An Analytical Framework for Understanding Knowledge-Sharing Processes in Online Q&A Communities

G. ALAN WANG, Virginia Tech
HARRY JIANNAN WANG, University of Delaware
JIEXUN LI, Oregon State University
ALAN S. ABRAHAMS and WEIGUO FAN, Virginia Tech

Online communities have become popular knowledge sources for both individuals and organizations. Computer-mediated communication research shows that communication patterns play an important role in the collaborative efforts of online knowledge-sharing activities. Existing research is mainly focused on either user egocentric positions in communication networks or communication patterns at the community level. Very few studies examine thread-level communication and process patterns and their impacts on the effectiveness of knowledge sharing. In this study, we fill this research gap by proposing an innovative analytical framework for understanding thread-level knowledge sharing in online Q&A communities based on dialogue act theory, network analysis, and process mining. More specifically, we assign a dialogue act tag for each post in a discussion thread to capture its conversation purpose and then apply graph and process mining algorithms to examine knowledge-sharing processes. Our results, which are based on a real support forum dataset, show that the proposed analytical framework is effective in identifying important communication, conversation, and process patterns that lead to helpful knowledge sharing in online Q&A communities.

Categories and Subject Descriptors: H.4.3 [**Information Systems Applications**]: Communications Applications; H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.1.2 [**Models and Principles**]: User/Machine Systems; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*

General Terms: Design, Languages

Additional Key Words and Phrases: Computer-mediated communication, online community, communication network, process mining, dialogue act, knowledge sharing

ACM Reference Format:
G. Alan Wang, Harry Jiannan Wang, Jiexun Li, Alan S. Abrahams, and Weiguo Fan. 2014. An analytical framework for understanding knowledge-sharing processes in online Q&A communities. ACM Trans. Manag. Inform. Syst. 5, 4, Article 18 (December 2014), 31 pages.
DOI: http://dx.doi.org/10.1145/2629445

1. INTRODUCTION

Communication plays an important role in knowledge sharing in online communities. According to knowledge management literature, people acquire knowledge through

This research was partially supported by the National Science Foundation under grant TUES-1122609 and a JPMorgan Chase Fellowship from the Institute for Financial Services Analytics at the University of Delaware.

Authors' addresses: G. A. Wang and A. S. Abrahams, Business Information Technology Department, Virginia Tech; email: alanwang@vt.edu; W. Fan, Accounting and Information Systems Department, Virginia Tech; email: wfan@vt.edu; H. J. Wang (corresponding author), Accounting and MIS Department, University of Delaware; email: hjwang@udel.edu; J. Li, College of Business, Oregon State University; email: jiexun.li@oregonstate.edu.

DOI: http://dx.doi.org/10.1145/2629445

reading and listening to others [McInerney 2002]. In addition, communication is used not only to exchange facts but also to transform and reshape them and engage in new trains of thought [Sharratt and Usoro 2003]. Computer-mediated communication research has long been studying communication patterns in online communities to understand the interpersonal interactions with implications for collaborative efforts [Walther 1996]. Interesting network structural characteristics including power law and scale-free are frequently observed for online social networks, such as Flickr, YouTube, LiveJournal [Cheng et al. 2008; Kumar et al. 2010; Mislove et al. 2007], and online blogs [Chau and Xu 2012]. These studies are done either at the individual level by studying users' egocentric network position and related measures such as closeness, centrality, and structural holes [Wassermann and Faust 1994], or at the community level by examining commonly occurred communication patterns such as reciprocity (direct or indirect) [Burke 1997; Flynn 2005; Molm et al. 2007; Putnam 2001] and preferential attachment [Faraj and Johnson 2011; Kadushin 2005; Newman 2003]. Very few studies examined thread-level communication patterns and their impacts on the effectiveness of knowledge-sharing activities. In addition, previous work [Wang et al. 2011] did not provide deeper consideration of the actual conversation process in each discussion. Conversation is important because it considers not only communication patterns but also communication contents.

In this study, we aim to examine thread-level knowledge sharing discussions in online Q&A communities from communication and process perspectives. An online Q&A community is a Web 2.0–based platform for users to exchange their knowledge in the form of asking and answering questions [Jin et al. 2012]. It consists of a large, loosely knit collection of geographically distributed individuals engaged in a shared practice through computer-mediated communication [Wasko and Faraj 2005]. Membership in an online Q&A community is often open to all users. In the community, a knowledge seeker can ask a specific question by starting a new discussion thread that may be responded to by other users as well as the seeker himself. A knowledge provider can share his or her knowledge in the form of a narrative describing a suggestion or a solution. Typically, question-answering threads in an online community begin with a question, posted by someone who needs help, that is followed by one or more postings made by those who try to provide or facilitate answers. A discussion may lead to one of the three possible outcomes: a satisfactory solution is obtained, helpful information is obtained but not an actual solution, or the discussion becomes futile or inconclusive.

Online Q&A communities are different from online technical support services provided by companies. They are interest-oriented communities [Spaulding 2010] in which community members spontaneously initiate and participate knowledge-sharing activities within a common knowledge domain such as software development (social.msdn.microsoft.com) or product knowledge (discussions.apple.com) [Sutanto et al. 2011]. Some online Q&A communities may be hosted or maintained by companies. However, host companies usually only maintain the normal operations of online communities without changing their crowd-intelligence nature. The whole idea of sponsoring an online Q&A community is to relieve the business from the burden of more traditional support mechanisms (e.g., call centers, live chats, customer service responses to user complaints) [Spaulding 2010] and more importantly the huge cost associated with it. Companies such as Dell have learned through years of experience that consumers are often more than willing to help each other in online Q&A forums [Bernoff and Li 2008]. These customer self-service communities are distinctly different from online technical support services such as Best Buy's Twelpforce on Twitter, where Best Buy employees are committed to answering customers' questions.

Online Q&A communities have become a popular and effective platform for knowledge seeking and sharing in recent years. Several studies have shown that organizations increasingly rely on external knowledge sources such as online Q&A communities because of increasing knowledge demands and limited availability of expertise and knowledge within their internal knowledge repositories [Constant et al. 1996; Wasko and Faraj 2005; Zhang and Watts 2003]. In addition to organizations and companies, individuals also seek knowledge related to product usage and problem solving in online communities [Lee et al. 2006]. A successful online Q&A community can accumulate a large knowledge repository and a large user base over time. Thus, it is important to adequately understand how to leverage and share knowledge in online Q&A communities to maximize the value of their resources.

The contribution of this research is twofold. First, we propose a new analytical framework to analyze and identify thread-level communication, conversation, and process patterns associated with effective knowledge sharing in online discussions. We explore how different discussion participants interact with one another in each individual discussion thread by building communication networks based on the "reply-to" relationships. Our goal is to identify those communication patterns that are best for effective knowledge sharing. To better understand the conversation process of knowledge sharing, we use a set of dialogue act tags [Kim et al. 2010] to capture the conversation purpose of each post in a discussion. We build a conversation network by connecting the dialogue acts based on the reply-to relationship. Standard communication network analysis techniques can be applied to conversation networks to reveal conversation patterns that may lead to effective knowledge sharing. By converting timestamped posts into event logs, we can also utilize existing business process modeling and mining techniques defined in process mining literature [Van der Aalst 1998, 2012; Van der Aalst et al. 2007] to discover the common conversation process patterns in effective knowledge-sharing discussions. Our second contribution is to examine online community discussions from three unique perspectives, such as communication networks, conversation networks, and process models, and discover patterns that lead to helpful knowledge sharing in online Q&A communities. The novel analytical approach of this study provides a potentially powerful way for future research to uncover interaction patterns embedded in online community data.

The rest of the article is organized as follows. Section 2 reviews related work on the network and process perspectives of online communities. Section 3 describes our research framework for understanding knowledge-sharing processes in online communities. Section 4 reports research findings and discuss research implications. Section 5 provides a conclusion and discusses limitations and future directions.

2. LITERATURE REVIEW

In this section, we provide the theoretical background of our research and a concise review of relevant research.

2.1. Theoretical Background

Online Q&A discussions are computer-mediated polylogues [Marcoccia 2004] where public users are engaged in online asynchronous discussions for knowledge construction. Social constructivist theory explains that the construction of knowledge is collaborative in nature and often triggered by a dialogue [Pena-Shaff and Nicholls 2004; Schrire 2004]. However, simply putting users in asynchronous discussions does not necessarily lead to effective knowledge co-creation [Weinberger et al. 2005]. Communication and cognitive theories show that there is a correspondence between the process structure of a discourse and the cognitive representation generated from the discourse [Pfister and Mühlpfordt 2002]. Clark and Brennan [1991] argue that speakers need to have a sense of grounding—that is, making sure that what they say is understood—to construct a coherent sequence of utterances and meaningful communication. The grounding methods mainly applied to face-to-face discussions include verbal feedback and nonverbal cues such as eye contact. However, existing computer-mediated communication tools are deficient in supporting grounding in asynchronous online knowledge-building discourses [Suthers 2001]. In the context of online learning, Suthers et al. [2001] found that collaborative knowledge construction is more effectively supported by knowledge representational tools such as knowledge maps. Those knowledge representational tools may not be applicable to online Q&A discussions where it would burden the users to use those tools.

A number of studies have particularly examined the role of conversation structures and participation patterns in the knowledge construction performance of computer-mediated discussions centered on learning. For example, Jeong [2003] found that interactions involving conflicting viewpoints promoted more discussion and critical thinking, and that the evaluation of arguments was more likely to occur as conclusions were being drawn. Hew and Cheung [2010] found that the number of participants in asynchronous online discussions was positively related to higher-level knowledge construction. However, Schellens and Valcke [2006] suggested that smaller groups were more likely to reach higher phases of knowledge construction because a large number of participants might suffer from cognitive overload. In addition, they examined three types of cognitive activities in task-oriented discussions, such as presentation of new information, explicitation (i.e., refining or elaboration of earlier ideas), and evaluation (i.e., confirmation or negation of earlier ideas). High proportions in communication about presentation and evaluation activities were observed in higher phases of knowledge construction, whereas explicitation activities were unexpectedly low. Those studies revealed practical implications for effective design of online discussions to better support knowledge sharing and construction.

Our study is motivated by those theories and empirical evidence in the cognitive and learning fields. Our goal is to identify conversation structures and communication patterns in online Q&A communities that can improve the efficacy of collaborative knowledge-sharing activities. One unique contribution of our research is to study the effectiveness of the cognitive process of online knowledge-sharing discussions using communication network analysis and process modeling techniques. To understand the cognitive meaning of each message in a discussion, we propose to use dialogue act in understanding the meaning of online Q&A discussion message so that conversation networks and process models can be constructed to represent the conversation structure and patterns of online discussions. We review dialogue acts, communication network, and process modeling in the next three sections.

2.2. Dialogue Acts: Understanding the Meaning of Online Q&A Discourse

In linguistic analysis, speech acts are often used to capture utterances or meaningful acts in human linguistic communication. Searle [1969] considers the illocutionary act as the basic unit of linguistic communication and defines five illocutionary acts such as representatives, directives, commissives, expressives, and declarations in his seminal work. Habermas [1979] claims that each speech act has consequences for the participants, leading to other immediate actions and commitments for future action. Winograd and Flores [1986] state that conversations contain a recurrent structure and can be formalized into a network of speech acts. The structure of interlinked speech acts in a conversation is described as a conversation network [Maier et al. 2009]. Although the speech act theory is ideal for us to understand the behavioral implications of each message in an online discourse, empirical evidence shows that Searle's speech act categorization does not work well when it is used to tag postings in online discussions. For example, Qadir and Riloff [2011] applied Searle's speech act categorization to an online message board text genre and found that declaration acts did not exist in message board forums. In addition, their effort to automatically classifying sentences to

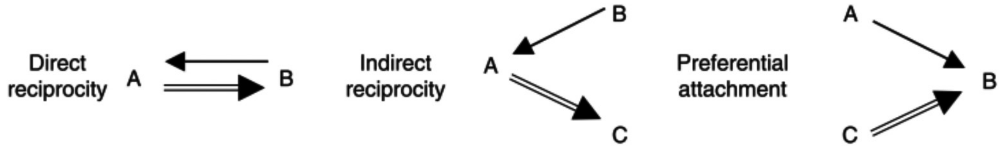


Fig. 1. Three common network exchange patterns (as in Faraj and Johnson [2011]). The presence of a relationship represented by a solid arrow increases the likelihood of a relationship represented by a double-lined arrow.

the five speech act categories found that some speech act categories such as assertives and commissives were difficult to recognize.

