Identifying Implicit and Explicit Relationships Through User Activities in Social Media

Christopher C. Yang, Xuning Tang, Qizhi Dai, Haodong Yang, and Ling Jiang

ABSTRACT: Social commerce has emerged as a new paradigm of commerce due to the advancement and application of Web 2.0 technologies including social media sites. Social media sites provide a valuable opportunity for social interactions between electronic commerce consumers as well as between consumers and businesses. Although the number of users and interactions is large in social media, the social networks extracted from explicit user interactions are usually sparse. Hence, the result obtained through the analysis of the extracted network is not always useful because many potential ties in the social network are not captured by the explicit interactions between users. In this work, we propose a temporal analysis technique to identify implicit relationships that supplement the explicit relationships identified through the social media interaction functions. Our method is based on the homophily theory developed by McPherson, Smith-Lovin, and Cook [31]. We have conducted experiments to evaluate the effectiveness of the identified implicit relationships and the integration of implicit and explicit relationships. The results indicate that our proposed techniques are effective and achieve a higher accuracy. Our results prove the importance of implicit relationships in deriving complete online social networks that are the foundation for understanding online user communities and social network analysis. Our techniques can be applied to improve effectiveness of product and friend recommendation in social commerce.

KEY WORDS AND PHRASES: Explicit relationships, homophily theory, implicit relationships, social commerce, social media, social network analysis, temporal analysis.

In recent years, a new paradigm of social commerce has emerged along with the advancement of Web 2.0 technology. Social commerce refers to the delivery of e-commerce activities and transactions through online social media platforms, mostly in social networks and with Web 2.0 technology [26]. Two essential elements of social commerce are identified: social media exchanges and commercial activities, where commercial activities include but are not limited to marketing, advertising, customer relations management (CRM), collaboration, and others [26]. Enabled by Web 2.0 technology, customers have been gradually taking a more active role in social media to contribute business value. This phenomenon is called co-creation. It includes the production of declarative content (e.g., knowledge compendia, consumer reviews, multimedia content, blogs, and virtual worlds), the development of social capital and appropriable relationship value, and word-of-mouth promotion [44]. One illustrative success story of social commerce is Starbucks’ efforts in leveraging social media sites and the online customer community to enhance sales. One example is My Starbucks Idea (http://mystarbucksidea.force.com), an online customer community where Starbucks invites customers to share ideas about coffee flavors and services. Another example is the initiative to allow customers to buy drinks from Starbucks for faraway friends through Facebook.
The underlying platforms of such social commerce and co-creation activities typically host online communities in which consumers or online users interact electronically or share common interests. For example, e-commerce today relies heavily on brand communities, which are online communities based on social relationships among brand-loyal customers [34]. These brand communities can be highly productive generators of social capital [44]. Hence, attracting and retaining members for online communities has become an important task for many businesses, especially social media sites.

The speedy growth and increasing importance of social commerce has led researchers to study a variety of aspects around the participation in and impacts of online communities, calling for the development of new theories [26]. One of these aspects of online communities relates to user relationships, and identifying user relationships online is essential to social commerce. There are two reasons for this. First, the interpersonal connections among users or members of online communities are the major driver for user participation and loyalty. For example, on social media sites such as Twitter and Facebook, users form online communities based on common interests around specific firms, brands, or products [10]. Generally, users’ intentions to use social commerce and social media sites are influenced by interpersonal relationships in the form of social support [27]. Furthermore, users become more loyal and attached to the online community when they feel more familiar with and similar to other users or form close interpersonal ties [38, 40].

Second, users’ interpersonal relationships greatly affect the performance of social commerce activities such as marketing and sales on social media sites. The interpersonal contacts on social media sites are very efficient in spreading news [11] and provide effective channels for delivering targeted advertisements [25]. More important, online reviews and electronic word-of-mouth communication (e-WOM) are significant demand drivers for purchasing intentions and buying decisions among online community members [4, 36]. In addition to these direct interpersonal relationships, the like-mindedness or common interests among online community users also support the effective dissemination of product news. One example is the diffusion of songs among users in an online music social network who have no explicit interpersonal interactions but have common interests in similar music revealed implicitly through linked individual playlists [13].

The findings from the above-mentioned research show that the relationships among users in online communities not only motivate user participation and sustain social networks but also enable social commerce activities. Hence, an important task for the sponsor of an online community, such as a social media site, is to identify a user’s relationships in the community so as to leverage these relationships to influence the user, encouraging participation and improving the performance of social commerce activities. For this task, researchers in the field of electronic commerce have been developing methods for detecting communities through user interactions [2, 5, 14, 43], recommending items or friends [19, 22, 30], and identifying opinion leaders or influential users [8, 16, 24].

In this field, social network analysis and mining (SNAM) techniques are often applied to an accurate and comprehensive social network, and the social
network is built mostly based on users’ direct interactions [7]. For example, if a user posts a photo and another user tags this photo later, these two users are connected to indicate a virtual relationship because of their direct interaction via the use of the tag. We refer to such direct interpersonal interactions in online communities as explicit relationships. Explicit relationships are captured by online actions that users carry out with social media functions such as retweet, reply, dig, comment, thumb up/down, and tag. Although explicit relationships are easy to observe, they do not capture all user relationships, especially when some user interactions are subtle and private.

