person would not be too costly. However, these works ignore the fundamental properties of virtual communities as social networks. Besides, many graph-based approaches can also play the important role to help people navigate and find information and knowledge in virtual communities. Next we shall explore the following two aspects for effective and efficient information navigation and knowledge discovery in virtual communities:

- Phenomena and Behavior Models
- Technically Algorithmic Approaches

Phenomena in Virtual Communities

Based on the graph representation of virtual communities, there are two surprising phenomena in social networks: “small world” and “power law.” They are the key to unraveling the information navigation and knowledge discovery, and to understanding behavior in virtual communities.

The power law (also called “scale-free”) phenomenon implies a highly skewed distribution that very few vertices maintain a large percentage of edges in the graph representations of social networks (Newman, 2003). This power law in the degree distribution reflects the presence of central community members who interact with many others and play a key role in relaying information and knowledge in virtual communities. At the social level, the work (Adamic, Lukose, Puniyani, & Huberman, 2001) supported the hypothesis that highly connected individuals do a great deal to improve the effectiveness of social networks in terms of access to relevant resources. In social networks of virtual communities, the origin of the power law phenomenon is well understood in terms of interactions among dynamic participants based on the graph growth and participants' preferential interactions, which is modeled as a “Barabási-Albert model” (Barabási & Albert, 1999).

The small-world phenomenon was first found in social networks by Milgram's experiment (Milgram, 1967), which is also known as “six degrees of separation.” Watts and Strogatz (1998) proposed a simple abstract model for the formation of the small-world phenomenon with two properties: being highly clustered like regular lattices, yet having small characteristic path lengths, like random graph. Therefore, the edges in the graph are divided into “local” and “long-range” contacts. The local neighborhoods lead to high clustering, while the long-range random connections lead to very short paths. This phenomenon reveals that behaviors in virtual communities are not uniformly random, but highly clustered and

with short average path between two vertices scaling logarithmically with graph size.

Information Navigation by Two Phenomena Properties

Social networks have the surprising property of being “searchable,” that is, ordinary people are capable of directing messages through their network of acquaintances to reach a specific but distant target person in only a few steps (Watts, Dodds, & Newman, 2002). Much work is focused on information navigation or seeking in social networks. For the intrinsically social networks property of virtual communities, they are important for designing and developing new algorithms or strategies of information navigation and knowledge discovery.

Based on the pioneering work of Watts and Strogatz (1998) who considered a “navigable” graph model, Kleinberg considered a more general problem of decentralized search in graphs with partial information about the underlying structure and performed algorithmic analysis. From the algorithmic perspective of small-world phenomenon, Kleinberg adopted the graph paradigm of Watts and Strogatz—rich in local connections, with a few random long-range connections, and began with the graph where vertices are placed on a specific two-dimensional lattice that has a simple “geographic” interpretation (Kleinberg, 2000). Based on this abstract model, an algorithm with local knowledge and limited structure of the graph can find the target in polylogarithmic time.

In the most general distributed search context, one may have very little information about the structure of graphs or the location of target. In virtual communities, one may have less information about geographic location. Therefore, Adamic (2001) proposed a number of efficient decentralized search algorithms that operate without structural information in power law graphs by exploiting the fact that the high connectivity vertices (with large degrees) play the important role of hubs in communication and networking.

Yet, these “navigable” models may not be satisfactory representations. The work by Watts (Watts et al., 2002) proposed a more realistic model that is based on plausible social structures and offered an explanation for the phenomenon of searchability. Social identity and similarity are the key components in this model. Their result suggests that searchability is a generic property of real-world networks. Their model may be applicable to the context of virtual communities, as there are reasonable assumptions.

Recently Dodds (Dodds, Muhamad, & Watts, 2003) conducted a global, Internet-based social-search experiment where more than 60,000 e-mail users attempted to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. Their results suggest that if individuals searching for remote targets do not have sufficient incentives to proceed, the small-world hypothesis will not hold. In contrast, even a slight increase in incentives can render social searches successfully under broad conditions. This is a promising result for virtual communities if the information and knowledge we search are clearly represented.