Similar to speech act, dialogue act provides an empirically based approach to the computational modeling of communication [ISO 2010]. Different dialogue act coding schemes have been proposed and customized to different application domains such as speech dialogue [Shriberg et al. 2004], task-focused email [Lampert et al. 2008], and instant messaging [Ivanovic 2008]. In an effort to understand the user behavior in newsgroups, Xi et al. [2004] closely examined a set of newsgroup messages and proposed five most prevalent message types, including question, answer, agreement or amendment, disagreement or argument, and courtesy. In an effort to analyze posts in troubleshooting-oriented Web forums, Kim et al. [2010] extended the five-way message classification into a set of 12 dialogue act tags. Online Q&A communities are similar troubleshooting-oriented Web forums. Therefore, we choose Kim et al.'s dialogue act tags to capture the illocutionary acts in online Q&A discussions.

2.3. Understanding Conversation Structures from a Network Perspective

Communication networks have been used to study the social exchanges among community participants. An online community is often viewed as a social network where the users who have a similar interest share information or knowledge within a knowledge domain. In this network, members of the community engage in continuous social exchanges online with diverse motivations. Some participate out of pure intrinsic motivation, whereas others participate more out of extrinsic factors such as recognition within the communities, rewards, or increased reputations, or a feeling of obligation to share because the member has received help from the community in prior exchanges [Constant et al. 1996; Von Hippel and von Krogh 2003; Peddibhotla and Subramani 2007; Preece 2000; Wasko and Faraj 2005].

Studying online communities from the communication network perspective in a natural setting has not happened until recently [Faraj and Johnson 2011]. Their perspective on online communities builds 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-hows, 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. Using the hybrid lens of both the social exchange theory [Cook and Rice 2001] and network exchange theory [Burke 1997] allows researchers to examine online communities systematically at the macro network level without being lost in the micro individual level. In their study, online community participation is viewed as a social phenomenon. As in any endeavor governed by human behavioral patterns, they argue that participants in online communities will exhibit nonrandom, intentional communication choices.

Faraj and Johnson [2011] propose three network exchange patterns (Figure 1) that they claim to be visible in all types of online communities: direct reciprocity (user A will reply to user B as a direct reciprocity for user B's prior help to user A), indirect reciprocity (after user B helps user A, user A will help user C as a general, indirect reciprocity), and preferential attachment (new actors choose to interact with already well-connected others; e.g., both A and C choose to interact with the prominent user B). Using five online discussion forums, they find consistent results across those communities during 27 months of data observation encompassing 38,483 interactions between 7,329 individuals. Their results show that the pattern of ties is consistent with the norms of direct reciprocity and indirect reciprocity and has a tendency away from preferential attachment.

Instead of focusing on social exchange patterns at the community level, we apply the communication network to the understanding of the conversation structures of individual discussion threads. After each message is tagged with a dialogue act, we can build a conversation network for each discussion thread with nodes being the dialogue act tags and links being the reply relationships. We can then apply network structural analysis methods [Wassermann and Faust 1994] to reveal characteristics of conversation patterns that emerge in online discussions. There are very few studies that focus on the structures of interactions of individual discussion threads. One exception is the work of Gómez et al. [2008], who visualized thread discussions using an intuitive radial tree representation and observed network structural properties such as self-similarity. Adamic et al. [2008] investigated the structural characteristics of networks, each of which was constructed based on a group of online discussions. Their intention was to find the characteristics of those users who were more likely to provide answers to questions. Although their research findings can be useful in terms of identifying good indicators of experts in social networks, there is little insight on the communication and conversation processes among users.

2.4. Process Modeling, Analysis, and Mining

The communication interactions of online discussions can also be understood from a process perspective. Many notations have been proposed to model business processes, such as UML activity diagrams, Business Process Modeling Notation (BPMN), event-driven process chains (EPCs), and Petri nets. Van der Aalst [1998] applied Petri nets to model and verify workflows. A number of Petri net extensions have been proposed, including colored Petri nets, timed Petri nets, and hierarchical Petri nets, to model workflow attributes, temporal behavior of workflow, workflow events, and subworkflows [Ha and Suh 2008]. Given that Petri nets have formal semantics and have been extensively used in workflow verification and process mining, we leverage Petri nets to formally define dialogue act–based process models in online communities. Process mining techniques can be roughly classified into three categories: process discovery, process conformance, and process performance analysis [Van der Aalst 2012]. Traditional process analyses are carried out using established methods such as “walk-throughs,” interviews, and workshops, which are extremely labor intensive and time consuming. Process mining techniques enable the automatic discovery of process models from historic data in various IT systems. Many algorithms have been developed for process discovery, such as the alpha algorithm, heuristic mining algorithm, multiphase mining algorithm, and fuzzy mining algorithm [Van der Aalst and Weijters 2004]. Besides discovering process models, useful information on the organizational aspect of business processes can also be discovered, such as the social network of participants in the processes, interaction patterns, and network of work transfers. When there is a predefined process model, process mining techniques can automatically check the conformance between the actual events and the existing model. For example, process mining techniques can recognize established paths that have not been followed and pinpoint the exceptional process instances. Root causes of those identified deviations can be easily traced down. Process mining can also provide insights on process performance by conducting process simulations. Typical process performance analyses include cycle time

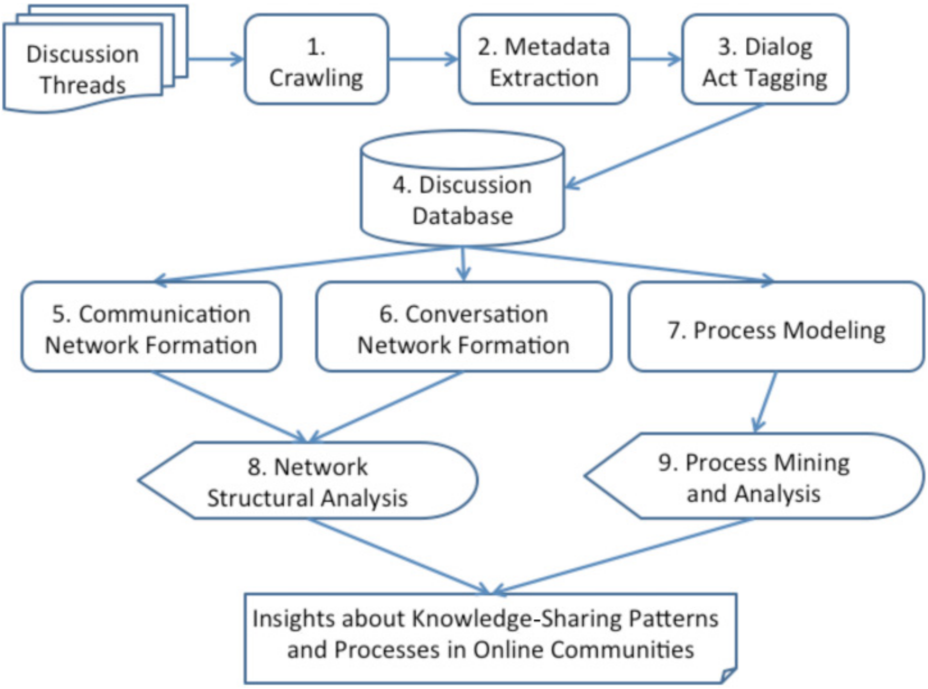

Discussion Threads
1. Crawling
2. Metadata Extraction
3. Dialog Act Tagging
4. Discussion Database
5. Communication Network Formation
6. Conversation Network Formation
7. Process Modeling
8. Network Structural Analysis
9. Process Mining and Analysis
Insights about Knowledge-Sharing Patterns and Processes in Online Communities

Fig. 2. An analytical framework for understanding knowledge-sharing patterns and processes in online communities based on dialogue acts.

analysis, bottleneck analysis, process cost analysis, and so forth, which enable process redesign and continuous process improvements. With widely used network technologies, virtual processes such as distance learning and online knowledge sharing have become an integral part of business operations and increasingly important in process mining. For example, Reimann et al. [2008] used process mining to identify models of group decision making based on online chat records. Gómez et al. [2008] proposed using a radial tree to represent the graphical structures of online discussion processes. Compared to existing studies, our work is unique in mining interaction patterns among different dialogue act tags that lead to effective knowledge sharing.

3. AN ANALYTICAL FRAMEWORK FOR UNDERSTANDING KNOWLEDGE-SHARING PROCESSES IN ONLINE COMMUNITIES

We propose an analytical framework for understanding the knowledge-sharing processes in online communities from three theoretical lens: communication network formation, conversation network formation, and process modeling (Figure 2). Steps 1 through 4 are the data capture stage, which gathers the data needed for modeling. Steps 5 through 9 are the data analysis and modeling stage, which seeks to uncover the underlying patterns for effective knowledge sharing in online knowledge communities. Using the analytical approaches proposed in this framework, we seek to identify important communication, conversation, and process patterns that lead to helpful knowledge sharing in online Q&A communities. In the rest of this section, we describe each component in this research framework in more detail.

3.1. Data Extraction

Online communities for knowledge sharing are often designed in the format of a discussion forum. According to Big Boards (http://1topix.com/statistics/?stat=engine), most

Table I. An Example of Metadata Extracted from Online Discussion Threads

Thread ID	Helpfulness Feedback	Post ID	Post Replied To	Participant	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

discussion forums use one of the 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 timestamp, a post ID, a thread ID, the 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 feedback: *solved* (a satisfactory solution is obtained), *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 pages. An example of online discussion metadata is provided in Table I.

To understand the conversation utterances in online discussions, we can assign a speech/dialogue act tag to each discussion post based on its conversation purposes, such as raising a question, providing an answer, and confirming an answer, in the context of problem solving. We choose not to use Searle's speech act classification because empirical evidence shows that the categorization does not work well when tagging postings in online message board texts [Qadir and Riloff 2011]. The nuanced manner in which humans express speech acts presents substantial challenge to computation mechanisms, and state-of-the-art computational techniques are unable to achieve high F1 scores (high precision and high recall, simultaneously) on a broad scale for multiple speech act types on diverse corpora [Carvalho and Cohen 2006]. The difficulty of automated speech act tagging is compounded when assessing threads that have complex graph relationships between speakers instead of single emails [Cohen et al. 2004]. Even for email threads, which possess lower message linkage than discussion threads, automated speech act tagging achieves satisfactory performance (F1 > 0.80) for only one of many speech act types (viz. "commissive" speech acts) and in only a restricted domain [Carvalho and Cohen 2005]. Automated speech act tagging for a wide assortment of speech acts on a general corpus with high message linkage therefore cannot currently be achieved with accuracy (high precision), sensitivity (high recall), and reliability (high agreement). We chose to use the dialogue act tags developed by Kim et al. [2010] because the tags were specifically developed to analyze Web forums similar to online Q&A communities and could be applied with satisfactory precision, recall, and agreement. We modified the definitions of two dialogue act tags, namely QUESTION-CONFIRMATION and ANSWER-CONFIRMATION, because they represent more than one meaning and can be confusing for tagging purposes. Table II summarizes the 11 tags that we adopt in this study and their definitions. It is worth noting that dialogue act tagging can be automated using statistical learning methods. However, the accuracy of the automated tagging approaches is not satisfactory [Kim et al. 2010]. Furthermore, our analysis of users' online communication patterns requires accurate tagging of posts. Therefore, in this study, we hire human annotators to manually read sample posts and assign dialogue act tags to them.

Table II. Dialogue Acts in Online Q&A Discussions

Short Tag	Full Tag	Explanation
Q-Q	QUESTION-QUESTION	The post raises a new question and expects an answer or explanation. In general, QUESTION-QUESTION is reserved for the first post in a thread.
Q-ADD	QUESTION-ADD	The post supplements the original question by providing additional information or by asking follow-up questions.
Q-CORR	QUESTION-CORRECTION	The post corrects errors in a question.
Q-CONF	QUESTION-CONFIRMATION	The post confirms the same question and is done by someone other than the thread initiator.
A-A	ANSWER-ANSWER	The post proposes an answer to a question.
A-ADD	ANSWER-ADD	The post supplements an existing answer by providing additional information or asking a follow-up question to the answerer.
A-CORR	ANSWER-CORRECTION	The post corrects errors in an answer.
A-CONF	ANSWER-CONFIRMATION	The post confirms an answer on experimental or theoretical grounds. This is done by someone other than the thread initiator.
A-OBJ	ANSWER-OBJECTION	The post objects to an answer on experimental or theoretical grounds.
RES	RESOLUTION	The person who initiated the question confirms that an answer works on the basis of implementing it.
OTH	OTHER	The post does not belong to any of the preceding dialogue act tags.