For example, let us assume there are two users A and B on Digg.com who share some common interests on given topics. User A may follow every story submitted by user B and submit his or her own stories but never “dig” or comment on B’s story. Such a common-interest-based connection between these two users is subtle and not shown as direct interpersonal interactions. We refer to such a connection as an implicit relationship in this paper. In other words, implicit relationships are drawn from the users’ related interests or activities that cannot be traced by any public records of interactions between two users on a social media platform. In summary, there are two types of relationships in social media sites and online communities: explicit relationships and implicit relationships. Current approaches for constructing online social networks are lacking methods for discovering implicit relationships, which may lead to an incomplete or inaccurate social network in terms of connections among users and hinder SNAM tasks. Such SNAM tasks include centrality measures, role identification, and community detection. Without properly identifying implicit relationships, the role identified for a particular user would be computed incorrectly or the boundary of the detected community would be inaccurately identified.

Therefore, in this paper, we propose a specific method to identify implicit relationships based on common interests according to McPherson, Smith-Lovin, and Cook’s homophily principle [31]. Our method is intended to complement the toolbox of SNAM techniques by constructing a social network with implicit relationships considered. According to the homophily principle, a contact between similar people occurs at a higher rate than among dissimilar people [31], which implies that users sharing common interests or similar behaviors toward external events are more likely to breed connections. This concept provides the theoretical foundation of our research.

Specifically, we represent a user by a user activity vector that captures a user’s daily activeness. We then utilize the spectral analysis technique, which is one type of temporal analysis, to quantify similarity between users. Two filtering methods are introduced to detect implicit user relationships from a user similarity matrix. In addition, we use logical operators to combine explicit and implicit relationships in order to gain a comprehensive view of the social network that can be used in SNAM. The effectiveness of our proposed method is verified on a testbed crawled from Digg.com.

The rest of the paper is organized as follows. The next section reviews related prior research on extracting user relationships. The third section presents the method we develop for identifying implicit relationships based on common interests. The fourth section describes our computational experiment
in testing and verifying our method, and the last section concludes with a discussion about contributions and future research opportunities.

Related Work

SNAM has been widely applied in studies of user interaction, community structure, information dissemination, and many other related domains. Scott [39] provided a comprehensive survey of social network analysis by reviewing its origins in sociology and graph theory, discussing network properties and widely used SNAM techniques for computing centrality and discovering components and cliques. One SNAM method is derived from an online analytical processing framework that can support repeated exploratory querying for changes in a large set of interacting entities [2]. A second method is a betweenness-based approach that efficiently detects user communities in a social network [6, 14]. In addition, user interaction patterns can be utilized to detect dynamic user communities and their evolutionary relationships [43].

Another method is collaborative filtering, which is widely adopted in designing recommendation systems. Various modifications have been proposed to incorporate additional features into the collaborative filtering approach. For example, a user similarity index based on the diffusion process [30] or topological features in a bipartite graph [19] can be included to enhance the collaborative filtering approach. One important step in the collaborative filtering approach is to measure linkages between users and discover influential users in social media [8, 16, 22, 24].

While most SNAM studies are based on direct user interactions or explicit user relationships, some researchers point out the need for studying the more subtle implicit user relationships that may be inferred from similarities between users. To improve the quality of a social network and tackle its sparsity issue, McPherson and colleagues introduced the notion of homophily and suggested that social relationships are likely to form between people of similar characteristics [31]. In addition, Provost and colleagues introduced methods to extract a “quasi-social network” from data on visitations to social networking pages [37]. A link is placed between two users in this quasisocial network depending on their visits to common Web pages. Similarly, Singla and Richardson proposed to study potential relationships between two users by investigating confounding factors such as zip code, age, and gender [41]. Anagnostopoulos and colleagues pointed out that it is an important indicator of a user relationship if a user’s action is triggered by another user’s action—namely, as induction [5]. Goyal et al. adopted the same idea to detect induction-based social correlations [17].

Link prediction is a technique for detecting implicit relationships among users. Given a snapshot of a social network, the link prediction technique infers which nonexistent interactions among users may occur in the future [28]. Liben-Nowell and Kleinberg categorized the existing link prediction approaches into two major types: methods based on node neighborhoods and methods based on the ensembles of paths [28]. One example of a neighborhood-based method is the approach that Newman presented, which computes the common
neighbors of two nodes in the collaborative network to represent their probability to collaborate in the future [35]. Adamic and Adar employed a score function that refines the simple counting of common neighbors by giving more weight to rarer neighbors [1]. In contrast, methods based on the ensemble of all paths take the distance between two nodes into account. For instance, Katz proposed a formula to measure the probability of a future link that sums over the paths between two nodes and gives greater weight to shorter paths [23]. This formula represents a social network as an adjacency matrix and defines the hitting time $H_{xy}$ as the expected number of random walks starting at $x$ to reach $y$, which can be used to represent their proximity. Liben-Nowell and Kleinberg proposed a normalized and symmetric version of hitting time to tackle the link prediction problem [28]. Besides the neighborhood-based and path-based approaches, a third approach to link prediction focuses on training a classifier that takes proximity, aggregation, and topological features into account to predict links in a social network [3, 29].

Although link prediction is useful in determining the potential ties between users, it should be differentiated from implicit user relationship detection. Most of the existing link prediction techniques need a social network as an initial input, and their goal is to predict future interactions given the snapshot of the existing network structure. In contrast, in the research we are presenting in this paper, we do not assume that such an initial social network is available when we start to discover the implicit relationships. We utilize temporal analysis to identify the implicit relationships to supplement the extracted explicit relationships. We also investigate the integration of the implicit relationships and explicit relationships to construct the social networks. Our method complements SNAM techniques.

