

# Examining micro-level knowledge sharing discussions in online communities

G. Alan Wang<sup>1</sup> · Xiaomo Liu<sup>2</sup> · Jianling Wang<sup>3</sup> ·  
Min Zhang<sup>4</sup> · Weiguo Fan<sup>5</sup>

Published online: 2 June 2015  
© Springer Science+Business Media New York 2015

**Abstract** Online communities of practice have become a popular knowledge source for both individuals and organizations. It is important to understand how to facilitate virtual knowledge sharing in online communities. Existing studies generally focus on system design factors or motivations behind knowledge sharing behavior. In this study we aim to investigate the knowledge sharing processes in online communities and identify process patterns that are indicative of effective knowledge sharing processes. We propose a computational framework to examine individual knowledge sharing

processes in online communities from a process perspective. Our empirical evaluations show that effective knowledge sharing processes have distinct structural characteristics and communication network patterns compared to unhelpful knowledge sharing processes. Our research findings have practical implications for online community practitioners.

**Keywords** Online community · Knowledge sharing · Computer-mediated communication · Communication network · Network structural analysis

✉ Min Zhang  
zhangmin792002@tju.edu.cn

G. Alan Wang  
alanwang@vt.edu

Xiaomo Liu  
xiaomo.liu@thomsonreuters.com

Jianling Wang  
wangjl@mail.xjtu.edu.cn

Weiguo Fan  
wfan@vt.edu

<sup>1</sup> Department of Business Information Technology, Virginia Tech, 1007 Pamplin Hall, 880 West Campus Drive, Blacksburg, VA 24060, USA

<sup>2</sup> The R&D Department, Thomson Reuters, 3 Time Square, New York, NY 10036, USA

<sup>3</sup> Department of Accounting and Finance, School of Management, Xi'an Jiaotong University, 28 Xianning West Road, Xi'an, Shanxi 710049, China

<sup>4</sup> Department of Industrial Engineering, College of Management and Economics, Tianjin University, 92 Weijin Road, Tianjin 300072, China

<sup>5</sup> Department of Accounting and Information Systems, Virginia Tech, 3007 Pamplin Hall, 880 West Campus Drive, Blacksburg, VA 24060, USA

## 1 Introduction

With the widely available network access, Internet based online communities have become a popular and effective knowledge seeking and sharing platform in recent years. Online communities intended for knowledge sharing and problem solving are known as electronic networks of practice (Wasko and Faraj 2005) or online Communities of Practice (Wenger et al. 2002) where their participants share a similar concern or passion about a topic and deepen their knowledge and expertise by interacting on an ongoing basis. They are different from those online communities where discussions are more open-ended, less goal-oriented, and often do not really have a formal termination (e.g., political discussion communities and dating communities). In the rest of this paper, we use the term “online communities” in the sense of “online Communities of Practice”. Several studies show that online communities have become a popular external knowledge source for organizations because of increasing knowledge demands and limited availability of expertise and knowledge within the knowledge repository of their own organization (Constant et al. 1996; Wasko and Faraj 2005; Zhang and Watts 2003). Besides organizations and companies, individuals also use online

communities to seek knowledge and solve problems. For example, the end-users of products and services often use online communities to share their experiences and learn from each other (Lee et al. 2006). Over time successful online communities have accumulated a useful and valuable knowledge repository, an active user base, and a dynamic community. It is important to adequately understand how to leverage and share knowledge in online communities to maximize its value.

Online communities are crowd-sourcing knowledge-intensive business services (KIBS) that function as a facilitator, carrier or source of knowledge exchange and innovation (Hertog 2000). Compared to traditional KIBS providers, online communities serve a much diverse client base, which is made up of knowledge seekers in the community, and relies on community members rather than employees to provide services. A community member can be a client at one time and a knowledge provider at another. The business goal here is to facilitate and support knowledge sharing among community members. Table 1 lists several online community examples that support knowledge sharing. The business processes are the knowledge sharing processes embedded in online discussions. The business processes can be very different from traditional business processes such as tech support. For example, organizational hierarchies usually do not exist in open online communities. In addition, the knowledge-sharing processes in online communities only contain a very limited number of tasks (mostly problem-solving tasks), activities (posting and replying), and participant roles (knowledge seeker and knowledge sharer). That will affect the specificity (Goedertier et al. 2011) of the mined process models that would reveal little insight about the processes other than the sequences of posting and replying activities. Therefore, it is difficult to apply existing business process modeling techniques to understand and analyze knowledge sharing processes in online communities.

Despite the increasing number of online communities, little is known about how to facilitate or support knowledge sharing in online discussions. Existing studies on knowledge sharing in online communities have primarily focused on system design factors (Sharratt and Usoro 2003) or motivations behind knowledge sharing participation (Chiu et al. 2006; Setia et al. 2012; Wasko and Faraj 2005). Recent studies have examined the contents of

online knowledge sharing and used content analysis to find experts in online communities (Wang et al. 2013) or to predict the helpfulness of online knowledge sharing (Wang et al. 2011) or reviews (Cao et al. 2011). However, very few studies have examined the communication process of online knowledge sharing where dynamic communication patterns can be an indicator of effective knowledge sharing discussions. Among them, Yu et al. examined the communication patterns of an online mailing list and found two-way communication important for supporting effective knowledge exchange, reuse, and construction (Yu et al. 2010). The study shows that communication patterns do impact the effectiveness of knowledge sharing. It is necessary to further examine communication patterns in online communities and have a deeper understanding on their impact on knowledge sharing.

In this study we aim to understand and analyze online knowledge sharing discussions from a communication process perspective. The contribution of this research is three-fold. Firstly, online community practitioners will benefit from this research by promoting those communication patterns that improve effective knowledge sharing experiences. Secondly, our findings can help online community users identify and engage in those discussions that are more likely to yield effective knowledge sharing outcomes. Lastly, our research will help better understand the knowledge sharing processes in online communities and call forth more research that focuses on those processes.

The rest of the paper is organized as follows. Section 2 reviews related work on knowledge sharing in online communities. Section 3 introduces a computational framework for analyzing individual knowledge sharing processes in online communities. Section 4 describes our empirical evaluations and report research findings. Section 5 provides a conclusion and a discussion on the practical implications of our research findings and the limitations of our research.