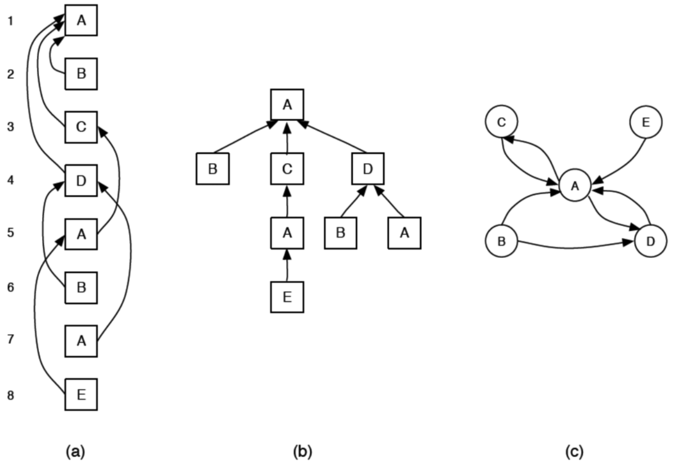


Fig. 3. Posts of a thread in a chronological order (a), a post tree (b), and a communication network (c). Squares denote posts in a thread. Circles denote post authors (A, B, . . . E). An arrow line represents the reply direction between two posts or users.

3.2. Communication Network Formation

Based on the metadata extracted from the previous section, we can build a communication network for each discussion thread based on its posts, authors of posts, and reply relationships between the posts. In most online discussion forums, posts of a thread are organized chronologically, from oldest to latest, as shown in Figure 3(a). Based on the reply relationships, we can first build a post tree with a tree-like view of all the posts in a thread, as shown in Figure 3(b). Furthermore, based on the post authorship,

Table III. An Illustration of Posts of a Thread Labeled with Dialogue Act Tags

Thread ID	Post ID	Participant	Reply To	Dialogue Act Tag
1	1.1	A	0	Q-Q
1	1.2	B	1.1	A-A
1	1.3	C	1.2	Q-ADD
1	1.4	D	1.3	A-A
1	1.5	E	1.4	A-OBJ
1	1.6	F	1.1	Q-CONF
1	1.7	G	1.1	Q-CONF
1	1.8	H	1.2	A-CORR
1	1.9	F	1.7	Q-ADD

we can convert the post tree into a communication network, which is a user-directed graph showing communications among the thread participants. We now provide formal definitions for a post tree and a communication network based on the graph theory.

Definition 1 (*Post Tree*). A post tree is a directed graph $T = \langle P, E \rangle$ 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 usually contains a question. A tree edge set E consists of all reply relationships between posts in the thread. A directed edge $e = \langle u, v \rangle$ indicates that post u replies to post v.

Definition 2 (*Communication Network*). A communication network is a directed graph $= \langle U, E \rangle$. Each node $u_i \in U$ corresponds to a thread participant, whereas each edge in this graph $e_i = (u_i, u_j)$ represents communication between two thread participants.

3.3. Conversation Network Formation

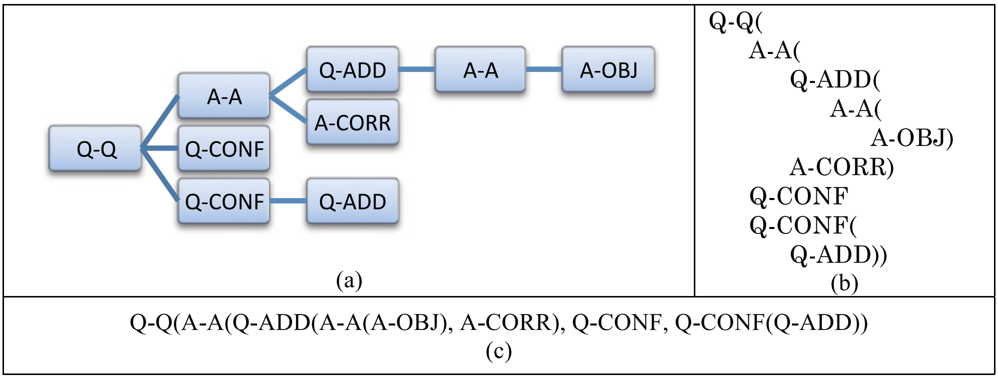
A communication network reveals the interactions between discussion participants; however, it only captures the structural patterns of communication and fails to understand the meanings embedded in the communication. Dialogue acts help us understand the conversation meanings. After assigning a dialogue act tag to each post in a discussion thread (described in Section 3.1), we can build a conversation network based on the "reply-to" relationships between posts. Unlike a communication network that focuses on participants, a conversation network consists of all posts in a thread and forms a tree. Such a conversation network is similar to the post tree defined in Section 3.2, except each node is represented by a dialogue act rather than a post ID. Here, we give a formal definition of a conversation network.

Definition 3 (*Conversation Network*). A conversation network is a directed graph $<A, E>$ with no cycles. A node set A contains all dialogue acts that appear in a discussion thread. The root node usually is a Q-Q, which denotes a new question in the starting post. A tree edge set E consists of all reply-to relationships between the dialogue acts in the thread.

For example, Table III is an example thread consisting of nine posts annotated by dialogue act tags. Based on the reply relationships, this thread can be presented in a conversation network (tree) as shown in Figure 4(a) through (c). Each figure is a different representation of the same conversation network.

3.4. Network Structural Analysis

Based on the extracted communication and conversation networks, we want to find out the structural patterns that are more likely to lead to effective discussions using



```
Q-Q(
    A-A(
        Q-ADD(
            A-A(
                A-OBJ)
        A-CORR)
    Q-CONF
    Q-CONF(
        Q-ADD))
```

(b)

Q-Q(A-A(Q-ADD(A-A(A-OBJ), A-CORR), Q-CONF, Q-CONF(Q-ADD))

(c)

Fig. 4. A conversation network in three different representations.

subgraph analysis. Subgraphs 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 substructures is crucial to understanding the interactions and dynamics at work in datasets. Subgraph mining algorithms such as gSpan [Yan and Han 2002] and FFSM [Huan et al. 2003] can be used to discover frequent graph-based patterns from a collection of communication and conversation networks. The family of subgraph mining algorithms is quite similar to the classic Apriori algorithm [Agrawal and Srikant 1994] on the transactional database. Generally, the problem of subgraph mining can be formalized as follows. Given a database with n graphs $G = \{G_1, \ldots, G_n\}$, $sup(g)$ is the number of graphs in G that contain the subgraph g:

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

Efficient algorithms can then be used to discover any possible g with the minimum support $sup(g) \geq \sigma$.

The skeleton of subgraph mining algorithms is presented in Algorithm 1. The most computationally intensive steps in the algorithm include the testing of subgraph isomorphism in the subroutine *subgraph-isomorphism*() and generating candidate subgraphs from size k to k+1 in the subroutine *candidate-generate*(). Existing algorithms invented various strategies to speed up these two steps. In this study, we choose one of the most commonly used algorithms, gSpan, to detect common patterns in communication and conversation networks.

3.5. Process Modeling, Analysis, and Mining

The conversation network in the previous section focuses on the interactions among different dialogue acts based on the reply-to relationship. Given that post timestamps can depict their temporal execution orders in a discussion thread, we can leverage process mining techniques to further investigate the knowledge-sharing processes. To facilitate the process structure analysis of different threads via process mining, we model the processes based on Petri nets given its formal mathematical foundation and process modeling semantics as defined in Van der Aalst [1998].

Definition 4 (*Petri net*). A Petri net *PN* is a triple (P, T, F):

—P is a finite set of places,
—T is a finite set of transitions ($P \cap T = \emptyset$), and
—$F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs.

ALGORITHM 1: Skeleton of subgraph mining algorithms

```
Input: a graph database $G = \{G_1, \ldots, G_n\}$, a minimum support $\sigma$
Output: $\{F_1, \ldots, F_k\}$ a set of frequent sub-graphs of cardinality 1 to $k$
$C_1 \leftarrow$ all sub-graphs of cardinality 1
$F_1 \leftarrow \{c \in C_1 | \mathbf{sup}(c) = c.count/n \geq \sigma\}$
for $(k = 2; F_{k-1} \neq \emptyset; k{+}{+})$ do
    $C_k \leftarrow$ candidate-generate $(F_{k-1})$
    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
    $F_k \leftarrow \{c \in C_k | \mathbf{sup}(c) = c.count/n \geq \sigma\}$
end
return $F \leftarrow \cup F_k$
```

A place p is called an *input place of a transition t* if and only if there exists a directed arc from p to t. Place p is called an *output place of a transition t* if and only if there exists a directed arc from t to p. $^{\bullet}t$ is used to denote the set of input places for a transition t; $t^{\bullet}$, $^{\bullet}p$, and $p^{\bullet}$ have similar meanings. To use Petri nets to model processes, additional requirements must be added, resulting in WorkFlow net (WF-net) definition [Van der Aalst 1998].

Definition 5 (*WF-net*). A Petri net $PN = (P, T, F)$ is a WF-net if and only if the following apply:

—PN has two special places: i and o. Place i is a source place: $^{\bullet}i = \emptyset$. Place o is a sink place: $o^{\bullet} = \emptyset$.
—If a transition t^* is added to PN to connect place o with i—that is, $^{\bullet}t^* = \{o\}$ and $t^{*\bullet} = \{i\}$—then the resulting Petri net is strongly connected.

This ensures that processes modeled as WF-nets must have one starting node, one ending node, and no dangling nodes. Various extensions to Petri nets have also been proposed to represent objects, such as process resources, time, and hierarchy [Van der Aalst 1998]. In process modeling, tasks are modeled using transitions. Routing constructs, such as fork, join, branch, and merge, are represented as transitions (fork and join) or places (branch and merge) [Van der Aalst 1998]. In this article, each thread (case) is an instance of the overall knowledge-sharing process, which includes a sequence of dialogue acts modeled as tasks. We can then define *dialogue act process* as follows.

Definition 6 (*Dialogue Act Process*). A dialogue act process $DAP = (P, T, F)$ is a WF-net, where T = {Q-Q, Q-ADD, Q-CORR, Q-CONF, A-A, A-ADD, A-CORR, A-CONF, A-OBJ, RES, OTH}.

Thus, a path pa from a node n_1 to a node n_k in a DAP $n_i \in P \cup T$ is a sequence $pa = <n_1, n_2, \ldots, n_k>$ such that $<n_i, n_{i+1}> \in F$ for $1 \leq i \leq k-1$. If a path has $n_1 = i$ (source place) and $n_k = o$ (sink place), it is then called a *full path*. Based on Definition 6, we also know that branch and merge are the only two routing constructs needed for DAPs given that parallelism and advanced workflow routing patterns [Van der Aalst et al. 2003] are not necessary for those processes made of dialogue act tags. For any DAP, there must be a place between two transitions according to the Petri net definition. To simplify the representation of dialogue act interactions from a temporal point of view, we formally define a *dialogue act process pattern* by removing all intermediate places from the full paths in a DAP as follows.

Definition 7 (*Dialogue Act Process Pattern*). A dialogue act process pattern (DAPP) is a sequence derived from a full path in DAP by removing all place nodes except for a source place i and a sink place o.

For example, <i, Q-Q, A-A, Q-ADD, o> is a DAPP, which represents a thread with three sequential posts tagged as Q-Q, A-A, and Q-ADD, respectively. To analyze DAPs from a process mining perspective, it is critical to know the five key inputs for process mining as follows [Van der Aalst and Weijters 2004]:

—*Process instance identifier/case identifier*: It is necessary to distinguish one process instance from another in business processes. The process instance identifier is domain specific. For example, the process instance identifier could be the patient ID in a hospital context. In our article, each DAP is treated as a process instance with a unique ID.