Specifically, extending the previous work by Chien and Immorlica [9] and He et al. [20], we utilize temporal analysis techniques to identify implicit relationships between online social media users. To detect aperiodic and periodic events, He and colleagues [20] represented each unique vocabulary by a time sequence and studied how to cluster correlated sequences to capture an event. Similarly, Chien and Immorlica [9] investigated the temporal correlations of pairs of vocabularies to identify semantically related search engine queries. Generally, Web users’ activeness may correlate with the occurrences of the events that they are interested in. In other words, if we represent a user by a time sequence of daily activeness on social media, we may identify highly related users based on the similarity between their time sequences, even if they do not interact directly with each other.

To better illustrate this idea, let us consider a hypothetical example. There are three online forum users A, B, and C. Users A and B are soccer fans, but user C is not. We measure a user’s daily activeness by the number of messages the user posts each day. In Figure 1, the diamond line represents the daily activeness of user A. The square line and triangle line represent the daily activeness of user B and user C, respectively. A’s daily activeness shows the same pattern as B’s from May 18 to September 14, but user C shows a different pattern. Hence, we can tell that it is more likely that users A and B share some common interests during this period than it is that users A and C share common interests. If we can identify external events that occurred when the
users’ activeness spiked, we may find out that both users A and B are following the same event, for example, a sports event such as the Olympic Games. This example demonstrates how we can identify implicit relationships via temporal analysis.

Employing temporal analysis, we consider user daily activeness as a multivariate point process and represent user activeness by a time sequence, namely, a user activity vector. We then measure user similarities by spectral coherency scores and develop filtering algorithms to identify implicit user relationships based on a user similarity matrix.

**Temporal Analysis**

As illustrated in Figure 1, strong temporal relatedness of user activeness sequences signals a high likelihood of the existence of implicit relationships between the users. In this section, we focus on quantifying user similarities by employing spectral analysis of signal processing techniques. Signal processing techniques have been applied in a wide variety of fields, including genetics [33], economics [18], and neuroscience [21, 15]. Signal processing techniques have also been used to extract semantically related search engine queries [9] and cluster words in news streams to detect popular events [20]. In this research, we adapt spectral analysis for analyzing user daily activeness. We perceive a user’s activeness on a given day as a signal, and the time sequence of the user’s activeness over a given period as a spectrum of the signal. Assuming there are two users $i$ and $j$, we first represent them by two user activity vectors. Then we compute an autospectrum for each individual vector and cross-spectrum of $i$ and $j$. Finally, the spectral coherency between the two users’ spectrums of signals measures their similarity, which indicates the likelihood of implicit relationship. For the sake of convenience, we include
in Table 1 all of the notations that will be used throughout this paper as well as a glossary (in Appendix A) that explains the key concepts and terms of temporal analysis that are used in the technical discussion that follows.

**User Activity Vector**

Given any online social media, let $T$ be the period of time (in days). We represent each user by a vector defined as

$$A_u = [A_u(1), A_u(2), ..., A_u(T)],$$

(1)

where each element $A_u(i)$ represents the activeness of user $u$ on the $i$th day. $A_u(i)$ can be defined in a very flexible manner according to the application domain. Several attributes can be used to quantify user activeness depending on which online social media one is studying, for example, the number of messages a user posts daily, the number of URLs a user tags daily, the number of tweets a user posts or retweets daily, or the number of videos a user clicks and comments on daily. In a simple manner, $A_u(i)$ can be defined as

$$A_u(i) = \frac{M_u(i)}{M(i)},$$

(2)

where $M_u(i)$ is the number of messages posted by user $u$ on day $i$ and $M(i)$ is the total number of messages posted by all users, including $u$, on day $i$. Note that $A_u(i)$ can be defined in more intricate ways when more information is given.
In this paper, we focus our research on studying the implicit relationships among users in online social media and use data collected from Digg.com as our testbed. Digg.com is a popular social news site where stories (or news) are submitted by Digg.com users. Users interested in a story can make comments or “dig” the story. When a user digs a story, it means the user likes it. For each story in Digg.com, we collect its story ID, submitter user ID, IDs of all the comments on the story, user IDs of all the associated commenters, and the corresponding time stamps. Given all this information, we define $A_u(i)$ as

$$A_u(i) = \frac{M_u(i)}{M(i)} \times \log\left( \frac{S}{S_u} \right),$$  

where $M_u(i)$ is the number of messages (either stories or comments) submitted by user $u$ on day $i$; $M(i)$ is the total number of messages contributed by all users, including $u$, on day $i$; $S_u$ is the total number of unique stories that user $u$ participated in (submitted or made comments to) over time $T$; and $S$ is the total number of unique stories participated in by all users over $T$. Equation (3) consists of two components: $M_u(i)/M(i)$ and $\log(S/S_u)$. $M_u(i)/M(i)$ measures the number of messages that user $u$ contributes on day $i$ normalized by the total number of messages of day $i$. The higher the $M_u(i)$, the more messages that $u$ contributes, leading to a larger activeness of $u$ at day $i$. The term $\log(S/S_u)$ represents a user penalty for participating in too many different stories, which may imply that this user lacks any specific focus.