## 2 Related work

Past research has used communication networks and structural network analysis to reveal aggregated community-level communication patterns in online communities. Very few studies focused at how communication patterns may affect the effectiveness of individual knowledge sharing discussions. Yu et al. (2010) showed that communication patterns were indeed important for supporting knowledge exchange, reuse, and construction. In this section, we review community-level communication network studies and structural network analysis methods, with which we propose to focus on studying communication patterns in individual knowledge sharing discussions instead of aggregated patterns at the community-level.

**Table 1** Examples of online communities supporting knowledge sharing

Online communities	Example
Consumer support communities	The Apple Support Community
Developer communities	The Oracle Technology Community
Corporate intranet communities	(Leshed 2009)

## 2.1 Communication networks in online communities

An online community is often viewed as a social network where users of similar interests share information or knowledge within a knowledge domain. In this network, members of the community engage in continuous social exchanges online with diverse motivations. The social cognitive theory (Bandura 1986) has been used to explain motivations behind user participation in online communities. Some participate out of pure intrinsic motivation such as friendship and loyalty, while others participate more out of extrinsic factors such as recognition within the communities, rewards, increased reputations, or a feeling of obligation to share (Constant et al. 1996; von Hippel and von Krogh 2003; Peddibhotla and Subramani 2007; Preece 2000; Wasko and Faraj 2005). Those motivation studies remain focused at individual-level analysis, addressing the question of when and why individuals participate in online communities. However, due to individual cognitive differences, some findings on participation motivation can cause conflicting explanations (Faraj and Johnson 2011). Communication networks are used to examine aggregated individual behaviors in online communities at the community-level. This perspective provides a complementary understanding to user participation behaviors and motivations in online communities by overlooking individual differences.

Communication network has been used to identify aggregated communication patterns in online communities at the community-level. For example, Nolker and Lina (2005) built a communication network based on discussions in an online community. They used network centrality measures to distinguish different user roles (e.g., leaders, motivators, and chaters). Moreover, De Moor (2006) used a communication network to identify frequent communication patterns, which is a set of communicative workflows and norms describing acceptable and desired communicative interactions within a community. Faraj and Johnson (2011) studied frequent communication exchange patterns in online communities from the communication network perspective. Their study is built on the dual aspect of online interactions: they are *social* exchanges that take place between participants within a *network* context. In other words, regardless of resources exchanged, facts, know-how, answers to questions, or social niceties, the interactions within online communities are social in nature. Moreover, online community interactions take place within the context of a social network, and are mediated by the communication network. All user posts are visible to all other participants and are organized in discussion threads. Guided by both the social exchange theory (Cook and Rice 2001) and the network exchange theory (Burke 1997), Faraj and Johnson proposed three network exchange patterns that they claimed to be visible in all types of online communities: direct reciprocity (user *A* will reply to user *B*'s post as a direct reciprocity for user *B*'s prior help to user *A*), indirect reciprocity (After user *B*

helped user *A*, user *A* will help user *C* as an indirect reciprocity), and preferential attachment (new actors choose to interact with already well-connected others: both user *A* and user *C* choose to interact with a prominent user *B*). Using five online discussion forums, they found consistent results across five online communities that the pattern of ties is consistent with norms of direct reciprocity and indirect reciprocity, and has a tendency away from preferential attachment. Their results across five online communities show the same network exchange tendencies (direct and indirect reciprocity, a tendency away from preferential attachment) but do so in different magnitudes.

Those studies show that communication network is a useful tool for finding frequent communication patterns without addressing individuals' motivational differences.

## 2.2 Network structural analysis methods

Interesting network structural characteristics such as power-law and scale-free are frequently observed for online social networks such as Flickr, YouTube, and LiveJournal (Cheng et al. 2008; Kumar et al. 2006; Mislove et al. 2007), and for blogs (Chau and Xu 2012). When analyzing online communities from a network perspective, network structural measures can reveal the importance of users in the social network, the characteristics of communication patterns, and social structures that emerge from online discussions (Wassermann and Faust 1994). Adamic et al. (2008) calculated network structural measures for a communication network constructed based on a group of online discussions. Their intention was to find those users who were more likely to provide answers to questions. These network structural characteristics can provide meaningful guidance for efficient information dissemination and effective knowledge sharing in network-based systems (Mislove et al. 2007).

## 2.3 Summary

Online community studies based on communication networks are mostly focused at the network-level or community-level analysis, which is built upon aggregated knowledge sharing processes in an online community. There are very few micro-level studies that focus on individual knowledge sharing processes (i.e., individual discussion threads). To the best of our knowledge, the only exceptions are Gómez et al. (2008) and Laniado et al. (2011). Both studies used tree graphs and network graphs to represent individual knowledge sharing discussions and observed discussion-level structural properties such as self-similarity. A tree graph is a special network type that captures the post-reply relationships in a discussion. In a tree graph, each node represents a discussion participant while each edge denotes a replying relationship between two participants. A tree graph can be transformed into a network graph

by merging participant nodes. Although those discussion-level analyses revealed interesting discussion patterns, they did not utilize the network structural measures and fail to show the significance of the discussion-level communication patterns related to effective knowledge sharing in online communities.

### 3 Examining online knowledge sharing processes: A micro-level analysis

In this study, we propose a new computational framework for analyzing individual knowledge sharing processes in online communities. The objective is to find communication patterns that are related to effective knowledge sharing in online communities. Our proposed framework, illustrated in Fig. 1, consists of five steps. Steps 1–2 are the data processing stage, which downloads and extracts data needed for modeling online knowledge sharing processes. Steps 3–5 are the stage of communication process modeling and structural analysis, which seeks to uncover the insights about online knowledge sharing patterns useful for online community practitioners. The crawling step downloads online discussion threads from an online knowledge community. An introduction to Web crawling can be found in (Baeza-Yates and Ribeiro-Neto 1999). In the rest of this section, we describe metadata extraction, communication process modeling, network structural analysis, and communication pattern mining in details.