—*Task/activity/event ID*: A task/activity/event can be informally defined as a well-defined step or status change that is performed in a business process. Typically, an event is represented by a stative verb (e.g., “registered,” “completed”) or locative phrase (e.g., “at specialist,” “in progress”) for status changes, or a verb–noun phrase for some actions or tasks, such as “submit form” and “approve application.” As mentioned earlier, we have 11 task/events (dialogue acts).

—*Timestamp*: A timestamp is needed for each task to determine the overall execution order for all tasks. Ideally, each task should have a starting timestamp and an ending timestamp so that the duration of the task can be derived. In this study, because each post only has a starting timestamp, the duration of each post is assumed to be zero seconds. Given that the duration of the posts (i.e., the time the posters take to write the posts) is not the focus of this work, we believe that this assumption is reasonable.

—*Originator/resource*: The originators are the organization resources responsible for executing the tasks, such as registrar, administration coordinator, email system, and authentication system. The post authors are considered as the resources in our study.

—*Data items*: The data items are the data related to tasks, such as forms, documents, and specific data values. Data items are useful information for mining business rules and checking process policy conformance. In this article, we mainly focus on the structural analysis of knowledge-sharing processes and defer the analysis of process data items to our future research.

Here, we use Disco (http://www.fluxicon.com/disco/) for process mining and analysis given its ease of use, rich features, and strong academic background. Disco was developed by leading academics behind the well-known open source process mining tool ProM (http://www.processmining.org). The process discovery feature of Disco was developed based on the fuzzy mining algorithm [Günther and van der Aalst 2007], which is one of the most widely used process mining algorithms in practice. In addition, as we will discuss in Section 4.4, the fuzzy mining algorithm is designed to handle event logs with high variations, which is one of the key characteristics of our dataset. The focus of this work is not on process mining algorithms, and therefore we refer those readers interested in the various parameters of fuzzy miners to Günther and van der Aalst [2007] and leave the analysis of our datasets using different process mining algorithms to our future research. After loading the datasets into Disco, we can first analyze the process mining data by looking into various statistics, such as total number of events, events per case distributions, resources frequencies, and case variations, which provide insights into detailed analysis about knowledge-sharing process. Then, we will mine the process maps for different types of threads, namely, “Unhelpful,” “Helpful,” and “Solved.” By examining the various paths among different tasks (tags), we conduct an analysis of the most frequent partial path pairs $<n_i, n_{i+1}>$ from DAPPs found in the discovered process maps. We then tag those path pairs as “Good,” “Neutral,” or “Bad,” according to whether the path pair is desirable for problem-solving processes based on their semantics and historical data, which we refer to as path pair quality rating (PPQR). For example, <Q-ADD, A-ADD> means that more information is added about the original question and then more information is added for existing answers. Adding information to existing answers is always good for problem solving, and as we will discuss in Section 5, this pair appears way more in effective threads than in ineffective threads. We thus give <Q-ADD, A-ADD> a PPQR of “Good.” <Q-ADD, o> is another example: here, the meaning is that the last post of the thread is still adding information for the original question, which implies that the original question has not been answered. Furthermore, <Q-ADD, o> appears more frequently in ineffective threads, which further confirms that its PPQR should be “Bad.” When a specific pair is not tagged due to the lack of data or its low frequency, we treat that pair as “Neutral.” Based on the PPQR of all path pairs, we can define the dialog act–based “golden paths” to problem solving in online communities as follows.

Definition 8 (*Golden Path*). Given a dialogue act process pattern *DAPP*, each unique subsequence pair in *DAPP* is assigned a *path pair quality rating* (PPQR) of either “Good,” “Neutral,” or “Bad”—that is, $<n_i, n_j>_{\text{good}}$, $<n_i, n_j>_{\text{neutral}}$, or $<n_i, n_j>_{\text{bad}}$. A *DAPP* is called a *golden path* if and only if the *golden path quality* (GPQ) defined as (α $*$ Total number of $<n_i, n_j>_{\text{good}}$ + β $*$ Total number of $<n_i, n_j>_{\text{neutral}}$ + γ $*$ Total number of $<n_i, n_j>_{\text{bad}}$)/Total number of $<n_i, n_j>$ is equal to or greater than θ, where α, β, γ are the coefficients to represent impacts of different types of path pairs on GPQ and θ is a predefined threshold for GPQ.

The values of α, β, γ, and θ are determined based on historical data. Given the formal quantitative definitions of *partial path quality rating* and *golden path quality*, we can calculate the GPQ value for all threads in the online forum and design advanced algorithms to detect paths that are deviating from golden paths, create mechanisms to increase those paths’ GPQ, and eventually convert those paths into golden paths, as we will discuss in later sections.

4. EMPIRICAL EVALUATIONS AND RESULTS

In this section, we describe an empirical evaluation of our proposed analytical framework using real data downloaded from a popular online community—Apple Support Communities. The evaluation shows that our analytical framework helps us better understand the knowledge-sharing processes in online Q&A communities and reveal insights on the communication and conversation processes.

4.1. Data Collection

To limit the scope of our study, we focused on one of the most active subforums—the iPhone subforum. Specifically, we crawled 49,343 online discussion threads from the iPhone subforum of the Apple Support Communities posted between July 1, 2007, and April 1, 2010. Table IV provides some insights about this online Q&A community. This is an active community, where an average of about 2,200 users generated about 8,500 posts per month. There were more knowledge contributors who replied to at least one question than those seekers who asked at least one question. There were only 2,695 users, about 9.7% of all knowledge contributors, who made at least 10 replying posts between 2007 and 2010. Those contributors participated in knowledge sharing in 42,178 discussion threads (85%). However, we should not discredit the efforts of

Table IV. Analysis of the Apple Forum Data

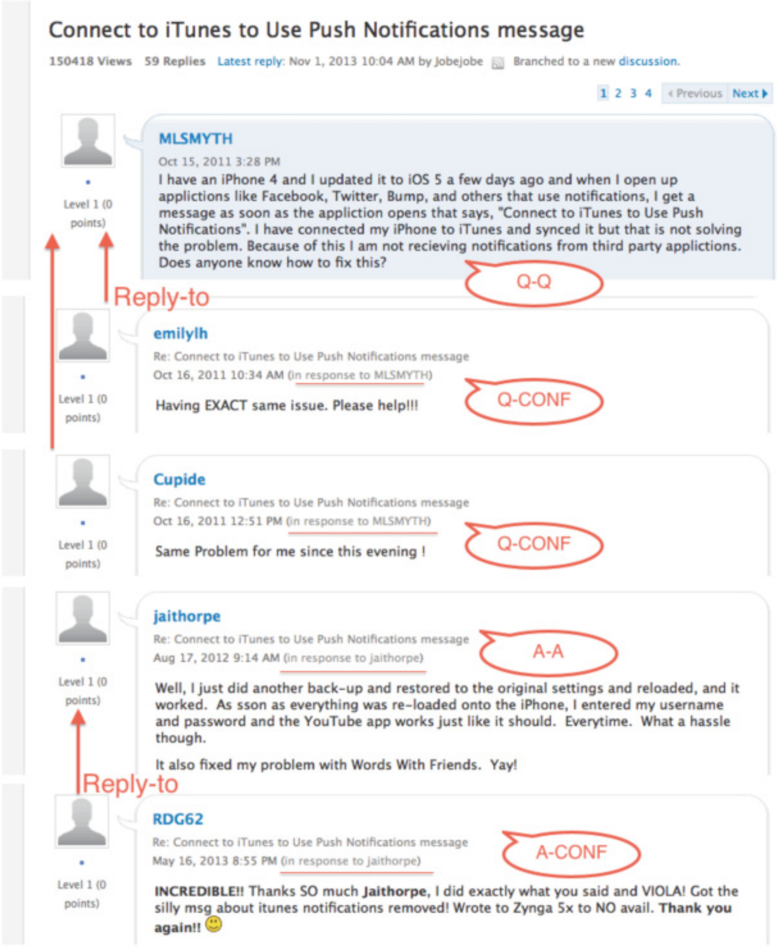
Community level:	
Registered users	55,108
Discussion threads	49,343
Posts	271,823
Average posts per month	8,494
Average unique users per month	2,217
Unique knowledge seekers	27,739
Unique knowledge contributors	35,681
Users both seek and contribute	16,643
Percentage of users who contributed 80% of replies	23.53%
Thread level:	
Average posts per thread	5.131
Average users per thread	4.154
Average words per thread	100.56

those one-time contributors, who replied to 16,270 discussion threads (33%). The Apple community is comparable to other large online Q&A communities, in that it is reported to have substantially more responses than questions and askers [Liu et al. 2008].

We removed 6,702 threads without any reply from the dataset because those threads do not have any communication. One unique feature in this online community is to let users provide feedback on user-generated contents. A knowledge seeker can rate an answer that he or she receives with either a “helpful” tag or a “solved” tag. There were 12,453 discussion threads with the solved tag and 1,140 threads with the helpful tag, whereas 29,048 threads did not have any user feedback. We randomly selected 40 solved discussion threads and 40 helpful threads. However, those threads without any user feedback do not necessarily mean that they contain no helpful knowledge or solutions. Past research has found that the user feedback participation is often very low in virtual communities [Cao et al. 2011]. It is likely that due to users’ negligence or unwillingness, some of those threads may contain helpful knowledge or even solutions. We decided to manually examine those threads without feedback and identify threads with no helpful knowledge. We randomly selected discussion threads without user feedback and asked two computer science graduate students to manually examine whether they contained solutions or helpful knowledge. We identified 1,017 threads of which both experts agreed that the threads contained no helpful information. We then randomly selected 40 “unhelpful” threads from them. In summary, our dataset contains 120 discussion threads including 40 solved discussion threads, 40 helpful threads, and 40 unhelpful threads. We then parsed each discussion thread to get its metadata as described in Section 3.1.

We did not select a larger sample because we needed to manually examine each discussion thread in the dataset and to decide on a dialogue act tag (defined in Table II) for each post. We tagged a total of 120 threads, each having an average of 5.32 posts. To ensure that our tagging was not biased, we asked two graduate students to independently tag 60 threads each. Before they started tagging, we provided them with a tagging tutorial explaining the definitions of the tags, differences between similar tags, and tagging examples. The kappa statistic was computed between the students’ tagging results and ours, with a value of 0.7425. According to existing literature [Landis and Koch 1977; Rietveld and Hout 1993], the kappa score indicates substantial agreement between the two sets of tagging results. Therefore, we did not pursue resolving disagreement between our tagging results and those of the students.

A thread example on the Apple support forum along with post tags is shown in Figure 5. In this example, user “MLSMYTH” asked a question (tagged with Q-Q). Two



Connect to iTunes to Use Push Notifications message

150418 Views 59 Replies Latest reply: Nov 1, 2013 10:04 AM by Jobejobe Branched to a new discussion.

1 2 3 4 ◂ Previous Next ▸

MLSMYTH

Level 1 (0 points)

Oct 15, 2011 3:28 PM

I have an iPhone 4 and I updated it to iOS 5 a few days ago and when I open up applictions like Facebook, Twitter, Bump, and others that use notifications, I get a message as soon as the appliction opens that says, "Connect to iTunes to Use Push Notifications". I have connected my iPhone to iTunes and synced it but that is not solving the problem. Because of this I am not recieving notifications from third party applictions. Does anyone know how to fix this?

emilylh

Level 1 (0 points)

Re: Connect to iTunes to Use Push Notifications message

Oct 16, 2011 10:34 AM (in response to MLSMYTH)

Having EXACT same issue. Please help!!!

Cupide

Level 1 (0 points)

Re: Connect to iTunes to Use Push Notifications message

Oct 16, 2011 12:51 PM (in response to MLSMYTH)

Same Problem for me since this evening !

jaithorpe

Level 1 (0 points)

Re: Connect to iTunes to Use Push Notifications message

Aug 17, 2012 9:14 AM (in response to jaithorpe)

Well, I just did another back-up and restored to the original settings and reloaded, and it worked. As sson as everything was re-loaded onto the iPhone, I entered my username and password and the YouTube app works just like it should. Everytime. What a hassle though.

It also fixed my problem with Words With Friends. Yay!