A group of $m$ users form an $m$-dimensional multivariate process. By employing the user activity vector defined above, the $m$-dimensional multivariate process can be denoted as $A$, where each row denotes a user and each column indicates the activeness of these users on a specific day:

$$A = (A_1, A_2, ..., A_m)^T = \begin{pmatrix} A_1(1) & \cdots & A_1(T) \\ \vdots & \ddots & \vdots \\ A_m(1) & \cdots & A_m(T) \end{pmatrix}. \quad (4)$$

**Spectral Analysis**

User behaviors in terms of multivariate time series are often rich in oscillatory nature. In the case of Digg.com, this means that the stories related to the real-world events and the Web users’ interests are usually volatile. Such a phenomenon can be naturally examined with spectral analysis. By computing the spectrum of user $i$, we can quantify the overall activeness of a user in online social media. To calculate spectral estimates of users (auto- or cross-spectrum and coherency), we first perform a Fourier transform on each user $A_i$. Given a finite-length user activity vector of discrete time process $A_i(t), t = 1, 2, ..., T$, the Fourier transform of the data sequence $A_i(t)$ is defined as follows:

$$\tilde{A}_i(f) = \sum_{t=1}^{T} A_i(t) \exp(-2\pi if t). \quad (5)$$
A simple estimate of the spectrum is taking the square of the Fourier transform of the data sequence, that is, \(|\hat{A}_i(f)|^2\). However, this estimate suffers from the difficulties of bias and leakage. To resolve these issues, we applied the multitaper technique to obtain smoother Fourier-based spectral density with reduced estimation bias [32]. In the multitaper technique, we apply \(K\) tapers successively to the \(i\)th user activity vector and take the Fourier transform as

\[
\hat{A}_i(f,k) = \sum_{t=1}^{T} w_i(k) A_i(t) \exp(-2\pi if t),
\]

where \(w_i(k)\ (k = 1, 2, \ldots, K)\) represent \(K\) orthogonal taper functions with appropriate properties. A particular choice of these taper functions, with optimal leakage properties, is given by the discrete prolate spheroidal sequences (DPSS) [42]. The multitaper estimates for the spectrum \(S_i(f)\) is defined as

\[
S_i(f) = \frac{1}{K} \sum_{k=1}^{K} |\hat{A}_i(f,k)|^2.
\]

Since the spectrum score of each user varies with different frequency, which implies that a user performs differently at different periods, in this work we define the dominant power spectrum of user \(i\) as its maximum spectrum value across all potential frequencies:

\[
S_i = \text{Max}_f (S_i(f)),
\]

which can be used to represent the overall activeness of user \(i\) in social media. Similarly, the cross-spectrum \(S_{ij}(f)\) between user behavior processes \(A_i\) and \(A_j\) is defined as follows:

\[
S_{ij}(f) = \frac{1}{K} \sum_{k=1}^{K} \hat{A}_i(f,k) \hat{A}_j(f,k)^*,
\]

where \(\hat{A}_i(f,k)^*\) denotes the complex-conjugate transpose of \(\hat{A}_i(f,k)\). We then have the spectral density matrix for the multivariate processes \(A\) as

\[
S(f) = \begin{pmatrix}
S_{1,1}(f) & \cdots & S_{1,p}(f) \\
\cdots & \ddots & \cdots \\
S_{p,1}(f) & \cdots & S_{p,p}(f)
\end{pmatrix},
\]

with the off-diagonal elements representing cross-spectrum and diagonal elements representing auto-spectrum.

### Spectral Coherency

Spectral coherency is a complex quantity that provides an estimation of the strength of coupling between two processes. We quantify the similarity of two user activity vectors by using spectral coherency. Spectral coherency for any pair of user activity vectors \(A_i\) and \(A_j\) is computed as
The absolute value of $C_{ij}(f)$ is defined as spectral coherency with a range from 0 to 1. A spectral coherency value of 1 indicates that the two signals have a constant relationship, and a value of 0 indicates the absence of any relationship. Although correlation can also indicate the coupling between two user activity vectors, we choose spectral coherency here since it not only tells us how similar two user activity vectors are but also tells us at which frequency these two user activity vectors are similar. In this study we obtain an overall spectral coherency score for each pair of users by summing their spectral coherency value across all frequencies:

$$C_{i,j} = \sum_{f} C_{i,j}(f).$$

(12)

It is important to note that this spectral coherency score between two users is employed to represent the similarity of these two users across all frequencies.

**Implicit Relationship Identification**

Once the similarity of each pair of users is computed, we obtain a user similarity matrix. In this subsection, we introduce two simple but effective filtering methods, simple filtering and mutual filtering, to extract implicit relationships.

**Simple Filtering**

Simple filtering is a static thresholding method. Concretely, during the thresholding process, each pair of users are marked as “related” if their similarity score (spectral coherency of two users) is within the top $K$ percent of user similarity scores in the entire similarity matrix and as “unrelated” otherwise. We then consider implicit relationships that exist between related users. Simple filtering is an easy-to-implement yet very effective method to uncover implicit relationships, and we demonstrate the method in the experiment section. If we tune the parameter $K$, this proposed method can consistently outperform the baseline approach. The implicit relationships detected by this approach supplement the explicit relationships to offer a more comprehensive view of how Web users are related and interact with each other.