#### 3.1 Metadata extraction

Online communities for knowledge sharing are often designed in the format of a discussion forum. According to the Big Boards (<http://www.big-boards.com/statistics/>), the majority

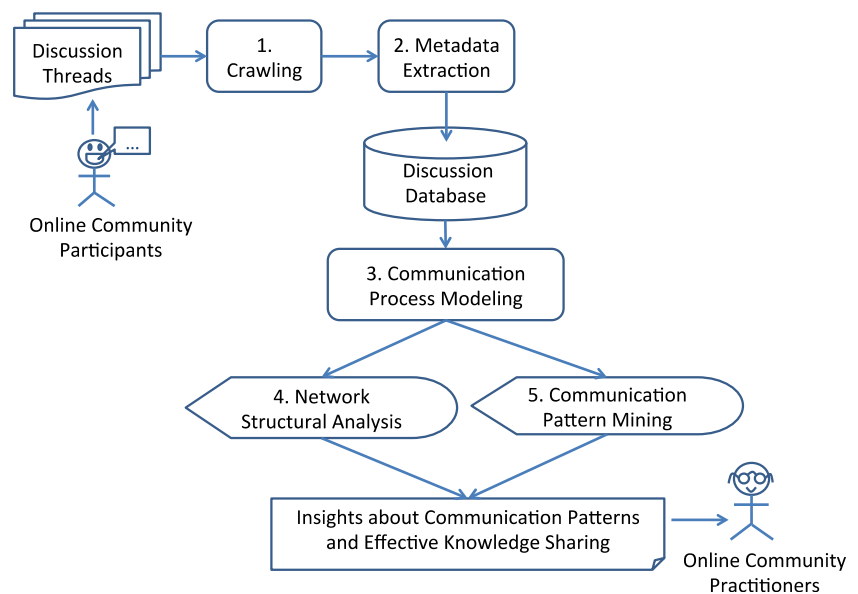
of discussion forums uses one of three service providers, namely vBulletin, Invision, and phpBB. All three forum providers organize online discussions into discussion threads, each of which has a starting post followed by replying posts. Each post contains not only text content, but also metadata including an author identifier (e.g., a user name), a time stamp, a post ID, a thread ID, the post ID of the previous post being replied to, and sometimes a user feedback indicator. Most advanced online communities such as the Apple Support Communities and Oracle's Java Programming Forum allow users to provide helpfulness feedback for reply posts. Each online community may have a different way of indicating users' helpfulness feedback. In this research we consider three types of user feedback: *solved* (a satisfactory solution is obtained in a discussion), *helpful* (helpful information is obtained but not an actual solution), and *unhelpful*. After crawling all web pages from an online discussion forum, we can acquire the metadata of online discussion threads by parsing those retrieved pages. Table 2 shows an example of online discussion metadata.

#### 3.2 Communication process modeling

Based on the metadata extracted from the previous section, we create post trees and communication networks for individual knowledge sharing processes based on extracted posts, authors of posts, and reply relationships between the posts.

In most online discussion forums, the posts of a thread are organized chronologically, from oldest to latest, as shown in Fig. 2a. Given a problem-solving oriented discussion thread, the knowledge seeker makes the first post describing the problem. Other knowledge sharers as well as the seeker himself are engaged in the discussion by posting to the same thread. We can then build a post tree that captures the replying

**Fig. 1** A micro-level computational framework for online knowledge sharing processes



**Table 2** An example of metadata extracted from online discussion threads

Thread ID	Helpfulness feedback	Post ID	Post replied to	Post author	Post type
1	Solved	1.1	N/A	A	Starting post
1		1.2	1.1	B	Reply
1		1.3	1.1	C	Reply
2	Helpful	2.1	N/A	D	Starting post
2		2.2	2.1	B	Reply

relationships among the posts using a tree-like view shown in Fig. 2b. That view helps us understand the conversational structure of knowledge sharing discussions. Furthermore, based on the post authorship and replying relationships, we can convert the post tree into a micro-level or discussion-level communication network, which illustrates the communication interactions among the discussion participants. The network view helps us identify important communication patterns that may improve the effectiveness of knowledge sharing. We now provide formal definitions for a post tree and a micro-level communication network based on the graph theory.

**Definition 1. (Post tree)** A post tree is a directed graph  $T=P, E$  with no cycles. A tree node set  $P$  contains all posts in a discussion thread. The root node of  $T$  denotes the starting post that contains a question. A tree edge set  $E$  consists of all reply relationships among all posts in the thread. A directed edge  $e=u, v$  indicates that post  $u$  replies to post  $v$ .

**Definition 2. (Micro-level communication network)** A micro-level communication network is a directed graph  $=U, E$  created for a discussion thread. Each node  $u_i \in U$

corresponds to a user while each edge in this graph  $e_i=(u_i, u_j)$  represents a reply relationship between two participants in a discussion thread.

### 3.3 Communication pattern mining

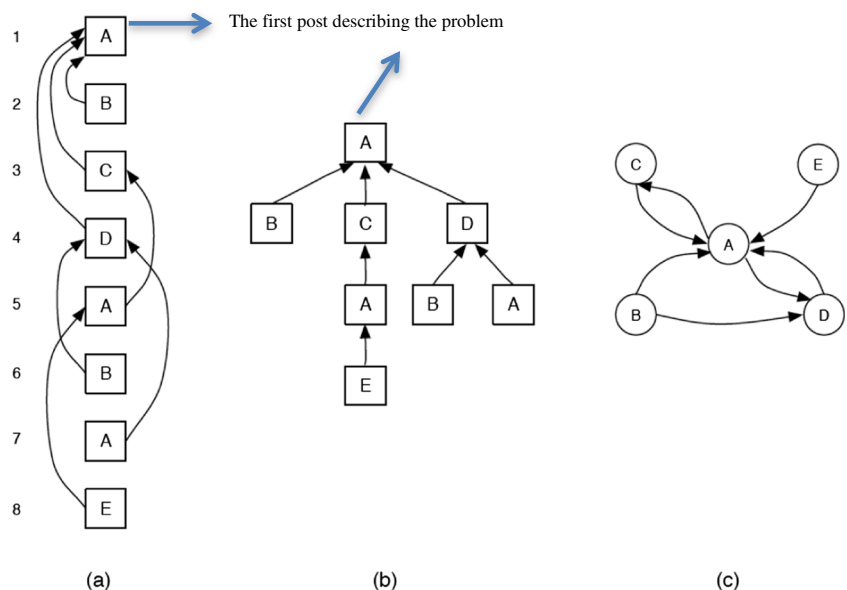
The post tree and micro-level communication network are used to capture an online knowledge sharing process in two different aspects. Our aim is to find the communication patterns embedded in the two process representations that may be associated with effective knowledge sharing leading to helpful knowledge or problem solutions. Network analysis methods have been commonly used to extract structural characteristics and patterns for graph-based representations. In the rest of this section, we describe several commonly used network analysis measures such as network structural characteristics and sub-graph mining for detecting important communication patterns.