These two phenomenon properties in virtual communities offer tantalizing leads about information navigation and knowledge discovery. However, studies that emerge in laboratory to track the real virtual communities are needed.

Knowledge Discovery by Graph-Based Approaches

These phenomena or properties of virtual communities and many graph-based approaches play an important role in knowledge discovery. Although these approaches originated from different areas, their methodologies and ideas are useful and can be borrowed for exploring virtual communities in terms of the abstract and mathematical expressive property of graphs.

- **Social Network Analysis (SNA):** There is a long history in this area. Of the academic disciplines, academics in the social sciences observe networks of people using sociometric techniques and have been developing sets of techniques to provide both visual and mathematical analysis of human relationships. SNA is entering the mainstream, thanks to its better techniques for tracing relationships. For the inherent social network property of the virtual community, SNA is playing the important role in knowledge discovery of virtual communities.
- **Data Mining Methods for Graphs:** There has been much work in the data mining area to explore useful patterns or knowledge in large and real-world graphs. The typical work is AGM (Apriori-Based Graph Mining) system (Inokuchi, Washio, & Motoda, 2003). It focused on mining frequent substructures and thus, the problem of finding frequent item sets is generalized to frequent subgraphs. After the proposal of AGM, a family of graph-based data mining based on similar principles has been proposed. Other similar work in-

cludes the SUBDUE system (Cook and Holder, 2000), a faster algorithm FARMER (Nijssen and Kok, 2001), and the MolFea approach to find characteristic paths from the graph data (Raedt and Kramer, 2001).

- **Link Analysis:** This is a newly emerging research area. Link analysis is the process of building up graphs of interconnected objects through various relationships in order to discover patterns and trends. The main tasks of link analysis are to extract, discover, and link together sparse evidence from vast amounts of data sources to represent and evaluate the significance of the related evidence, and to learn patterns to guide the extraction, discovery, and linkage of entities. The most famous example of exploiting link structure is the use of links to improve information retrieval results. They include both the well-known PageRank measure (Lawrence et al., 1999) and Authority and Hub scores (Kleinberg, 1998).
- **Spectral Analysis:** The spectrum of a graph is a component of spectral analysis. Spectral graph theory is the study of relationship between a graph and the eigenvalues of matrices naturally associated to a graph (Chung, 1997). Eigenvectors of a graph is another component of spectral analysis, which has been successfully applied to link analysis of the Web for improving information retrieval.
- **Kernel Function-Based Analysis:** Kernel methods, especially the support vector machine, have become a popular tool for machine learning and data mining. They are computationally attractive because of the low cost of explicitly computing the feature map. Researches in kernel methods have begun investigating kernel functions defined on graphs (Gärtner, 2003). This is a new area for knowledge discovery in graphs.
- **Customer Relationship Analysis to Achieve Business Objectives:** Relationships among customers are of potential exploits in business strategies. Models and processes in analysis of the customer relationships are of importance to develop algorithms for knowledge discovery of interactions among community members and their similar social network properties. Ong proposed two types of customer profiles: active and passive (or potential), where potential customers are discovered for target marketing (Ong, Ng, & Lim, 2002). Also, Domingos and Richardson (2001) proposed the concept of "customer's network value," where they viewed the market as a social network and is modeled as a Markov random field.

CONCLUSION

Due to the two inherent properties of virtual communities, the function of information navigation and knowledge discovery in virtual communities is pursued by considering the graph representation of virtual communities, by analyzing the information navigation in terms of two important phenomena properties, and by investigating various graph-based approaches for knowledge discovery. These perspectives are reviewed from the views of information retrieval and knowledge discovery in graphs. Therefore, they may provide new insights for virtual communities.