**Mutual Filtering**

Different from simple filtering, which considers global threshold, mutual filtering takes each user’s local information into account and determines a threshold independently for each individual user. First, given a similarity matrix, for each
user $i$ we sort the similarity scores between $i$ and all other users in descending order and retrieve the top $K$ percent users according to their similarity score with $i$, denoted as $\text{Candidate}(i)$. Therefore, user $i$ has relatively higher similarity with the users in $\text{Candidate}(i)$. Second, for each user $i$ in $\text{Candidate}(i)$, we check whether $i$ also belongs to $\text{Candidate}(j)$ to ensure that $j$ has relatively high similarity with $i$ as well. If that is true, we retain $j$ in $\text{Candidate}(i)$; otherwise we remove $j$ from $\text{Candidate}(i)$. We ensure that each user will be associated with at least one other user. As a result, if $\text{Candidate}(i)$ is an empty set, we correlate $i$ with the user $p$ that has the highest coherency with $i$. Finally, except for the relationships in $\text{Candidate}(i)$ where $i = 1, 2, \ldots, N$, we set all other elements in the similarity matrix equal to zero. By considering the original similarity matrix as a fully connected network, this filtering step removes edges with relatively lower weights (similarity scores). The remaining sparse matrix indicates the identified implicit user relationships. Box 1 shows the pseudocode of the mutual filtering algorithm, where $N$ denotes the total number of users:

Different from simple filtering, mutual filtering takes each user’s local information into account and determines a threshold independently for each individual user, which has the potential to increase mutual filtering method’s robustness.

**Experiment**

As stated in the introduction, the major goal of this paper is to capture the user relationships in social media. In particular, we identify implicit user relationships, which to some extent can supplement explicit relationships and improve the performance of SNAM tasks. For our experiment, we recruited human annotators to go through the messages in the selected online social
media setting and manually label user relationships. The labeled user relationships were treated as gold standard to verify our proposed techniques. In this section we first introduce our testbed and gold standard. Then, we discuss two baseline methods that utilize only explicit user interactions to determine user relationships. Finally, we compare our proposed techniques to two baselines by using the F1 score as a measurement.

### Data Set

In this experiment, we built our testbed by collecting messages from Digg.com. (Refer to Appendix B for a detailed description of Digg.com. In this work, we utilized the public Digg Application Programming Interface (API) to collect data from Digg.com. Specifically, we arbitrarily selected five leading newswires nationwide: CNN, BBC, NPR, the *Washington Post*, and Yahoo! News. During a three-month period from March 1, 2011, to May 31, 2011, we collected all “top stories” (which are defined and recommended by Digg.com) from these five news services, their comments, and all associated user and time stamp information. For each story that we collected, we recorded its story ID, submitter ID, a brief abstract of the story, all comments and the commenters’ IDs, and the corresponding time stamps of each story and each individual comment. The collected items and their corresponding meanings are summarized in Table 2. The whole data set had 12,742 stories, 13,531 comments (590 stories had comments) and 286 active users (we defined the users who were active on more than ten days during this three-month period as active users and filtered the inactive ones).

### Gold Standard

Two annotators were recruited to analyze the testbed for generating our gold standard. These two annotators were both doctoral students in informatics at a university in the northeastern United States, and their research interests focus on social network analysis. Therefore, their expertise could contribute to building a reliable gold standard. Specifically, we generated our gold standard as follows:

<table>
<thead>
<tr>
<th>Items</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Story ID</td>
<td>Unique ID of a story on Digg.com</td>
</tr>
<tr>
<td>Comment ID</td>
<td>Unique ID of a comment on Digg.com</td>
</tr>
<tr>
<td>Submitter ID</td>
<td>Unique ID of a Digg.com user who posted a story</td>
</tr>
<tr>
<td>Commenter ID</td>
<td>Unique ID of a Digg.com user who commented on a story</td>
</tr>
<tr>
<td>Story abstract</td>
<td>Abstract of a story provided by the story submitter</td>
</tr>
<tr>
<td>Comment content</td>
<td>Content of a comment</td>
</tr>
<tr>
<td>Time stamp</td>
<td>Time stamp of a message posted by a Digg.com user; a message may be a new story or a comment to an existing story</td>
</tr>
</tbody>
</table>
1. Fifty pairs of Digg users in our data set were randomly selected.
2. For each pair of selected users, two annotators independently re-viewed what stories these two users participated in and what their common interests were in order to decide whether these two users should be connected or not. Specific judging rules were as follows: For each user in the pair, human annotators went through all the stories they participated in, classified all the stories into different subjects (e.g., politics, terrorism, business, and so on), and counted the number of stories under each subject. If the two users in the pair shared more than three subjects, and the total number of stories in all the shared subjects exceeded ten, then the two users were considered connected. For example, Table 3 shows two pairs of users and a summary of their stories in different subject categories.
3. The weighted kappa measure was computed to investigate the reliability of the generated gold standard. When disagreement happened, the two human annotators would discuss their diverging opinions until they reached a consensus.

As mentioned previously, to determine the reliability of the gold standard produced by the two annotators separately, we used a weighted kappa measure to compute the interrater agreement. The weighted kappa measure is a statistical measure extended from the kappa measure for computing the agreement between two ordered lists [12]. It has a maximum value of 1, which indicates a perfect agreement between two raters, and a value of 0 when the agreement is not better than by chance. In general, a weighted kappa measure value larger than 0.8 is considered to indicate a substantial agreement between the two raters [39]. In our experiment, the weighted kappa measure was 0.84, which means that the two annotators had a very high level of agreement and the gold standard was reliable.