#### 3.3.1 Network structural analysis

**Depth and width** Given a post tree  $T$ , its depth is the length of its path from its root to the deepest node of the tree. Its width is simply defined as the number of branches (i.e., degree) that  $T$ 's root node has. For example, the depth and width of the post tree in Fig. 2b are both 3.

**$h$ -index** To overcome the bias of post quantity, Gomez et al. proposed a modified version of  $h$ -index to measure the controversy of a thread discussion (2008). The  $h$ -index was originally invented to measure the scientific contribution of a researcher based on his/her publication quantity and the number of citations each publication received. The  $h$ -index of a discussion thread is defined similarly: Given

**Fig. 2** a Posts in a discussion thread with arrows representing reply relationships; b a post tree; and c a communication network. Squares denote posts in a thread. Circles denote post authors (A, B, ..., E). Arrow lines represent the reply direction between two posts or users



the post tree of a thread, the  $h$ -index value  $h$  is its maximum depth  $h$  that has  $i$  replies where  $i \geq h$ . For example, the  $h$ -index of the post tree in Fig. 2b is 3 because the depth level 3 ( $h$ ) has three posts ( $i$ ).

**Dialogue length** This is a heuristic-based measure. We argue that a thread would be more likely to reach a solution if the knowledge seeker and sharers have a long conversation between them. Therefore, we define a dialogue as a series of continuous replies between a knowledge seeker and a sharer. The dialogue length is the number of replies in the longest conversation. For example, the dialogue length of the post tree in Fig. 2b is 4.

**Reciprocity** Reciprocity measures the tendency to form a mutual connection between vertex pairs in a directed network. In other words, it defines how often bi-directional links occur between any two nodes. It is important to study two-way communications because it has shown to be important in knowledge exchange, reuse, and construction (Yu et al. 2010). Traditionally, the reciprocity  $r$  is defined as the ratio of the number of bi-directional edges  $L$  to the total number of edges  $L$  in a directional network (Wassermann and Faust 1994).

$$r = \frac{\overset{\leftrightarrow}{L}}{L}$$

When  $r=1$ , it is a fully connected network with all connections bidirectional. When  $r=0$ , it is a network where bi-directional links do not exist at all. Because of some conceptual problems with  $r$ , Garlaschelli and Loffredo (2004) defined a reciprocity coefficient  $\rho$  as the correlation coefficient between the entries of the adjacency matrix of a directed graph,

$$\rho = \frac{\sum_{i \neq j} (a_{ij} - \bar{a})(a_{ji} - \bar{a})}{\sum_{i \neq j} (a_{ij} - \bar{a})^2} = \frac{r - \bar{a}}{1 - \bar{a}}$$

where  $a_{ij}=1$  if a link from  $i$  to  $j$  exists and 0 otherwise. Both reciprocity measures are used in this study.

**Cluster coefficient** The local cluster coefficient of a vertex in a graph measures how well its neighbors are connected to be a clique, in which every two neighbors are connected. Consider a graph  $G$  with vertices  $i=1, \dots, n$ . Let  $k_i$  be the number of neighbors that vertex  $i$  has and  $n_i$  be the number of edges between its neighbors. The local cluster coefficient for vertex  $i$  is defined as

$$C_i = \frac{n_i}{\binom{k_i}{2}}$$

The global cluster coefficient for the whole graph is calculated as the average of the local cluster coefficients of all  $n$  vertices (Watts and Strogatz 1998),

$$\bar{C} = \frac{1}{n} \sum_{i=1}^n C_i.$$

### 3.3.2 Network sub-graph mining

We use the sub-graph mining technique to identify frequent communication patterns embedded in post trees and micro-level communication networks that are more likely to lead to effective knowledge sharing. Sub-graphs are basic structural elements of many types of natural networks such as chemical compounds (Yan and Han 2002), protein structures (Huan et al. 2004), and social networks (Lahiri and Berger-Wolf 2008). Studies have demonstrated that mining common sub-structures in graphs is crucial to understanding the interactions and dynamics at work in those graphs. Generally, the problem of sub-graph mining can be formalized as follows. Given a database with  $n$  graphs  $G = \{G_1, \dots, G_n\}$ , the support level of a sub-graph  $g$ ,  $\text{sup}(g)$  is the percentage of graphs in  $G$  that contains the sub-graph  $g$ ,

$$\text{sup}(g) = \frac{|\delta_G(g)|}{n}.$$

Sub-graph mining algorithms can then discover all sub-graphs with the minimum support ( $\sigma$ ) where  $\text{sup}(g) \geq \sigma$ . The pseudocode of the sub-graph mining algorithm is presented in Fig. 3. The most computationally intensive parts in the algorithm are testing sub-graph isomorphism in the subroutine *subgraph-isomorphism()* and generating candidate sub-graphs from size  $k$  to  $k+1$  in the subroutine *candidate-generate()*. Existing sub-graph mining algorithms employ various strategies to speed up these two processes.

We propose to use an efficient sub-graph mining algorithm named gSpan (Yan and Han 2002) for frequent communication pattern discovery in online discussions. gSpan is a DFS (Depth-First-Search) based algorithm. In a comparative survey of algorithms for frequent sub-graph discovery, DFS-based algorithms were found to be much more efficient than other algorithms (Krishna et al. 2011). In their experiment, the search of sub-graphs was completed within minutes given a dataset with more than 42,000 graphs.