RDG62

Level 1 (0 points)

Re: Connect to iTunes to Use Push Notifications message

May 16, 2013 8:55 PM (in response to jaithorpe)

INCREDIBLE!! Thanks SO much **Jaithorpe**, I did exactly what you said and VIOLA! Got the silly msg about itunes notifications removed! Wrote to Zynga 5x to NO avail. **Thank you again!!** 🙂

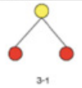
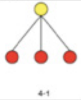
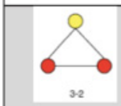
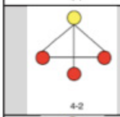
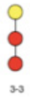
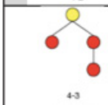
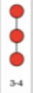
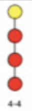
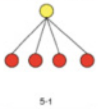
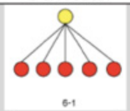
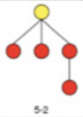
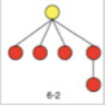
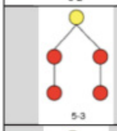
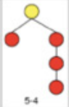
Fig. 5. An example thread on the Apple support forum.

other users (i.e., “emilylh” and “Cupide”) replied to user MLSMYTH and confirmed the same problem, which were tagged as question confirmation (Q-CONF). User “jaithorpe” provided an answer to the original question (tagged as A-A). User “RDG62” replied to user jaithorpe to confirm that his or her solution/answer works, which was tagged as answer confirmation (A-CONF).

4.2. Frequent Patterns in Communication Networks

Table V shows all frequent subgraphs discovered from the communication networks built using Apple’s online discussions with the minimum support value of 0.1, which is the default setting of the gSpan algorithm and is also used in other subgraph mining studies [Huan et al. 2003; Yan and Han 2002]. The support value of a subgraph is defined as “given a set of graphs, the percentage of graphs that contain the substructure.” Subgraphs with support values of at least 0.1 are frequent patterns that may provide insights into conversation patterns among online users. The discovered n-node subgraphs contain local interaction patterns between one knowledge seeker and up to

Table V. Frequent Subgraphs in Apple Communication Networks

3-Node Pattern	Support Value			4-Node Pattern	Support Value		
	Unhelpful	Helpful	Solved		Unhelpful	Helpful	Solved
	69%	67%	66%		22%	36%	33%
	5%	19%	21%		3%	15%	13%
	33%	32%	36%		20%	26%	29%
	5%	16%	15%		5%	15%	15%
5-Node Pattern				6-Node Pattern			
	11%	18%	18%		2%	10%	12%
	12%	20%	20%		4%	13%	10%
	2%	8%	11%				
	3%	14%	12%				

five knowledge sharers. They include a variety of interaction structures and uncover how helpful knowledge is produced during the knowledge-sharing process.

The support values in Table V show the percentages of each pattern found in the three types of discussion threads. We first compared the frequencies of patterns in unhelpful threads with those in helpful and solved threads. For three-node patterns, subgraphs 3-1 and 3-3 represent two basic formats of interactions: (1) two repliers discuss with the original poster individually or (2) two repliers have a sequential dialogue with the original poster. Both patterns appeared frequently in all three types of online Q&A discussions. Furthermore, their frequencies are almost the same regardless of the helpfulness level. Subgraphs 3-2 and 3-4 are the special patterns that are only frequent in helpful and solution threads. Their frequency differences compared to unhelpful threads were more than 10%. These two patterns exhibit the interactions between knowledge sharers only. They indicate that encouraging discussions among replying participants can be beneficial for effective knowledge sharing. For subgraphs with four nodes or more, they all appeared more frequently in helpful and solved threads than in

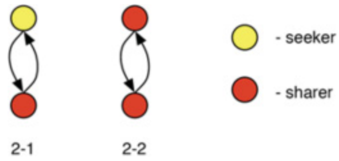


Fig. 6. Reciprocal patterns in communication networks.

Table VI. Frequency of Reciprocal Communication Patterns

Pattern	Unhelpful	Helpful	Solved
2–1	33%	67%	58%
2–2	4%	10%	10%

unhelpful threads. The frequency differences were more than 5%, some were even more than 10%. These results indicate that unhelpful threads had very different structures when compared with helpful and solved threads. There were also certain frequency differences between patterns in helpful and solved threads. However, the differences are not statistically significant. Thus, these communication patterns are not sufficient to be used to determine if a discussion thread yields a solution.

The gSpan algorithm has one limitation in that it can only search subgraphs in undirected graphs. Thus, we used a simple heuristic to extract reciprocal patterns as shown in Figure 6. Pattern 2-1 denotes the reciprocal interactions between a knowledge seeker and a knowledge sharer, whereas pattern 2-2 represents interactions between two knowledge sharers. Frequencies of these two patterns in our dataset are reported in Table VI. We observed that both reciprocal patterns (especially 2-2 for seeker and sharer) in helpful and solved threads were much more frequent than in unhelpful threads. Helpful and solved threads have significantly higher reciprocity communication patterns than unhelpful threads. In other words, reciprocity behavior in a Q&A discussion is more likely to lead to a helpful discussion.

4.3 Frequent Patterns in Conversation Networks

Although communication networks reveal the interactions between online discussion participants, they fail to reveal the conversational meanings embedded in the communication. A conversation network uses dialogue act tags to capture those meanings in an online discussion and shows their relationship over the course of the conversation. With each post annotated with a dialogue act tag, a discussion thread is represented as a conversation network. In our study, we used the same subgraph mining algorithm, gSpan, used for analyzing communication networks, to discover frequent substructures in conversation networks.

We examined frequent subgraph patterns extracted from all conversation networks. Table VII summarizes subgraph patterns with at least three nodes and with overall frequency greater than 10%. We also compared these patterns' frequencies across the three types of threads. For each pattern, we highlight the highest frequency in a bold font. Then, we conducted difference-of-proportion tests between the thread type with the highest frequency and the other two types, respectively. The frequencies that are significantly lower than the highest frequency are annotated with *(p-value < 0.1) or **(p-value < 0.05).

According to our tests, only four patterns showed significant differences in frequency between unhelpful and helpful/solved threads. All four patterns showed significantly lower occurrences in unhelpful threads than in the other two types. These four significant subgraph patterns are shaded in Table VII. Moreover, there was no significant difference in frequency between helpful and solved threads. Therefore, we have reasons

Table VII. Frequent Conversation Subgraphs of Three or More Nodes

ID	Subgraph	Overall	Unhelpful (0)	Helpful (1)	Solved (2)
#3.1	Q-Q(A-A, A-A)	19.82%	2.78%**	21.05%	**35.14%**
#3.2	Q-Q(A-A(Q-ADD))	18.02%	11.11%*	**26.32%**	16.22%
#3.3	Q-Q(A-A, A-ADD)	18.02%	13.89%	13.16%	**27.03%**
#3.4	Q-Q(Q-CONF, A-A)	13.51%	**16.67%**	15.79%	8.11%
#3.5	Q-Q(Q-ADD, A-A)	12.61%	11.11%	**15.79%**	10.81%
#3.6	Q-ADD(A-A(A-ADD))	11.71%	0.00%**	**21.05%**	13.51%
#3.7	Q-Q(Q-ADD(Q-ADD))	10.81%	11.11%	7.89%	**13.51%**
#4.1	Q-Q(A-A(Q-ADD(A-ADD)))	10.81%	0.00%**	**18.42%**	13.51%

$^{*}p$-value < 0.1.
$^{**}p$-value <0.05.

to believe that these four subgraphs represent conversation patterns that are more likely to result in effective knowledge sharing in online discussions. For example, pattern #3.1 Q-Q (A-A, A-A) starts with an initial question followed by two answers from different users. It implies that if multiple knowledge sharers participate in answering a question, there is a better chance of having a helpful discussion. This finding is in line with the work of Hew and Cheung [2010] in that the number of participants in asynchronous online discussions is positively related to higher-level knowledge construction. The finding also backs the conclusion of Jeong [2003], where conflicting viewpoints are found to promote discussion, critical thinking, and a conclusive result. However, our finding disagrees with the conclusion of Schellens and Valcke [2006] that smaller groups are more likely to reach higher phases of knowledge construction because a large number of participants might suffer from cognitive overload. The disagreement may be caused by the context of online communities where generally the number of participants in each discussion thread is relatively small.

Patterns #3.2, #3.6, and #4.1 show that the original question needs to be amended after receiving an answer. If the thread initiator (i.e., the knowledge seeker) can raise a follow-up question and possibly receive additional information from another user, then this thread is more likely to result in either helpful or solved knowledge sharing. The follow-up posts improve the explicitation of questions and answers that are more likely to result in helpful discussion. This finding is a good supplement for the conclusion of Schellens and Valcke [2006] that only a small portion of explicitation activities was observed compared to other cognitive activities such as presenting new information and evaluation. In their study, discussion participants were specifically asked to ground their contributions in the resources made available. Therefore, the participants were not motivated to refine or elaborate on earlier ideas. Our finding shows that explicitation plays an important role in effective knowledge sharing.

We also conducted the subgraph analysis for two-node patterns as shown in Table VIII. Some patterns appear more often in effective than in ineffective discussions. We will further discuss how this bi-gram subgraph analysis result can help with process mining in the next section.

4.4. Process Mining and Analysis Results

The basic statistics of the datasets according to the process mining input requirements are shown in Table IX. We can see that on average those threads tagged as "Helpful" and "Solved" have more events per case (220/38 = 5.79 and 229/37 = 6.19, respectively) than those tagged as "Unhelpful" (139/36 = 3.86). In addition, as shown in Figure 7, "Helpful" and "Solved" threads tend to have much more cases with five or more events. Therefore, threads with more posts are more likely to be helpful. In general, having

Table VIII. Subgraphs with Only Two Nodes

ID	Subgraphs	Overall	Unhelpful (0)	Helpful (1)	Solved (2)
#2.1	Q-Q(A-A)	77.47%	63.89%	**86.84%**	81.08%
#2.2	Q-Q(Q-ADD)	28.83%	**36.11%**	26.32%	24.32%
#2.3	Q-ADD(A-A)	27.03%	16.67%*	**42.11%**	21.62%
#2.4	Q-Q(Q-CONF)	24.32%	**33.33%**	21.05%	18.92%
#2.5	Q-Q(A-ADD)	20.72%	13.89%	15.79%	**32.43%**
#2.6	Q-ADD(A-ADD)	20.72%	5.56%**	**28.95%**	27.03%
#2.7	Q-ADD(Q-ADD)	13.51%	13.89%	10.53%	**16.22%**
#2.8	A-A(RES)	10.81%	0.00%**	13.16%	**18.92%**

Table IX. Basic Statistics of Process Mining Input Datasets

	All	Unhelpful	Helpful	Solved
Cases (Threads)	111	35	38	38
Events (Posts)	588	135	220	233
Activities (Tags)	11	9	10	11
Resources (Posters)	309	94	115	127

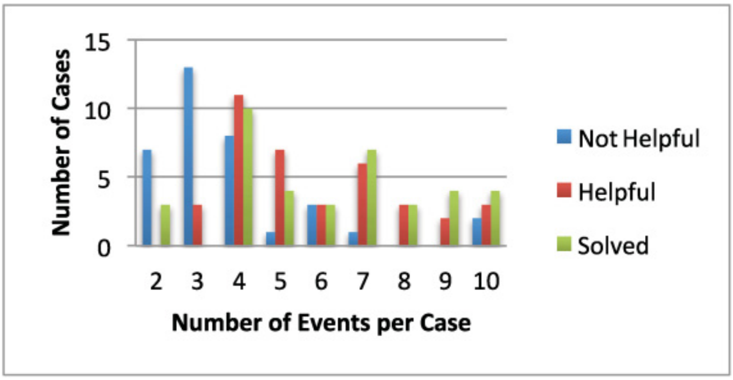


Fig. 7. Number of events per case distribution.

more posts per threads implies more participation and interaction from community users. Communities should try to increase the "length" of the knowledge-sharing processes for more effective problem solving.

We then summarize the distribution of different dialogue act tags for different types of threads in Table X. From this table, we can define the dialogue act Question-Answer ratio (Q/A ratio) as the total number of posts related to questions divided by the total number of posts related to answers:

Q/A ratio = (total number of posts with Q-Q, Q-ADD, Q-CORR, Q-CONF tags)/(total number of posts with A-A, A-ADD, A-CORR, A-CONF, A-OBJ, RES tags).