**Baseline**

Explicit user interactions are widely adopted to represent user relationships for constructing social networks. In our first baseline, namely the $1 \times N$ baseline, two users A and B are connected if A submitted a story and B made a comment

<table>
<thead>
<tr>
<th>Connected</th>
<th>Unconnected</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User1</strong></td>
<td><strong>User2</strong></td>
</tr>
<tr>
<td>Business (5*)</td>
<td>Business (3)</td>
</tr>
<tr>
<td>Politics (3)</td>
<td>Politics (4)</td>
</tr>
<tr>
<td>Terrorism (4)</td>
<td>Terrorism (5)</td>
</tr>
<tr>
<td>Music (1)</td>
<td>Medicine (1)</td>
</tr>
<tr>
<td>Pets (1)</td>
<td></td>
</tr>
</tbody>
</table>

* Number of stories in the corresponding subject category.
on it or B submitted a story and A made a comment on it. The rationale of this approach is simple. If two users have a common interest, they will explicitly interact with each other by making comments. In the second baseline, namely the $N \times N$ baseline, users within a same story posted in Digg.com are fully connected regardless of whether they explicitly interact with each other.

**Evaluation Measurement**

Given the gold standard, we evaluated the performance of the baselines and our proposed approaches by F1 score. Specifically, given a prediction method, if two users are labeled as relevant by annotators and are also connected by this prediction method, then we consider that as a true positive (TP). Similarly, if two users are labeled as irrelevant by annotators but are connected by a prediction method, it is counted as a false positive (FP). However, if two users are labeled as relevant by annotators but are not connected by a prediction method, we claim that as a false negative (FN). Finally, if two users are labeled as irrelevant by annotators and are also disconnected, it is counted as a true negative (TN). Given these definitions, precision ($P$) is computed as: $P = TP / (TP + FP)$. Similarly, recall ($R$) is computed as: $R = TP / (TP + FN)$. Finally, the F1 score is defined as

$$F = 2 \times P \times R / (P + R)$$

The higher the F1 score, the better a prediction method performs.

**Experimental Results**

In the methodology section, we first proposed to employ spectral coherency to quantify user similarity and then introduced two filtering techniques to detect implicit user relationships. The first filtering method is called simple filtering and the second one is called mutual filtering. As discussed earlier, simple filtering relies on a parameter $K$ that ensures that only the relationships with the top $K$ percent spectral coherency score will be retained. Similarly, mutual filtering also depends on a parameter $K$ to filter out relatively weaker relationships for each individual user but two users have to be the top $K$ of each other. In this experiment, we compare simple filtering and mutual filtering with two baseline approaches. Since baseline methods capture explicit relationships and simple and mutual filtering retrieve implicit relationships, to supplement explicit relationships with implicit relationships, we employ the standard logical operators OR/AND to integrate explicit and implicit relationships. Specifically, if a relationship is captured by either simple filtering or the $N \times N$ baseline, the simple-filtering-OR-$N\times N$ method considers that this relationship is valid. If a relationship is captured by both simple filtering and $N \times N$ baseline, the simple-filtering-AND-$N\times N$ method considers this relationship to be valid. We define the other two methods, mutual-filtering-OR-$N\times N$ and mutual-filtering-AND-$N\times N$, similarly.
As shown in Figure 2, two baseline methods (explicit relationship identification) were not affected by any parameters and thus performed consistently. Among them, the $1 \times N$ baseline performed worse, and the $N \times N$ baseline achieved a relatively higher performance in terms of F1 score. This implies that only considering direct interactions is not enough to detect user relationships. The $N \times N$ baseline performed better than the $1 \times N$ baseline because it connects users within a same story even though they may not directly interact with each other.

By comparing the implicit relationship identification using simple filtering and mutual filtering to two baselines, we observed that these implicit relationship identifications consistently outperformed the $1 \times N$ baseline no matter what value of $K$ is predefined. However, these two filtering methods performed slightly worse than the $N \times N$ baseline when parameter $K$ was small. However, when $K > 0.2$, simple filtering outperformed the $N \times N$ baseline. Similarly, when $K > 0.15$, mutual filtering also outperformed the $N \times N$ baseline. Moreover, the improvements of both simple filtering and mutual filtering in comparison to the $N \times N$ baseline become greater along the increase of parameter $K$. With a small $K$, filtering methods only take user relationships with relatively higher spectral coherency into account, so that many real implicit relationships are discarded. In addition, mutual filtering performed slightly better than simple filtering, but the difference is not substantial. We conducted a $t$-test, which showed that there is no significant difference between the performance of mutual filtering and that of simple filtering ($p < 0.01$).

Furthermore, we utilized logical operators to combine explicit and implicit relationships. First, we found that using the OR operator can substantially improve filtering methods’ performance. Both simple-filtering-OR-$N \times N$ and mutual-filtering-OR-$N \times N$ outperformed baselines and filtering methods even when parameter $K$ was small. Second, the F1 scores of simple-filtering-OR-$N \times N$ and mutual-filtering-OR-$N \times N$ were relatively stable comparing to filtering methods. We also conducted a $t$-test that showed that there is no
significant difference between the performance of simple-filtering-OR-$N \times N$ and that of mutual-filtering-OR-$N \times N$ ($p < 0.01$). However, using the AND operator cannot increase but rather decreases filtering methods’ performance. Given these results, we consider that explicit relationships detected by direct interactions and implicit relationships detected by strong spectral coherency scores are actually quite different, so that using the AND operator to combine them leads to worse performance. However, it implies that our detected implicit relationships might be a good supplement to explicit relationships to gain a more comprehensive view of social networks.