Sub-Graph Mining Algorithm

Input: a graph database  $G = \{G_1, \dots, G_n\}$ , a minimum support  $\sigma$

Output:  $g = \{g_1, \dots, g_k\}$  a set of frequent sub-graphs of cardinality 1 to  $k$

$C_1 \leftarrow$  all sub-graphs of cardinality 1

$g_1 \leftarrow \{c \in C_1 \mid \text{sup}(c) = c.\text{count}/n \geq \sigma\}$

for ( $k=2$ ;  $g_{k-1} \neq \emptyset$ ;  $k++$ ) do

**Fig. 3** Sub-graph mining algorithm

```

for ( $k = 2; g_{k-1} \neq \emptyset; k++$ ) do
     $C_k \leftarrow$  candidate-generate( $g_{k-1}$ )
    for each graph  $G_i \in G$  do
        for each candidate  $c \in C_k$  do
            if subgraph-isomorphism( $c, G_i$ ) then
                 $c.count++$ 
            end
        end
    end
     $g_k \leftarrow \{c \in C_k | \text{sup}(c) = c.count/n \geq \sigma\}$ 
end
return  $g \leftarrow \cup g_k$ 

```

```

 $C_k \leftarrow$  candidate-generate ( $g_{k-1}$ )
for each graph  $G_i \in G$  do
    for each candidate  $c \in C_k$  do
        if subgraph-isomorphism( $c, G_i$ ) then
             $c.count++$ 
        end
    end
end
 $g_k \leftarrow \{c \in C_k | \text{sup}(c) = c.count/n \geq \sigma\}$ 
end
return  $g \leftarrow \cup g_k$ 

```

## 4 Empirical evaluations and results

### 4.1 Data collection

We evaluated our computational framework using real data downloaded from two large online communities: the *Apple Support community* for consumer product support and the *Sun's Forums* for software development support. These two communities represent two major types of online communities: consumer communities and developer communities. Both communities are IT related, problem-solving oriented, and large in size. The Apple Support Communities, which attract a large number of Apple product consumers, has a much diverse user base than the Sun's Forums in terms of the users' technical background and technology literacy. Users in the Sun's Forums are mainly IT developers who are familiar with technology and share a relatively focused interest.

To limit the scope of our study, we focused on one of the most active sub-forums from each of the two communities.

Specifically, we crawled 49,343 online discussion threads published between June 29, 2007 and April 6, 2010 from the *Using iPhone* forum in the *Apple Support Community*, and 70,488 threads published between June 19, 2001 and June 6, 2009 from the *Java programming* forum in the *Sun's Forums*. We removed 6702 and 6979 discussion threads from the iPhone and Java data sets respectively because they did not have any reply.

A unique feature provided by these two online communities is the ability to let knowledge seekers rate replies provided by knowledge sharers. The seeker can rate a reply with either a "This helped me" (helpful) tag or a "This solved my question" (solved) tag based. Table 3 shows the total number of threads with at least one reply and the number of threads with different user feedback labels in each dataset. Those threads without any user feedback do not necessarily mean that they contain no helpful knowledge or solutions. Past research has found that user feedback participation is often very low in virtual communities (Cao et al. 2011). Due to users' negligence or unwillingness, many discussion threads in the two forums that are without any user feedback, however, may contain helpful knowledge or even problem solutions. We decided to manually examine those threads without feedback and identify those discussion threads that failed to be productive in knowledge sharing.

From each data set, we randomly selected unlabeled discussion threads without replacement and asked two Computer Science graduate students to manually examine whether they contained solutions or helpful information. From each dataset we identified 1017 threads that both experts agreed upon not containing any helpful knowledge to the question being discussed. We then randomly selected 500 solved discussion threads, 500 helpful threads, and 500 "unhelpful" threads.

**Table 3** Statistics of discussions threads with different user feedback labels

Dataset	Number of threads with replies	Number of solved threads	Number of helpful threads	Number of threads without user feedback
iPhone	42,641	12,453 (29.2 %)	1140 (2.7 %)	29,048 (68.1 %)
Java	63,509	5777 (9.1 %)	1260 (2.0 %)	56,472 (88.9 %)

During the selection process, we made sure that each selected thread was problem solving oriented. Table 4 shows some basic statistics of the three types of discussion threads.

We can observe that effective knowledge sharing discussions, i.e., the discussions threads labeled with a solved or helpful tag, have significantly more posts and participants than unhelpful discussions (i.e., threads manually identified as unhelpful). The observation is intuitive. A discussion thread with many posts indicates an active knowledge sharing process, which often leads to a positive outcome. Similarly, if a discussion attracts more participants, it will be more likely to have an active sharing process as well as diverse knowledge sources.

We also observe differences between the two datasets. The iPhone forum has slightly more users participating in each discussion thread on average. However, the average number of posts in the iPhone forum is fewer than that in the Java forum. It shows that Java forum participants, who have less diverse technical background and share a more focused interest, are more committed to discussions and contribute more than iPhone forum participants.

## 4.2 Results

We parsed each selected discussion thread to get its metadata before we constructed a post tree and a micro-level communication network. We then calculated the seven network structural metrics described in Section 3.3.1 for all 3000 threads in our sample datasets. We notice that the number of posts in a thread may significantly affect most metrics. Therefore, discussion threads with different lengths cannot be directly comparable. We grouped the selected threads by their thread type (solved, helpful, and unhelpful) and length (i.e., number of posts). Discussion threads with 12 or more posts were discarded because they were few and did not have enough

**Table 4** Basic statistics of the three types of discussion threads

Dataset	Solved threads		Helpful threads		Unhelpful threads	
	Average number of posts	Average number of users	Average number of posts	Average number of users	Average number of posts	Average number of users
iPhone	6.37	4.03	6.54	4.04	3.91	2.99
Java	7.38	3.64	7.92	3.87	4.69	2.95

data points to estimate accurate means. We then calculated the mean values of those metrics for the threads with the same length in each thread group. Finally, we computed the average network structural metrics across thread groups with different lengths for each thread type.

### 4.2.1 Differences in network structural characteristics

Table 5 shows the comparison on the seven network structural metrics across the three thread types. The significance level  $\alpha=0.05$  was used to determine the statistical significance of the differences between solved/helpful threads and unhelpful threads. All structural metrics except *h*-index showed significant differences when comparing solved or helpful threads to unhelpful threads.