The Q/A ratios for "Unhelpful," "Helpful," and "Solved" threads are 1.77, 0.70, and 0.79, respectively. This shows that Unhelpful threads have more question-related posts than answer-related posts, whereas Helpful and Solved threads have more answer-related posts than question-related posts. The A-ADD dialog act is especially important among all answer-related acts. When an answer (A-A) is present, 89% of the threads with an A-ADD act are helpful or solved, whereas only 10% are unhelpful. The findings suggest that online Q&A communities should promote more answer-related posts, especially those supplementing an existing answer. The promotion can be realized by

Table X. Dialogue Act Tag Distribution Summary

	All	Unhelpful	Helpful	Solved
Q-Q	111	35	38	38
Q-ADD	103	31	33	39
Q-CORR	3	0	0	3
Q-CONF	43	17	14	12
A-A	129	27	50	52
A-ADD	95	10	47	38
A-CORR	7	1	4	2
A-CONF	9	3	2	4
A-OBJ	21	6	7	8
RES	25	0	11	14
OTH	41	5	14	22

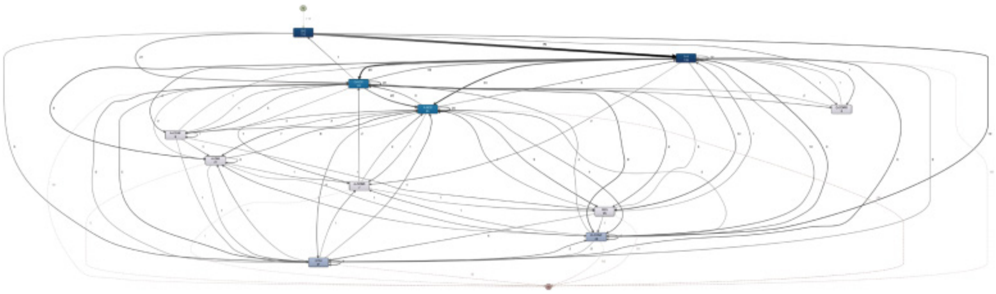


Fig. 8. A “Spaghetti” process model for the whole dataset.

many existing incentive features in online communities, such as reputation points being awarded to users who provide helpful information (answers) to questions [Cheng and Vassileva 2005].

After investigating the individual dialogue act tags, we studied the interactions among those tags in different threads via DAPP as defined in Definition 7. Overall, there are 85 unique DAPPs in our datasets, only 11 (12.9%) of which have been shared by at least two cases. This means that our dataset has a very high variation in case behavior, which may lead to large and complex process models. The fuzzy mining algorithm [Günther and van der Aalst 2007] used in Disco is designed to simplify “Spaghetti” process models mined out of diverse and less-structured event logs, which is the ideal algorithm for our study in this work compared with other process mining algorithms. We loaded five datasets into Disco, which included (1) All threads, (2) “Unhelpful” threads, (3) “Helpful” threads, (4) “Solved” threads, and (5) “Helpful” and “Solved” threads. When we chose to show all activities and paths of the discovered process models, we got Spaghetti process models as expected due to the high variation of case behavior. Figure 8 shows such a Spaghetti process model for the whole dataset, which is hard to interpret and analyze. Note that the Spaghetti process models are not necessarily incorrect—in other words, they are the results of precisely representing every detail of the highly diverse case behavior found in the datasets [Günther and van der Aalst 2007]. To simplify the process models for further analysis, we filter out some edges based on significance and correlation as described in Günther and van der Aalst [2007]. More specifically, we chose 0.5 as the parameter to filter out some less significant paths based on a trial-and-error approach, where we try to preserve as many paths as possible to get a more holistic view and also strive to achieve a certain level of process abstraction for better model readability and interpretation. Figure 9 shows the

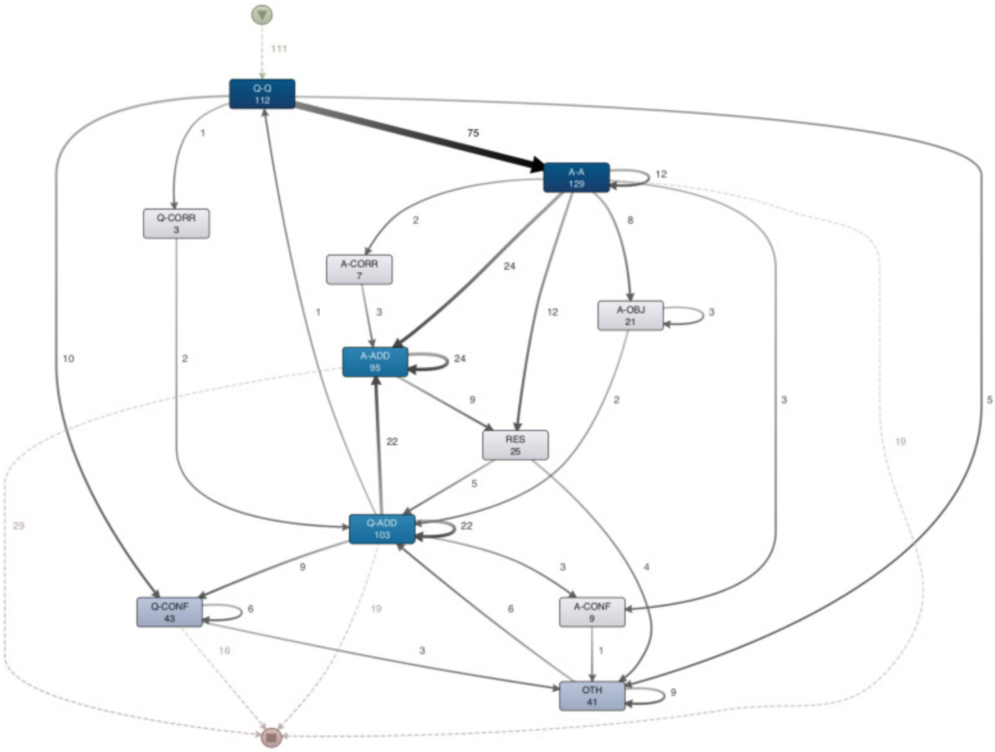


Fig. 9. Simplified process model for the whole dataset.

simplified process model for the whole dataset, which is much easier to analyze and can quickly help identify important paths among different dialogue act tags. The simplified process maps for other four datasets (Figures 10 through 13) are shown in the Appendix.

Based on the simplified process models, we can quickly identify high-frequency path pairs, such as <Q-Q, A-A>, and <A-A, A-ADD> in Figure 9. Table XI shows an analysis of the top high-frequency path pairs. The percentage in Table XI is calculated using the total frequency of the path pair divided by the total number of cases in each category—for example, <Q-Q, A-A> appears 17 times for 35 threads tagged as "Unhelpful," and its percentage is 17/35 = 48.57%. We also group "Helpful" and "Solved" threads as effective threads and treat Unhelpful threads as ineffective ones.

Next, we assign a PPQR for each path pair shown in Table XII in three steps: (1) we analyze the semantics of the path pair to see whether it is desirable for problem solving; (2) we then look at the data for that pair in Table XI to see whether the data is consistent with the semantics; and (3) we resolve conflicts between (1) and (2), if any, and assign a PPQR. Note that the result from the subgraph analysis of two nodes (see Table VIII) shows another type of correlation between dialogue act tags based on the reply-to relationship, which is different from the direct temporal execution relationship used in Table XI and provides another source of information for assignment of PPQR values. We compare the results shown in Tables VIII and XI and find that the subgraph analysis result is consistent with process mining analysis result. The detailed results of PPQR values for the top frequent path pairs are shown in Table XII.

To calculate GPQ scores based on Definition 8, we need to determine the parameter values for α, β, γ, and θ, which should be learned based on historical data. In this article,

Table XI. Summary of Top-Frequency Path Pairs

	All	Unhelpful/Ineffective		Helpful		Solved		Effective
		Total (T)	*Percentage* (P)	T	P	T	P	P
<Q-Q, A-A>	75	17	48.6%	31	81.6%	27	71.1%	76.3%
<A-ADD, o>	29	3	8.6%	9	23.7%	17	44.7%	34.2%
<A-A, A-ADD>	24	5	14.3%	8	21.1%	11	28.9%	25.0%
<A-ADD, A-ADD>	24	1	2.9%	10	26.3%	13	34.2%	30.3%
<Q-ADD, A-ADD>	22	1	2.9%	8	21.1%	13	34.2%	27.6%
<Q-ADD, Q-ADD>	22	6	17.1%	10	26.3%	6	15.8%	21.1%
<Q-Q, Q-ADD>	21	10	28.6%	4	10.5%	7	18.4%	14.5%
<A-A, o>	19	6	17.1%	7	18.4%	6	15.8%	17.1%
<Q-ADD, o>	19	12	34.3%	3	7.9%	4	10.5%	9.2%
<Q-ADD, A-A>	18	4	11.4%	7	18.4%	7	18.4%	18.4%
<Q-CONF, o>	16	9	25.7%	3	7.9%	4	10.5%	9.2%
<A-A, A-A>	12	0	0.0%	3	7.9%	9	23.7%	15.8%
<A-A, RES>	12	0	0.0%	4	10.5%	8	21.1%	15.8%
<RES, o>	10	0	0.0%	6	15.8%	4	10.5%	13.2%
<Q-Q, Q-CONF>	10	7	20.0%	0	0.0%	3	7.9%	3.9%

we use a heuristic-based approach to assign those values based on the frequency of path pairs found in our dataset as shown in Table XI. More specifically, we have very few path pairs with high frequency, which results in most of the pairs being assigned a PPQR of "Neutral." Among the high-frequency pairs, much more pairs are tagged as "Good" than "Bad." Therefore, we assign $\alpha = 1.5$, $\beta = 0.25$, and $\gamma = -2.5$, to roughly represent the data characteristics. Table XIII shows the patterns with highest frequency in each type of thread with corresponding normalized GPQ scores. We can clearly see that if we define normalized $\theta = 0.7$, all three most frequent paths for effective threads are correctly identified as golden paths. In addition, we investigate Figures 5 through 7 and find a few DAPPs with GPQ value of 1, which means that those paths are composed from all good path pairs, such as <*i*, Q-Q, A-A, RES, *o*>; <*i*, Q-Q, A-A, A-ADD, RES, *o*>; and <*i*, Q-Q, A-A, A-ADD, *o*>. Those ultimate golden paths are the most effective paths for problem solving.

4.5. Discussions

Our process mining approach to analyzing conversation processes in online knowledge-sharing discussions opens the field to business process modeling and mining techniques that have been mainly focused on operational business processes. Our innovative use of dialogue acts to convert online knowledge-sharing activities into "event logs" makes the application of process modeling and mining theories and techniques possible. We expect that more process-oriented research be done in the future to understand and theorize the process patterns of online knowledge-sharing activities.

The discussion characteristics and patterns identified and the process patterns discovered in Sections 4.2, 4.3, and 4.4 can be of use to researchers and practitioners in numerous ways, for instance, by enabling the following:

—*Automated detection of helpful versus unhelpful threads*: Automated helpfulness assessment supplements traditional information retrieval tools and algorithms, such as PageRank [Brin and Page 1998], which often rely primarily on a combination of relevance and centrality (authority). Users may then be able to more rapidly discover postings that are not just relevant and authoritative but also helpful rather than having to manually sift through content that is relevant but not helpful. Researchers can build novel information retrieval systems that supplement

Table XII. PPQR Analysis of Top Frequent Path Pairs

Path Pair	Meaning	Data Support?	Conflict Comments	PPQR
<Q-Q, A-A>	Right after a question is asked, an answer is provided.	Yes	No conflict	Good
<A-ADD, o>	The last post in the thread is either adding additional information to the answer or asking a follow-up question to the existing answer. Nevertheless, this last post is related to existing answers.	Yes	No conflict	Good
<A-A, A-ADD>	Right after an answer is posted, additional information of that answer is added or a follow-up related to that answer is asked.	Yes	No conflict	Good
<A-ADD, A-ADD>	Additional information about the answer is repeatedly added.	Yes	No conflict	Good
<Q-ADD, A-ADD>	Information is added about the original question and existing answers.	Yes	No conflict	Good
<Q-ADD, Q-ADD>	Additional information about the original question is repeatedly added, which does not help with solving the question per se.	No	Given that the percentage values for ineffective and effective are very close, Neutral is assigned.	Neutral
<Q-Q, Q-ADD>	Right after a question post, additional information about the question is posted.	Yes	No conflict	Bad
<A-A, o>	The last post of the thread is an answer, which is good.	No	Although the data is neutral, we assign Good based on clear semantics.	Good
<Q-ADD, o>	The last post of the thread is still adding additional information for the original question, which implies that the original question has not been answered.	Yes	No conflict	Bad
<Q-ADD, A-A>	After more information about the question is posted, an answer is provided.	Yes	No conflict	Good
<Q-CONF, o>	The last post of the thread is confirming the original question, which implies that the original question has not been answered.	Yes	No conflict	Bad
<A-A, A-A>	Two answers are provided back to back to the original question.	Yes	No conflict	Good
<A-A, RES>	Two answers are provided, and the second one solves the original question.	Yes	No conflict	Good
<RES, o>	The last post is the answer to the original question.	Yes	No conflict	Good
<Q-Q, Q-CONF>	Right after the original question is posted, another user confirms the question, which should be good.	No	The data conflicts with the semantics, but we believe that this pair should be tagged as Good.	Good

Table XIII. GPQ of Most Frequent DAPPs

	Most Frequent DAPP	Normalized GPQ
Unhelpful	<*i*, Q-Q, A-A, Q-ADD, *o*>	0.59
Helpful	<*i*, Q-Q, A-A, A-ADD, A-ADD, *o*>	0.93
	<*i*, Q-Q, A-A, Q-ADD, A-ADD, *o*>	0.88
Solved	<*i*, Q-Q, A-A, A-ADD, A-A, *o*>	0.88

relevance statistics with helpfulness estimates obtained using the characteristics and patterns that we have identified, even in the absence of explicit helpfulness assessments from users. Our GPQ scores can be a very useful input for such novel automated helpfulness assessment feature.