In summary, two baseline approaches in this experiment represent two straightforward ways to discover explicit user relationships. The first baseline approach ($1 \times N$) assumes that one user is related to another user if a user explicitly comments on another user’s post. The second baseline approach ($N \times N$) assumes that participants in a post should all be related with regard to their interests. Figure 2 demonstrates that baseline ($N \times N$) outperformed baseline ($1 \times N$). This is no surprise, because baseline ($1 \times N$) only considers the interactions between commenters and the story submitter, and it ignores the fact that commenters will also interact with each other. In addition, it is encouraging to observe that when K is larger than 0.15 (0.2), mutual filtering (simple filtering) starts outperforming two baseline approaches, which indicates that our proposed implicit relationship identification methods are effective in identifying unobservable user relationships. In addition, as shown in Figure 2, mutual filtering outperformed those two baseline methods earlier than simple filtering (0.15 vs. 0.2), which may indicate the robustness of the mutual filtering method. Furthermore, neither filtering approach achieved obvious higher F1 score when K is larger than 0.3. It indicates that increasing threshold K may induce more noise and so more false positives may occur. In practice, it is better to choose a more conservative value for K instead of a large value. However, how to choose an optimal value for K is still an open question for us that we will further study in the future. Last but not least, it is important to understand to what extent the implicit relationships detected by our filtering methods overlap with the explicit relationships detected by the baseline methods, since our primary goal in this work is to supplement explicit user relationships with implicit user relationships. From Figure 2 we observed that, by using the OR operator, we achieved an even higher F1 score, whereas using the AND operator leads to a much lower F1 score. This result tells us that these filtering methods detected very different relationships from those in the baseline methods. In practice, we should use the OR operator to better supplement explicit user relationships with implicit user relationships. In short, our proposed spectral analysis is effective in detecting implicit relationships. It is beneficial to combine explicit relationships and implicit relationships to construct a more accurate social network.

By identifying implicit and explicit relationships, we can make valuable recommendations to users. These recommendations may encourage interactions among social media users, which would not be possible otherwise, due to the huge and sparse social networks. Increasing user interactions on a social media platform will generate business value because the high volume of traffic can attract sponsors in placing advertisements on the platform. The
business models of most social media platforms rely critically on revenue from advertisements. As a result, making effective recommendations based on both implicit and explicit user relationships is a valuable organizational capability in the social commerce field.

**Conclusion**

Our experimental results offer several valuable lessons. First, as demonstrated by the two baseline approaches, explicit relationships by themselves are insufficient to capture all the important user relationships in online social media. Second, our proposed method is effective in identifying implicit user relationships. In addition, by using the OR operator to combine implicit and explicit relationships, we can achieve even better performance. As a result, it is important that we supplement explicit relationships with implicit relationships in real-world SNAM tasks.

The technique we developed in this paper and our findings provide valuable implications for research on social commerce. Social commerce concerns social media technologies, user community interactions, and commercial activities [26]. Our research on implicit user relationships directly expands our understanding of user communities by showing the user activeness as one dimension for identifying communities of interests. This makes it possible for us to build a more inclusive user community and social network when we examine the formation, structure, evolution, and impacts of online communities. Our technique can be adopted for social commerce research as well as social network analysis in management and sociology. Furthermore, the implicit relationships we revealed in this paper also point out an effective channel for information diffusion on social media sites. Our technique for identifying such relationships provides a tool for studying the trajectories by which information is diffused, even when there are no explicit direct interpersonal interactions. Taking into consideration such diffusion processes will enhance research efforts in knowledge management and Internet marketing strategies.

In addition, our research has important implications for business practice. Our proposed technique has the potential to create business value by identifying users with common interests, which serves as the foundation of many real-world applications, such as common-interest-based product and friend recommendations. Social media sites can apply our technique to detect prospective users and recruit members for their communities, improving the performance of user community detection. Currently, most methods for community detection and friend recommendation rely only on explicit user relationships. In other words, they either detect a user community based on how users interplay with each other explicitly or recommend friends based on the observable friends of the focal user. By employing our techniques to detect implicit user relationships and supplement explicit user relationships, businesses engaging in social commerce, such as social media sites, can identify users with common interests but without observable direct interactions. By so doing, they are able to recruit user prospects for a particular community
and grow the community base effectively because these prospects do share common interests, with both implicit and explicit relationships considered. In addition, our method can be applied to recommend friends to a user based on implicit common interests. This is likely to expand the pool of potential friends and increases the acceptance rate of the recommendations.

In the meantime, social media sites can implement our method to better understand user interests by including users who quietly follow certain events or topics so that they can better design the functions and contents of their sites to increase user loyalty. Moreover, with the technique we have presented here, social media sites can obtain more comprehensive and accurate understanding of user social networks with implicit user relationships included. With such understanding, they can identify opinion leaders more accurately, and thus can expect to better match online advertisements with users and also improve the performance of the advertisements. At the same time, they can extend the reach of the advertisements to the users who simply follow the development of events quietly.

Along with the implications for research and practice, our research has limitations. One limitation is that the current experiment was conducted using the social activities on one social media site, Digg.com. The extracted activities depend on the functionalities of Digg.com. So, one possible direction for extending our work in the future is to apply our techniques to user activity data from other social media sites. Moreover, because of the limitation of the experiment data, the scalability of our techniques is unclear. Hence, in the future we plan to study the scalability of our techniques with big data and resolve the related issues. Last but not least, based on the homophily theory, our method is designed to detect potential social relationships between users with common behavior or interests. However, there are many other types of homophily-based relationships, such as relationships between users with common attitudes, abilities, or beliefs. In the future we are also interested in exploring new approaches to discover these social relationships to further facilitate co-creation in social media.