The results show that effective knowledge sharing discussions do have distinct structural characteristics compared to unhelpful discussions. An effective knowledge sharing discussion tends to have more depth and less width. It suggests that a knowledge sharing process with a limited number of engaged dialogues is more likely to yield productive knowledge sharing. The dialogue length values suggest that helpful or solved discussions are more likely to have a long conversation between the knowledge seeker and at least one particular knowledge sharer than unhelpful discussions. Results in reciprocity and clustering coefficient show that effective knowledge sharing processes have higher reciprocity (two-way communications) and more cohesive discussions among

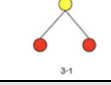
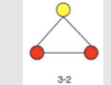
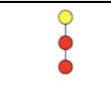
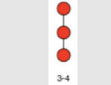
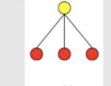
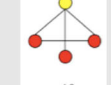
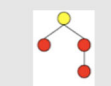
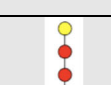

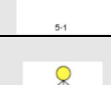
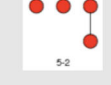
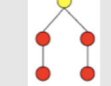

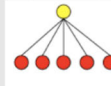
**Table 5** The differences in network structural characteristics across different thread types and datasets

Metrics	iPhone			Java		
	Solved threads	Helpful threads	Unhelpful threads	Solved threads	Helpful threads	Unhelpful threads
Depth	<b>3.02*</b>	<b>3.18*</b>	2.77	<b>4.11*</b>	<b>4.17*</b>	3.91
Width	<b>2.40*</b>	<b>2.26*</b>	2.53	<b>1.63*</b>	<b>1.58*</b>	1.78
<i>h</i> -index	2.42	2.41	2.41	2.28	2.29	2.27
Dialogue	<b>2.25*</b>	<b>2.43*</b>	2.04	<b>2.84*</b>	<b>2.89*</b>	2.43
Reci. Ratio	<b>0.46*</b>	<b>0.51*</b>	0.38	<b>0.64*</b>	<b>0.64*</b>	0.52
Reci. Coef.	<b>0.02*</b>	<b>0.07*</b>	-0.04	<b>0.13*</b>	<b>0.14*</b>	0.05
Clus. Coef.	<b>-0.04*</b>	<b>0.01*</b>	-0.1	<b>0.05*</b>	<b>0.08*</b>	-0.04

\*:  $p$ -value<0.05



**Table 6** Support values of sub-graph patterns (a light-colored node represents a knowledge seeker, a dark-colored node represents one who replied)

#	Pattern Illustration	Explanation	iPhone			Java		
			U	H	S	U	H	S
3-1		Communications between an seeker and two sharers	69%	67%	66%	64%	68%	66%
3-2		Communications between an seeker and two sharers, and between the two sharers	5%	<b>19%</b>	<b>21%</b>	9%	<b>23%</b>	<b>22%</b>
3-3		A seeker communicates with one sharer who communicates to another	33%	32%	36%	34%	35%	33%
3-4		Three sharers communicate with one another	5%	<b>16%</b>	<b>15%</b>	8%	<b>14%</b>	<b>15%</b>
4-1		A seeker communicates with three sharers	22%	<b>36%</b>	<b>33%</b>	23%	<b>35%</b>	<b>35%</b>
4-2		A seeker communicates with three sharers, two of which have communication	3%	<b>15%</b>	<b>13%</b>	7%	<b>15%</b>	<b>17%</b>
4-3		A seeker communicates with two sharers, one of which communicates to another sharer	20%	<b>26%</b>	<b>29%</b>	23%	<b>28%</b>	<b>28%</b>
4-4		A seeker and three sharers are engaged in a single dialogue	5%	<b>15%</b>	<b>15%</b>	8%	<b>12%</b>	<b>13%</b>
5-1		A seeker communicates with 4 different sharers	11%	<b>18%</b>	<b>18%</b>	8%	<b>16%</b>	<b>17%</b>
5-2		A seeker communicates with 3 different sharers, one of which communicates with another sharer	12%	<b>20%</b>	<b>20%</b>	11%	<b>16%</b>	<b>18%</b>
5-3		A seeker and 4 sharers are engaged in two parallel dialogues	2%	<b>8%</b>	<b>11%</b>	3%	<b>6%</b>	<b>5%</b>
5-4		A seeker and 4 sharers form a long dialogue and a short one	3%	<b>14%</b>	<b>12%</b>	5%	<b>9%</b>	<b>10%</b>
6-1		A seeker communicates with 5 sharers	2%	<b>10%</b>	<b>12%</b>	2%	<b>7%</b>	<b>7%</b>
6-2		A seeker communicates with 4 sharers, one of which communicates with another sharer	4%	<b>13%</b>	<b>10%</b>	4%	<b>9%</b>	<b>10%</b>

**Bold font:** Support levels in helpful and solved threads are higher than those in unhelpful threads

different discussion participants than unhelpful processes. The above findings are consistent across the two datasets.

The differences in the metrics between solved and helpful threads were statistically insignificant on the Java dataset. For the same comparisons on the iPhone data set, only the dialogue, reciprocity and cluster coefficient metrics had significant *p*-values. Therefore, there is no conclusive evidence that these structural metrics can be used distinguish solved threads from helpful threads.

Comparing the metrics between the two datasets, solved and helpful threads in the iPhone forum have significantly smaller depth, larger width, shorter dialogue, less reciprocity, and lower clustering coefficient than those in the Java forum. That can be explained by the fact that users in the Java community, compared to those in the iPhone forum, have a focused interest and are more committed to problem-solving discussions.

#### 4.2.2 Differences in communication network patterns

We used the gSpan algorithm (Yan and Han 2002) to discover sub-graph patterns from the communication networks with a minimum support level of 10 %, which is the default threshold used by gSpan. The minimum support level of 10 % means that a sub-graph is considered as a frequent pattern if at least 10 % of the communication networks contain the sub-graph. Table 6 lists support values for those sub-graph patterns in the three types of discussion threads. We compared the frequencies of patterns in “unhelpful” with those appeared in “helpful” and “solved” threads.

First, we found that those patterns with at least four nodes all had higher supports in helpful and solved threads than in unhelpful threads. It suggests that having diverse participants (at least three different knowledge sharers) improves the likelihood of having a helpful discussion.

Second, increased interactions among knowledge sharers can improve the chance of having a helpful discussion. Among 3-node patterns, patterns 3–1 and 3–3 had almost the same level of support across the three thread groups. However, patterns 3–2 and 3–4, both of which showed interactions between knowledge sharers, had much higher support in helpful and solved threads than unhelpful threads. A similar trend exists in 4-node patterns. The support values of those patterns with more interactions between sharers (e.g., patterns 4–2 and 4–4) had a significant increase in helpful and solved threads over the corresponding support values in unhelpful threads. The finding suggests that it is important to have interactions among knowledge sharers in problem-solving oriented discussions. It is consistent with group learning literature that fully-connected or cliquish communication patterns improve group learning performance (Chen et al. 2003).