REFERENCES

- Adamic, L. A., Lukose, R. M., Puniyani, A. R., & Huberman, B. A. (2001). Search in power-law networks. *Physical Review E*, 64, 046135.
- Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- Borgatti, S. P., & Cross, R. (2003). A relational view of information seeking and learning in social networks. *Management Science*, 49(4), 432–445.
- Chen, S. (2002). Conceptualizing online communities with social network analysis. *Proceedings of the International Sunbelt Social Network Conference*, New Orleans.
- Chung, F. (1997). *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.
- Contractor, N. S. (1999). *Management Communication Quarterly*, 13(154).
- Cook, D. J., & Holder, L. B. (2000). Graph-based data mining. *IEEE Intelligent Systems*, 15(2), 32-41.
- Cross, R., & Borgatti, S. (2000). *The ties that share: Relational characteristics that facilitate knowledge transfer and organizational learning*. Working paper of the Carroll School of Management, Boston College. Boston: Boston College.
- Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, 301, 827-829.
- Domingos, P., & Richardson, M. (2001). Mining the network value of customers. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 57-66).
- Gärtner, T. (2003). A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1), 49-58.
- Hagel, J., & Armstrong, A. G. (1997). *Net gain: Expanding markets through virtual communities*. Boston: Harvard Business School Press.
- Inokuchi, A., Washio, T., & Motoda, H. (2003). Complete mining of frequent patterns from graphs: Mining graph data. *Machine Learning*, 50(3), 321-354.
- Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 668-677).
- Kleinberg, J. M. (2000). The small-world phenomenon: An algorithmic perspective. *Proceedings of the 32nd ACM Symposium on Theory of Computing* (pp. 163-170).
- Lawrence, P., Sergey, B., Rajeev, M., & Terry, W. (1999). *The Pagerank citation ranking: Bringing order to the Web* (Technical Report). Stanford, CA: Stanford Digital Library Technologies Project.
- Milgram, S. (1967). The small-world problem. *Psychology Today*, 2, 60-67.
- Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2), 167-256.
- Nijssen, S., & Kok, J. N. (2001). Faster association rules for multiple relations. *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 891–896).
- Ong, K.-L., Ng, W.-K., & Lim, E.-P. (2002). Mining relationship graphs for effective business objectives. *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 561-566).
- Raedt, L. D., & Kramer, S. (2001). The Levelwise Version Space algorithm and its application to molecular fragment finding. *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, (pp. 853-862).
- Rheingold, H. (1999). *The virtual community: Homesteading on the electronic frontier*. Cambridge, MA: MIT Press.
- Watts, D. J., Dodds, P. S., & Newman, M. E. J. (2002). Identity and search in social networks. *Science*, 296, 1302-1305.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature* 393, 440-442.

Wellman, B. (1997). An electronic group is virtually a social network. In S. Kiesler (Ed.), *Culture of the internet* (pp. 179-205). Mahwah, NJ: Lawrence Erlbaum.

Wellman, B. (2001). Computer networks as social networks. *Science*, 293, 2031-2034.

Wellman, B., & Gulia, M. (1999). Virtual communities as communities: Net surfers don't ride. In M. A. Smith & P. Kollock (Eds.), *Communities in cyberspace*. London: Routledge.

KEY TERMS

Computer-Mediated Communication: The process by which people create, exchange, and perceive information using networked telecommunications systems (or nonnetworked computers) that facilitate encoding, transmitting, and decoding messages. Examples include Usenet, e-mail, and also cover some real-time chat tools such as lily, IRC, and even video conferencing.

Data Mining: A class of database applications or data processing that discovers hidden patterns and correlations in a group of data or large databases that can be used to predict future behavior.

Power Law Distribution: A probability distribution function, $P[X=x] \sim cx^{\pm}$, where constants are $c > 0$ and $\pm > 0$.

Small-World Phenomenon: A fact in some networks: most pairs of vertices are connected by a short path through the network, and most neighbors of each vertex are connected.

Social Networks: A set of people or groups of people with some pattern of contacts or interactions between them. These patterns can be friendships between individuals, business relationships between companies, and intermarriages between families.

Social Network Analysis: Social network analysis refers to the study of uncovering the patterning of people's interaction.

Social Capital: Social capital refers to the collective value of all social networks and the inclinations that arise from these networks to do things for each other.