—*Automated detection of constructive versus destructive users*: Constructive users may be defined as those playing an active role in threads that are predominantly helpful, where helpfulness of each thread is estimated using the communication characteristics and patterns that we identified in Section 4. Destructive users may be defined as those playing an active role in threads that are predominantly unhelpful. Users who often associated with paths with low GPQ scores can also be potential destructive users. Forum owners can reward, promote, or leverage constructive users, and ban or moderate destructive users. For instance, to leverage constructive users, such users can be directed toward "high prospect" threads where golden paths are nearing completion and away from "low prospect" threads (threads exhibiting indicators of futility). This would lead to lower average time to resolution for issues, and higher productivity and satisfaction for constructive users. Destructive users can be directed away from the community and potentially to competitors; transference of burdensome "demon" customers can have a significant impact on profitability [Selden and Colvin 2003]. Researchers can investigate alternative definitions and metrics for constructive versus destructive users, as well as the success of a number of candidate strategies in addressing these users. The development of the strategies must be guided by user motivation theories in the context of participation in online communities, such as uses and gratifications theory and organizational commitment theory.

—*Automated detection of functional versus dysfunctional communities (forums or subforums)*: Functional communities may be defined as those with a large proportion of helpful threads, where helpfulness of each thread is again estimated from thread characteristics and discussion patterns, using the procedures that we have suggested. Dysfunctional communities may be defined as those with a small proportion of helpful threads. The total GPQ scores for all threads can also be used as a good indicator to evaluate whether the whole community is functioning well or not. In particular, the GPQ scores can be dynamically updated whenever new posts appear, which provides a real-time snapshot of the overall functional health of the community. Researchers and practitioners can rapidly identify sets of functional communities and dysfunctional communities, and then compare and contrast these communities to identify and possibly remediate sources of weakness. For example, dysfunctional communities may employ certain forum features (functions), configuration parameters, and/or style sheets that could be adjusted to better conform to benchmarks derived from functional communities. For businesses, marketing and product development resources could be focused on functional communities, which may deliver the greatest bang-per-buck for dollars invested in advertising, requirements elicitation, best practice elicitation, or other community-centric tasks. Standards organizations, industry trade groups, and search engines may use our computational mechanisms to rapidly rate communities and guide consumers to high-functioning communities, thus facilitating expedited growth of productive communities. Dysfunctional communities can be identified for possible remediation effort.

5. CONCLUSIONS

In this article, we proposed an analytical framework for understanding knowledge-sharing processes in online Q&A communities. Using the proposed framework, we examined the communication and process patterns of knowledge-sharing activities in online communities. Our analysis results showed that the proposed analytical framework was effective in revealing interesting findings from three different perspectives. By analyzing the communication exchanges of online discussions based on reply relationships, we found that online discussions with reciprocity communication patterns were more likely to be helpful for problem solving. By examining the conversational exchanges of online discussions based on both reply relationships and dialogue act tags, we observed that online discussions with answers from different users and information supplemental to original questions and answers are more likely to produce helpful knowledge. By modeling online discussions as event-based processes, we discovered that online discussions with too many question-related posts were less likely to achieve a problem solution.

Limitations of this research are related to the sample size, the generalizability of our research findings to other online Q&A communities, and an assumption that we implied in our analysis methods. Considering that manually tagging the posts in online discussions is time consuming, we only tagged a relatively small number of discussion threads for our analysis. An increased sample would improve the validity of our findings, albeit requiring more tagging efforts. In addition, all sample threads were selected from a single online community. Examining more online Q&A communities would definitely improve the external validity of our research findings. As a remedy, we provided quantitative descriptions of the user base and activity characteristics of the Apple discussion forum. One can compare those to the characteristics of a different community when applying our findings to it. Last, we simplified our analysis by implicitly assuming that the nature of the question would not have an effect on the conversation structure and process patterns, as well as on the helpfulness of the discussion. However, the nature of the questions, such as difficulty levels, will definitely affect the dynamics of conversations and the outcome of knowledge sharing because fewer people would know the answers. To recognize the nature of the question, text mining techniques can be used to identify hard problems, which may have unique conversation and process patterns.

We are planning to extend our work in a number of directions. First, the reply-to relationships represent strong correlations among different dialogue act tags. The existing fuzzy mining algorithm does not take such correlation into consideration when discovering dialogue act–based process models. We plan to enhance the fuzzy mining algorithm to create a unique DAP mining algorithm for better process discovery for the online community. Second, we are working on mechanisms to automatically tag posts once they appear in the online community, which can greatly increase our sample size because manual dialogue act tagging is time consuming. This requires text analysis of the post contents. We are planning to apply various text mining techniques to analyze the semantics of the posts. In addition, the post semantics analysis results also provide us more insights as to why a particular dialogue act tag sequence happens, which in turn help us to further craft the PPQR value for the sequence. Based on that, we can also develop algorithms to dynamically calculate GPQ scores for threads. Finally, in addition to network-related and process-related information, we may consider other types of information, such as the information quality of post content and the expertise and trustworthiness of post authors, to further reveal the differences between effective and ineffective knowledge sharing. From the network exchange perspective, additional information can be captured as attributes of nodes and links. Advanced subgraph mining algorithms such as those in Tong et al. [2007] can be used to extract frequent patterns that consider both structural characteristics and node properties. From the process perspective, online discussion participants with various levels of expertise are invaluable resources to the community. Therefore, each participant in an online discussion process can be considered as an event resource, which can be integrated nicely in process modeling. The goal will be to identify process patterns related to effective knowledge sharing with a minimum consumption of human capital.

APPENDIX

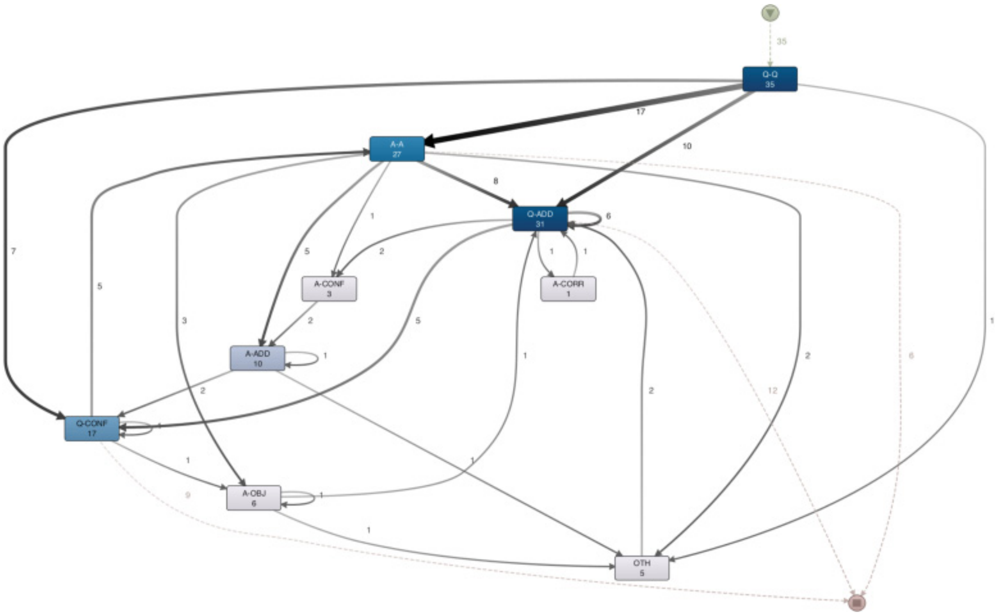


Fig. 10. Process model for the “Unhelpful” dataset.

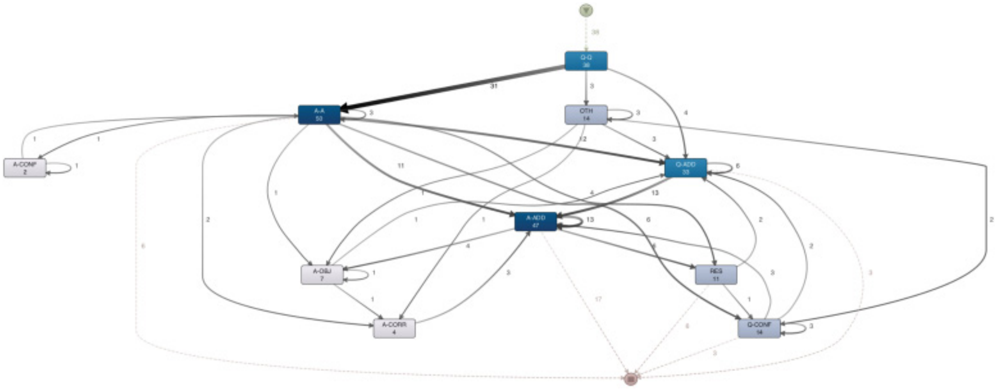


Fig. 11. Process model for the “Helpful” dataset.

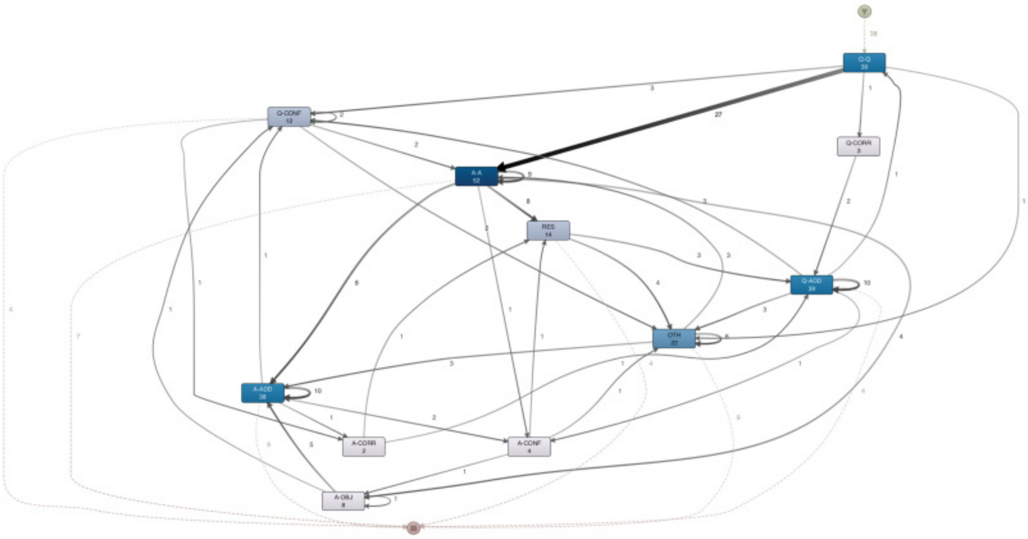


Fig. 12. Process model for the "Solved" dataset.