Lastly, unhelpful threads do not exhibit distinct communication patterns. Patterns 3–1 and 3–2 are common in all three

types of discussion threads. It suggests that unhelpful threads are often caused by insufficient user participation and insufficient engagement in discussion dialogues. Both the difficulty level of the question and lack of communication can contribute to insufficient participation. In this research, we focus on the problem of lack of communication. To avoid unhelpful discussions, we call for online community practitioners to design mechanisms for encouraging user participation.

## 5 Conclusions and discussions

Online communities designed to support knowledge sharing have become an important knowledge source. While most existing studies on online communities focus on system design factors or motivations behind user participation, we aim to decipher the online knowledge sharing processes and identify communication patterns associated with effective knowledge sharing. In this paper we proposed a computational framework to examine individual knowledge sharing processes in online communities from a communication process perspective. Our empirical evaluations showed that, compared to unhelpful knowledge sharing processes, effective knowledge sharing processes exhibit distinct structural characteristics using post tree and communication network representations.

Our research findings on knowledge sharing processes in online communities have practical implications for online community practitioners. The practitioners should encourage those structural characteristics and communication network patterns that we found more likely to appear in effective knowledge sharing processes. For example, we suggest that knowledge sharing processes with a limited number of engaged discussion dialogues is more likely to have effective knowledge sharing. Online community practitioners can choose up to two replies directly replying to the original question post based on relevance and filter out the rest of direct replies that are less relevant. Therefore, discussion dialogues can be focused on two selected reply posts. We also suggest that effective knowledge sharing has a longer dialogue length. Online community may develop a reputation mechanism that rewards activities contributed to long dialogues (e.g., with at least three continuous replies in one dialogue). Similarly, we found that effective knowledge sharing had higher reciprocity and more cohesive discussions among discussion participants than unhelpful discussions. A reputation mechanism can be used to reward reciprocal postings and those posts that improve the cluster coefficient of a discussion process (i.e., increase the number of interactions among discussion participants in the same thread).

Our study still has several limitations and can be extended in a number of directions. First, our computational framework only examines the structural characteristics and communication patterns of knowledge sharing processes without directly

considering the content and information quality of individual posts. Although user feedback ratings can be considered as quality inputs, they are only available for very few posts in a discussion thread if there is any. Some posts may be irrelevant to the problem raised in the original posting or may have low information quality. Those posts contribute little to problem solving and knowledge sharing, but potentially introduce biases into the structural metrics calculation and communication patterns. Second, we currently do not consider the nature of the problem in a discussion, specifically the difficulty level of the problem. A knowledge sharing process may not be productive due to the problem being too difficult or complex to resolve rather than the patterns in the knowledge sharing process. Removing those problems would further improve the validity of our findings. Lastly, although we conducted empirical evaluation using two different online communities in two different domains, we could still further improve the external validity of our research by evaluating more knowledge sharing processes in other online communities. We expect that similar result would be obtained from other tech support forums. However, some online communities that have unique characteristics should be further examined. For example, communication patterns may change in those online communities where users are identifiable and offline relationships may affect online communication patterns (e.g., corporate intranet communities (Leshed 2009)).

## References

- Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and Yahoo answers: everyone knows something. Proceedings of the 17th International Conference on World Wide Web—WWW '08, Beijing, China, pp. 665–674.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: Addison-Wesley.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs: Prentice-Hall.
- Burke, P. (1997). An identity model for network exchange. *American Sociological Review*, 62, 134–150.
- Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the “Helpfulness” of online user reviews: a text mining approach. *Decision Support Systems*, 50(2), 511–521.
- Chau, M., & Xu, J. (2012). Business intelligence in blogs: understanding consumer interactions and communities. *MIS Quarterly*, 36(4), 1189–1216.
- Chen, G. D., Wang, C. Y., & Ou, K. L. (2003). Using group communication to monitor web-based group learning. *Journal of Computer Assisted Learning*, 19(4), 401–415.
- Cheng, X., Dale, C., & Liu, J. (2008). Statistics and social network of youtube videos. Proceedings of the 16th International Workshop on Quality of Service, Enschede, pp. 229–238.
- Chiu, C. M., Hsu, M. H., & Wang, E. T. G. (2006). Understanding knowledge sharing in virtual communities: an integration of social capital and social cognitive theories. *Decision Support Systems*, 42(3), 1872–1888.
- Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: the usefulness of electronic weak ties for technical advice. *Organization Science*, 7(2), 119–135.
- Cook, K., & Rice, E. (2001). Exchange and power: Issues of structure and agency. In J. H. Turner (Ed.), *Handbook of sociological theory* (pp. 699–719). New York: Kluwer Academic Publishers.
- De Moor, A. (2006). Community memory activation with collaboration patterns. Proceedings of the 3rd International Community Informatics Conference (CIRN 2006), Monash Centre, Prato, Italy, pp. 1–18.
- Faraj, S., & Johnson, S. L. (2011). Network exchange patterns in online communities. *Organization Science*, 22, 1464–1480.
- Garlaschelli, D., & Loffredo, M. I. (2004). Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26), 268701.
- Goedertier, S., De Weerd, J., Martens, D., Vanthienen, J., & Baesens, B. (2011). Process discovery in event logs: an application in the telecom industry. *Applied Soft Computing*, 11(2), 1697–1710.
- Gómez, V., Kaltenbrunner, A., & López, V. (2008). Statistical analysis of the social network and discussion threads in slashdot. Proceedings of the 17th International Conference on World Wide Web (WWW), Beijing, China, pp. 645–654.
- Hertog, P. D. (2000). Knowledge-intensive business services as co-producers of innovation. *International Journal of Innovation Management*, 04(04), 491–528.
- Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., & Tropsha, A. (2004). Mining protein family specific residue packing patterns from protein structure graphs. *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology—RECOMB '04* (pp. 308–315). San Diego: ACM Press.
- Krishna, V., Ranga Suri, N. N. R., & Athithan, G. (2011). A comparative survey of algorithms for frequent subgraph discovery. *Current Science*, 100(2), 190–198.
- Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. In P. S. Yu, J. Han, & C. Faloutsos (Eds.), *Link mining: Models, algorithms, and applications* (pp. 337–357). New York: Springer.
- Lahiri, M., & Berger-Wolf, T. Y. (2008). Mining periodic behavior in dynamic social networks. Proceedings of Eighth IEEE International Conference on Data Mining—ICDM '08, Pisa, Italy, pp. 373–382.
- Laniado, D., Tasso, R., Volkovich, Y., & Kaltenbrunner, A. (2011). When Thewikipedians talk: Network and tree structure of wikipedia discussion pages. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11), Barcelona, Spain.
- Lee, M. K. O., Cheung, C. M. K., Lim, K. H., & Sia, C. L. (2006). Understanding customer knowledge sharing in web-based discussion boards: an exploratory study. *Internet Research*, 16, 289–303.
- Leshed, G. (2009). Silencing the Clatter: Removing Anonymity from a Corporate Online Community. In T. Davies & S. P. Gangadhara (Eds.), *Online deliberation: Design, research, and practice* (pp. 243–251). Stanford: CSLI Publications.
- Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement—IMC '07, ACM Press, New York, New York, USA, p. 29.
- Nolker, R. D., & Lina, Z. (2005). Social computing and weighting to identify member roles in online communities. Web intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on, pp. 87–93.
- Peddibhotla, N. B., & Subramani, M. R. (2007). Contributing to public document repositories: a critical mass theory perspective. *Organization Studies*, 28, 327–346.
- Preece, J. (2000). Online communities: Designing usability and supporting sociality.