To sum up, we developed a method to represent online social media users by activity vectors to capture their temporal activeness. We then applied spectral analysis techniques on user activity vectors to detect implicit relationships between users. Furthermore, we utilized logical operators to combine explicit and implicit relationships to construct a more accurate and comprehensive social network. In our experiment we recruited annotators to manually label user relationships as gold standards. Our method was then compared with two baseline methods. The experiment results demonstrate that our method substantially outperformed the baseline methods in effectively identifying online user social networks. Our work has great implications for social commerce research and practice because the techniques we developed allow researchers and businesses to effectively construct more accurate social networks for online users that often offer the foundation for social network analysis, online community detection, and product and friends recommendation. Future research on this aspect could look into the scalability of our method and further extend our method to include other types of homophily-based relationships.
REFERENCES


10. Culnan, M.J.; McHugh, P.J.; and Zubillaga, J.I. How large U.S. companies can use Twitter and other social media to gain business value. MIS Quarterly Executive, 9, 4 (December 2010), 243–259.


15. Goebel, R.; Roebroeck, A.; Kim, D.; and Formisano, E. Investigating directed cortical interactions in time-resolved fMRI data using vector auto-
regressive modeling and Granger causality mapping. Magnetic Resonance Imaging, 21, 10 (December 2003), 1251–1261.


Appendix A: Glossary

*Auto-spectrum:* Auto-spectrum is the forward Fourier transform of cross-correlation of a process with itself.

*Cross-spectrum:* Cross-spectrum is the forward Fourier transform of cross-correlation. It provides a statement of how common activity between two processes is distributed across frequency.

*Fourier transform:* It is a mathematical transform that transforms a mathematical function of time into a new function whose argument is frequency.

*Multitaper technique:* It offers a different way for spectral estimation that can obtain smoother Fourier-based spectral estimation with reduced estimation bias.

*SNAM:* Social network analysis and mining (SNAM) is a process of quantitative and qualitative analysis of a social network.

*Spectral analysis:* In statistical signal processing, spectral analysis is the technical process of decomposing a complex signal into simpler parts, which helps to detect any periodicities in the data, by observing the peaks at the frequencies corresponding to these periodicities.

*Spectral coherence:* The spectral coherence is a measure by which two signals are compared in the frequency domain. It is a complex quantity that provides estimation of the strength of coupling between two processes.

*Spectral estimation:* The objective of spectral estimation is to estimate the power spectrum of a random signal from a time sequence. It can help to detect any periodicities in the data.

*Temporal analysis:* Temporal analysis is primarily concerned with the study of the time variation of physical processes.

Appendix B: An Overview of Digg.com

Digg.com is a popular social news Web site that allows users to share and discover content from the Web. The content may come from popular newswires such as CNN or the BBC, video-sharing Web sites such as YouTube, or other personal blogs. The most important function of Digg is allowing users to vote a story up or down, namely “digging” or “burying.” During this process, Digg users can discover interesting stories, discuss topics that they are passionate about, and connect with like-minded people.

Figure B1 is a screenshot of a story posted on Digg.com. A Digg user can submit any story to the system. Once the story is submitted, the title and a brief abstract of the story can be seen on Digg. Other Digg users can click the green “thumb up” button to dig this story or the red “thumb down” button to bury the story. In addition, Digg users can make comments on the story and also dig or bury other users’ comments.
CHRISTOPHER C. YANG (chris.yang@drexel.edu) is an associate professor at the College of Computing and Informatics, Drexel University. His recent research interests include health-care informatics, security informatics, social intelligence and technology, Web search and mining, knowledge management, and information visualization. He received his Ph.D. in computer engineering from the University of Arizona. Yang has published over 220 referred journal and conference papers in such venues as *ACM Transactions on Intelligent Systems and Technology*, *Journal of the American Society for Information Science and Technology*, various *IEEE Transactions*, *Decision Support Systems*, *International Journal of Electronic Commerce*, and *Information Processing and Management*. He is the associate editor-in-chief of *Security Informatics* and the co-editor of *Electronic Commerce Research and Applications*. He has edited special issues in various journals and frequently serves as an invited panelist in NSF and other government agency review panels.

XUNING TANG (xt24@drexel.edu) received a Ph.D. in information science from Drexel University. His research interests include social computing, data mining, and social commerce, in particular topic detection and tracking, dynamic community modeling, identifying influential users, recommender systems, and fraud detection. He has published widely in peer-reviewed journals and conferences proceedings, including various *ACM* and *IEEE Transactions*.

QIZHI DAI (qd24@drexel.edu) is an associate professor of management information systems at the LeBow College of Business, Drexel University. She earned her Ph.D. at
the University of Minnesota. Her research interests range from strategic and economic impacts of information systems to IT standards and standardization. Her work has appeared in *MIS Quarterly, Journal of Management Information systems, International Journal of Electronic Commerce, Electronic Markets, Information Technology and Management*, and other journals. She has served as an associate editor for journals and conferences on information systems.

HAODONG YANG (hy95@drexel.edu) is a Ph.D. candidate at the College of Computing and Informatics, Drexel University. He holds a master’s degree in information studies, text mining, and health informatics, especially drug safety signal detection from online health communities. His work has been published in the transactions of a variety of conferences on health-care informatics and electronic commerce.

LING JIANG (lj328@drexel.edu) is a Ph.D. candidate at the College of Computing and Informatics, Drexel University. She holds a master’s degree in information studies from Beijing Normal University. Her research interests include social media, health informatics, and consumer health vocabulary. Her work has been published in the transactions of a variety of conferences on health-care informatics and electronic commerce.
Copyright of International Journal of Electronic Commerce is the property of M.E. Sharpe Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.