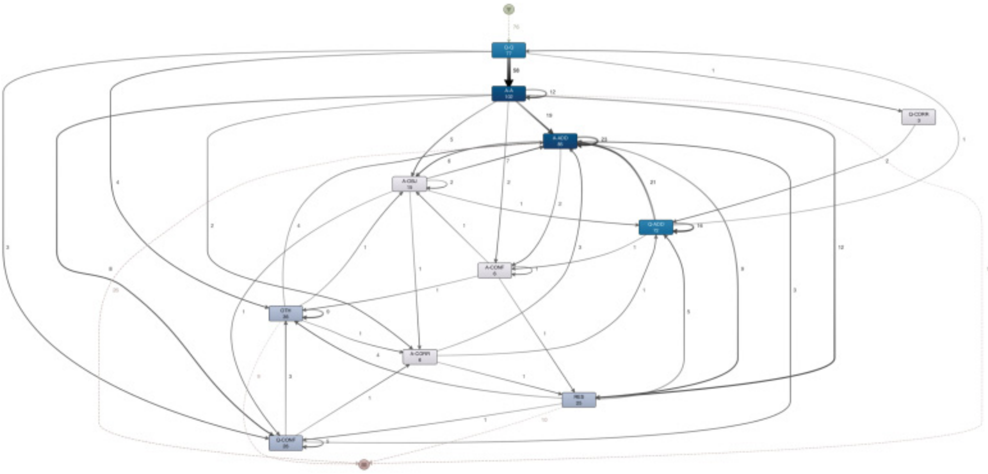


Fig. 13. Process model for the integrated "Helpful" and "Solved" dataset.

REFERENCES

W. Van der Aalst. 2012. Process mining: overview and opportunities. *ACM Transactions on Management Information Systems* 3, 2, 7.

W. M. P. Van der Aalst. 1998. The application of Petri nets to workflow management. *Journal of Circuits, Systems, and Computers* 8, 1, 21–66.

W. M. P. Van der Aalst, A. H. M. Ter Hofstede, B. Kiepuszewski, and A. P. Barros. 2003. Workflow patterns. *Distributed and Parallel Databases* 14, 3, 5–51.

W. M. P. Van der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. Van Dongen, A. K. Alves De Medeiros, M. Song, and H. M. W. Verbeek. 2007. Business process mining: An industrial application. *Information Systems* 32, 713–732.

W. M. P. Van der Aalst and A. Weijters. 2004. Process mining: A research agenda. *Computers in Industry* 53, 3, 231–244.

L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. 2008. Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. 665–674.

R. Agrawal and R. Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*. 487–499.

J. Bernoff and C. Li. 2008. Harnessing the power of the oh-so-social Web. *MIT Sloan Management Review* 49, 3, 36–42.

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks* 30, 107–117.

P. Burke. 1997. An identity model for network exchange. *American Sociological Review* 62, 1, 134–150.

Q. Cao, W. Duan, and Q. Gan 2011. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems* 50, 2, 511–521.

V. R. Carvalho and W. W. Cohen. 2005. On the collective classification of email "speech acts." In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*. 345.

V. R. Carvalho and W. W. Cohen. 2006. Improving email speech acts analysis via n-gram selection. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*. 35–41.

M. Chau and J. Xu. 2012. Business intelligence in blogs: Understanding consumer interactions and communities. *MIS Quarterly* 36, 4, 1189–1216.

R. Cheng and J. Vassileva. 2005. User motivation and persuasion strategy for peer-to-peer communities. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. 193a.

X. Cheng, C. Dale, and J. Liu. 2008. Statistics and social network of YouTube videos. In *Proceedings of the 16th International Workshop on Quality of Service*. 229–238.

C. Chiu, M. Hsu, and E. Wang. 2006. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems* 42, 3, 1872–1888.

H. Clark and S. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley (Eds.). American Psychological Association, Washington, DC, 127–149.

W. Cohen, V. Carvalho, and T. Mitchell. 2004. Learning to classify email into "speech acts." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 309–316.

D. Constant, L. Sproull, and S. Kiesler. 1996. The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization Science* 7, 2, 119–135.

K. Cook and E. Rice. 2001. Exchange and power: Issues of structure and agency. In *Handbook of Sociological Theory*, J. H. Turner (Ed.). Springer, New York, NY, 699–719.

S. Faraj and S. L. Johnson. 2011. Network exchange patterns in online communities. *Organization Science* 22, 6, 1464–1480.

F. Flynn. 2005. Identity orientations and forms of social exchange in organizations. *Academy of Management Review* 30, 4, 737–750.

V. Gómez, A. Kaltenbrunner, and V. López. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*. 645–654.

C. W. Günther and W. M. P. Van Der Aalst. 2007. Fuzzy mining: Adaptive process simplification based on multi-perspective metrics. In *Proceedings of the 5th International Conference on Business Process Management*. 328–343.

S. Ha and H.-W. Suh. 2008. A timed colored Petri nets modeling for dynamic workflow in product development process. *Computers in Industry* 58, 2, 193–209.

J. Habermas. 1979. What is universal pragmatics? In *Communication and the Evolution of Society*. Beacon Press, Boston, MA, 1–68.

K. F. Hew and W. S. Cheung. 2010. Higher-level knowledge construction in asynchronous online discussions: An analysis of group size, duration of online discussion, and student facilitation techniques. *Instructional Science* 39, 3, 303–319.

E. Von Hippel and G. von Krogh. 2003. Open source software and the "private-collective" innovation model: Issues for organization science. *Organization Science* 14, 2, 209–223.

J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. 2004. Mining protein family specific residue packing patterns from protein structure graphs. In *Proceedings of the 8th Annual International Conference on Computational Molecular Biology (RECOMB'04)*. 308–315.

J. Huan, W. Wang, and J. Prins. 2003. Efficient mining of frequent subgraphs in the presence of isomorphism. In *Proceedings of 3rd IEEE International Conference on Data Mining (ICDM'03)*. 549–552.

ISO. 2010. Language resource management—semantic annotation framework—part 2: Dialogue acts. In *ISO/DIS 24617–2*, Geneva, Switzerland.

E. Ivanovic. 2008. *Automatic Instant Messaging Dialogue Using Statistical Models and Dialogue Acts*. Master's Thesis. University of Melbourne, Melbourne, Australia.

A. C. Jeong. 2003. The sequential analysis of group interaction and critical thinking in online threaded discussions. *American Journal of Distance Education* 17, 1, 25–43.

X. Jin, Z. Zhou, M. Lee, and C. Cheung. 2012. Why users keep answering questions in online question answering communities: A theoretical and empirical investigation. *International Journal of Information Management* 33, 1, 93–104.

C. Kadushin. 2005. Networks and small groups. *Structure and Dynamics* 1, 1, Article No. 5.

S. N. Kim, L. Wang, and T. Baldwin. 2010. Tagging and linking Web forum posts. In *Proceedings of the 14th Conference on Computational Natural Language Learning*. 192–202.

R. Kumar, J. Novak, and A. Tomkins. 2010. Structure and evolution of online social networks. In *Link Mining: Models, Algorithms, and Applications*, P. S. Yu, J. Han, and C. Faloutsos (Eds.). Springer, New York, 337–357.

M. Lahiri and T. Y. Berger-Wolf. 2008. Mining periodic behavior in dynamic social networks. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*. 373–382.

A. Lampert, R. Dale, and C. Paris. 2008. The nature of requests and commitments in email messages. In *Proceedings of the AAAI Workshop on Enhanced Messaging*. 42–47.

J. Landis and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1, 159–174.

M. K. O. Lee, C. M. K. Cheung, K. H. Lim, and C. L. Sia. 2006. Understanding customer knowledge sharing in Web-based discussion boards: An exploratory study. *Internet Research* 16, 3, 289–303.

Y. Liu, J. Bian, and E. Agichtein. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. 483.

R. Maier, T. Hädrich, and R. Peinl. 2009. *Enterprise Knowledge Infrastructures* (2nd ed.). Springer, Secaucus, NJ.

M. Marcoccia. 2004. On-line polylogues: Conversation structure and participation framework in Internet newsgroups. *Journal of Pragmatics* 36, 1, 115–145.

C. McInerney. 2002. Knowledge management and the dynamic nature of knowledge. *Journal of the American Society for Information Science and Technology* 53, 12, 1009–1018.

A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC'07)*. 29.

L. Molm, J. Collett, and D. Schaefer. 2007. Building solidarity through generalized exchange: A theory of reciprocity. *American Journal of Sociology* 113, 1, 205–242.

M. Newman. 2003. The structure and function of complex networks. *SIAM Review* 45, 2, 167–256.

N. B. Peddibhotla and M. R. Subramani. 2007. Contributing to public document repositories: A critical mass theory perspective. *Organization Studies* 28, 3, 327–346.

J. Pena-Shaff and C. Nicholls. 2004. Analyzing student interactions and meaning construction in computer bulletin board discussions. *Computers and Education* 42, 3, 243–265.

H. Pfister and M. Mühlpfordt. 2002. Supporting discourse in a synchronous learning environment: The learning protocol approach. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*. 581–582.

J. Preece. 2000. *Online Communities: Designing Usability and Supporting Sociability*, Wiley, New York, NY.

R. Putnam. 2001. *Bowling Alone: The Collapse and Revival of American Community*. Touchstone Books, New York, NY.

A. Qadir and E. Riloff. 2011. Classifying sentences as speech acts in message board posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 748–758.

P. Reimann, J. Frerejean, and K. Thompson. 2009. Using process mining to identify models of group decision making in chat data. In *Proceedings of the 9th International Conference on Computer Supported Collaborative Learning*. 98–107.

T. Rietveld and R. Van Hout. 1993. *Statistical Techniques for the Study of Language Behaviour*. Walter de Gruyter, Berlin, Germany.

T. Schellens and M. Valcke. 2006. Fostering knowledge construction in university students through asynchronous discussion groups. *Computers and Education* 46, 4, 349–370.

S. Schrire. 2004. Interaction and cognition in asynchronous computer conferencing. *Instructional Science* 32, 6, 475–502.

J. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language, Syntax, and Semantics*. Cambridge University Press, Cambridge, UK.

L. Selden and G. Colvin. 2003. *Angel Customers and Demon Customers: Discover Which Is Which and Turbo-Charge Your Stock*. Portfolio Hardcover, New York, NY.

M. Sharratt and A. Usoro. 2003. Understanding knowledge-sharing in online communities of practice. *Electronic Journal of Knowledge Management* 1, 2, 187–196.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. *The ICSI Meeting Recorder Dialog Act (MRDA) Corpus*, Berkeley, CA.

T. Spaulding. 2010. How can virtual communities create value for business? *Electronic Commerce Research and Applications* 9, 1, 38–49.

J. Sutanto, A. Kankanhalli, and B. C. Y. Tan. 2011. Eliciting a sense of virtual community among knowledge contributors. *ACM Transactions on Management Information Systems* 2, 3, 14.

D. Suthers. 2001. Collaborative representations: Supporting face to face and online knowledge-building discourse. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*.

D. Suthers, R. Vatrapu, and R. Medina. 2001. Beyond threaded discussion: Representational guidance in asynchronous collaborative learning environments. *Computers and Education* 50, 4, 1103–1127.

H. Tong, C. Faloutsos, B. Gallagher, and T. Eliassi-Rad. 2007. Fast best-effort pattern matching in large attributed graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*. 737.

J. B. Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research* 23, 1, 3–43.

G. Wang, X. Liu, and W. Fan. 2011. A knowledge adoption model based framework for finding helpful user-generated contents in online communities. In *Proceedings of the 2011 International Conference on Information Systems*. 15.

M. Wasko and S. Faraj. 2005. Why should I share? Examining social capital and knowledge contribution in electronic network of practice. *MIS Quarterly* 29, 1, 35–57.

S. Wassermann and K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, New York, NY.

A. Weinberger, M. Reiserer, and B. Ertl. 2005. Facilitating collaborative knowledge construction in computer-mediated learning environments with cooperation scripts. In *Barriers and Biases in Computer-Mediated Knowledge Communication*, R. Bromme, F. W. Hesse, and H. Spada (Eds.). Springer, 15–37.

T. Winograd and C. Flores. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Addison-Wesley, Norwood, NJ.

W. Xi, J. Lind, and E. Brill. 2004. Learning effective ranking functions for newsgroup search. In *Proceedings of the 27th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 394–401.

X. Yan and J. Han. 2002. Gspan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM)*. 721–724.

W. Zhang and S. Watts. 2003. Knowledge adoption in online communities of practice. *Systemes d'Information et Management* 9, 1, 81–102.

Received May 2013; revised March 2014; accepted April 2014