- Setia, P., Rajagopalan, B., Sambamurthy, V., & Calantone, R. (2012). How peripheral developers contribute to open-source software development. *Information Systems Research*, 23(1), 144–163.
- Sharratt, M., & Usoro, A. (2003). Understanding knowledge-sharing in online communities of practice. *Electronic Journal on Knowledge Management*, 1(2), 187–196.
- von Hippel, E., & von Krogh, G. (2003). Open source software and the “Private-collective” innovation model: issues for organization science. *Organization Science*, 14, 209–223.
- Wang, G., Liu, X., & Fan, W. (2011). A knowledge adoption model based framework for finding helpful user-generated contents in online communities. Proceedings of 30th Second International Conference on Information Systems, Shanghai, China.
- Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). Expertrank: a topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, 54(3), 1442–1451.
- Wasko, M. M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly*, 29(1), 35–57.
- Wassermann, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘Small-world’ networks. *Nature*, 393, 440–442.
- Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business Press.
- Yan, X., & Han, J. (2002). Gspan: Graph-based substructure pattern mining. Proceedings of 2002 I.E. International Conference on Data Mining (ICDM), Maebashi City, Japan, pp. 721–724.
- Yu, M. Y., Lang, K. R., & Kumar, N. (2010). Supporting better communication in academic communities of practice: an empirical study of Ais/Isworld. *Communications of the Association for Information Systems*, 26, 305–328.
- Zhang, W., & Watts, S. (2003). Knowledge adoption in online communities of practice. *Systemes d’Information et Management*, 9, 81–102.
- G. Alan Wang** is an Associate Professor of Business Information Technology. He received his Ph.D. in Management Information Systems from the University of Arizona, Tucson, AZ, in 2006, an M.S. in Industrial Engineering from Louisiana State University, Baton Rouge, LA, in 2001, and a B.E. in Industrial Management & Engineering from Tianjin University, P.R. China, in 1995. His research interests, such as data and text mining, heterogeneous data management, information retrieval, knowledge management and discovery, quality management, and service management, cover a wide spectrum of application areas in social media analytics, business intelligence, managerial decision support, engineering, and security informatics. His work has appeared in prestigious journals, such as *Production and Operations Management*, *Journal of Business Ethics*, *Decision Support Systems*, *Communications of the ACM*, *IEEE Transactions of Systems, Man and Cybernetics (Part A)*, and *IEEE Computer*.
- Xiaomo Liu** is a Research Scientist at Thomson Reuters R&D. He received a Ph.D. in Computer Science from Virginia Tech in 2013, an M.S. in Complex Adaptive Systems from Chalmers University of Technology in 2005, and a B.E. in Computer Science and Technology from Nanjing University of Science & Technology in 2003. His research interests include data mining, information retrieval and social computing. He has published more than 10 refereed journal and conference papers. His research appeared in journals such as *Information Systems Frontiers*, and *IEEE Computer*, and conferences proceedings such as *International Conference on Information Systems*, *Frontiers in Education Conference*, and *China Summer Workshop on Information Management*.
- Jianling Wang** is an Associate Professor of Accounting and Finance in the School of Management, Xi’an Jiaotong University. She received a Ph.D. in Management in 2005 and a M.S. in Economics in 2000 from Xi’an Jiaotong University. She received a B.S. in Economics from Shanxi Finance Institute in 1994. Her research interests include corporate finance and accounting, corporate managerial issues, and knowledge management.
- Min Zhang** is a Professor in the College of Management and Economics, Tianjin University. She received her Ph.D. in Management Science and Engineering from Tianjin University, China in 2006. Her research interests focus on statistical quality control, Six Sigma management, process monitoring applications, and service quality management. Her research has appeared in journals such as *Journal of Business Ethics*, *International Journal of Production Economics*, *International Journal of Production Research*, *Managing Service Quality*, *Total Quality Management & Business Excellence*, and *Quality & Reliability Engineering International*.
- Dr. Weiguo Fan** is a Full Professor of Accounting and Information Systems and Full Professor of Computer Science (courtesy) at Virginia Tech. He received his Ph.D. in Business Administration from the Ross School of Business, University of Michigan, Ann Arbor, in 2002, a M. Sc in Computer Science from the National University of Singapore in 1997, and a B. E. in Information and Control Engineering from the Xi’an Jiaotong University, P.R. China, in 1995. His research interests focus on the design and development of novel information technologies—information retrieval, data mining, text/web mining, business intelligence techniques—to support better business information management and decision making. He has published more than 140 refereed journal and conference papers. His research has appeared in journals such as *Information Systems Research*, *Journal of Management Information Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *Information Systems*, *Communications of the ACM*, *Journal of the American Society on Information Science and Technology*, *Information Processing and Management*, *Decision Support Systems*, *ACM Transactions on Internet Technology*, *Pattern Recognition*, *IEEE Intelligent Systems*, *Pattern Recognition Letters*, *International Journal of e-Collaboration*, and *International Journal of Electronic Business*. His research has been cited more than 3600 times (h-index:33, i10-index:72) Google Scholar. His research has been funded by five NSF grants, one PWC grant, and one KPMG grant.