1-1-2008

Expert Recommendation Via Semantic Social Networks

Daning Hu  
University of Arizona, hud@email.arizona.edu

J. Leon Zhao  
University of Arizona, lzhao@eller.arizona.edu

Follow this and additional works at: http://aisel.aisnet.org/icis2008

Recommended Citation


This material is brought to you by the International Conference on Information Systems (ICIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICIS 2008 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
EXPERT RECOMMENDATION VIA SEMANTIC SOCIAL NETWORKS

Recommandation experte via les réseaux sociaux sémantiques

Completed Research Paper

Daning Hu
Department of Management Information Systems, Eller College of Management
University of Arizona, Tucson, AZ 85721
hud@email.arizona.edu

J. Leon Zhao
Department of Management Information Systems, Eller College of Management
University of Arizona, Tucson, AZ 85721
lzhao@eller.arizona.edu

Abstract

The use of social network analysis (SNA) in the design of expert recommendation systems is becoming increasingly popular. However, the experts recommended from such systems often do not meet users’ needs since the network semantic information is largely ignored. In this study, we used conditional logistic analysis to quantitatively examine the semantics of two social networks in a large open source community called Ohloh. It was found that homophily in nationality, location, programming language preference, and community reputation are determinants for forming evaluation and collaboration relationships among the Ohloh members. Moreover, past collaborations and mutual acquaintances are also found to significantly affect the formation of evaluation links but not collaboration links. In addition, we demonstrated how to embed the discovered network semantics into the design of expert recommendation systems through two mechanisms - user-based link prediction and Top-N most recognized mechanism.

Keywords: Semantic, social network analysis, expert recommendation, open source community

Résumé

Les recommandations expertes basées sur l’analyse des réseaux sociaux ne répondent pas aux besoins des utilisateurs puisque l’information des réseaux sémantiques est largement ignorée. Dans cette étude, nous utilisons une analyse conditionnelle logistique pour mettre à jour les sémantiques de deux réseaux sociaux. Nous montrons également comment inclure les sémantiques de réseau dans la conception des systèmes de recommandation experte au travers de deux mécanismes.
Introduction

The notion of social networks and the methods of social network analysis (SNA) are becoming increasingly popular in the design of expert recommendation systems, most notably in ReferralWeb (Henry et al. 1997), Yenta (Leonard 1997), and Expertise Recommender (David et al. 2000). These systems search the users’ social networks for people with appropriate expertise to answer questions, solve problems or provide collaborations (Jun et al. 2005). However, users often either find that the recommended experts do not meet their needs or that those experts are difficult to collaborate with. This is mainly because social network analysis embedded in existing expert recommendation systems so far has largely ignored network semantics and has only focused on global topological measures.

In this study, the semantics of social networks are defined as the social factors that significantly influence the relationship formation among the individuals. In social network analysis, these factors are called the determinants of link formation (Daning et al. 2008; Kossinets et al. 2006; Powell et al. 2005). For instance, two programmers’ past collaboration experiences in a software development project may have an influence on their mutual positive evaluations.

These network semantics can lead to significant differences in users’ relations with recommended individuals and users’ perceptions of individual expertise, hence affecting the perceived usefulness of expert recommendation systems. Therefore, network semantics are greatly needed to support better SNA based expert recommendation. According to prior research (Kossinets et al. 2006; McPherson et al. 2001; Powell et al. 2005), these semantics include:

- **Individual attributes**: People in a social network vary in their attributes such as age, gender and experiences. Unlike nodes in theoretically constructed networks, a person in real-world networks forms relationships (links) by considering individual attributes of the candidates.

- **Shared affiliations**: The links (i.e. relationships) among people in a real-world social network are highly meaningful and vary greatly (Wasserman et al. 1994). The link formation process is largely driven by the shared affiliations of their members (e.g., attending the same class) (Kossinets et al. 2006).

However, SNA in existing expert recommendation systems has largely ignored network semantics mainly due to the lack of 1) effective methods to quantitatively identify network semantics, 2) large data sets about semantics of social networks, and 3) methods to embed network semantics into the system design. These challenges naturally lead to two research questions:

- How to quantitatively discover semantics in social networks?
- How to embed the discovered semantics into the design of expert recommendation systems?

To answer these two questions, we used multiple methods to study the semantics of two social networks in an online open source software (OSS) community – Ohloh. Ohloh hosts more than 11,800 projects that involve over 94,330 developers. Firstly we conducted topological analysis on the two social networks in Ohloh community. Secondly, we statistically examined the determinants of link formations in these two networks using a conditional (fixed-effects) logistic model (CLM).

In addition, we demonstrated how to embed the discovered network semantics in the design of expert recommendation systems through two mechanisms. The first mechanism is used to quantitatively reflect individual users’ perspectives on the recommended experts. The second mechanism is to construct semantic social networks which provide contextual information for more accurate expert recommendation.

The main contribution of this paper is to 1) propose a computational approach that derives useful semantics of social networks, and 2) embed these semantics into the design of expert recommendation systems to improve their perceived usefulness. The approach we used in this study may be generalized to the design of other SNA based information systems.

The remainder of this paper is organized as follows. In the next section, we provide a review of literature on SNA, expert recommendation and OSS community analysis. The third section introduces the dataset for this study. Then we describe the research design and the experimental results. After that, we demonstrate how to embed the discovered network semantics into two expert recommendation mechanisms. At last, we conclude and suggest directions for future work.
Related Work

Social Network Analysis

Social network analysis is originally developed and used in sociology research to analyze patterns of relationships and interactions among social actors, aiming to discover the underlying social structure. It has been widely used to study various real-world networks (Albert et al. 1999; Boissevain 1974; Kossinets et al. 2006; McPherson et al. 2001; Newman 2001b), including networks in open source software communities (Crowston et al. 2003; Grewal et al. 2006; Jin et al. 2005; Madey 2002; Wagstrom et al. 2005).

There are mainly two types of SNA studies. One focuses on the topological characteristics of social networks. In such studies, the structural properties of the nodes and links are examined to describe and explain how network topologies affect the functions and behaviors of complex systems (Albert et al. 2002). Another line of SNA research studies the mechanisms and determinants behind the network dynamics. They mainly use statistical methods to model different network mechanisms. These models are then tested to account for the structural changes of network topologies.

However, while both types of SNA studies have made great efforts in analyzing the network topologies, little attention is given to the semantics of social networks mainly due to the lack of appropriate analytical methods and network semantic information. The existing analytical methods of SNA are introduced in the following sub-sections.

Topology Analysis

Several quantitative SNA measures are developed to describe network topologies at both individual and network level. At the individual node level, network centrality measures are used to identify key members and interaction patterns between sub-groups. One of the most commonly used centrality measure—a node’s degree—is defined by Freeman (1979) as the number of direct links this node has. It measures how active a particular node is. A network member with a high degree can be the leader or “hub” in a network.

On the other hand, several network level SNA measures such as average degree, clustering coefficient, average path length, and degree distribution are developed to describe and distinguish different network topology models. Three models have been employed to characterize complex networks: random graph model (Erdos et al. 1960), small-world model (Watts et al. 1998), and scale-free model (Barabasi et al. 1999). In random networks, each node has roughly the same number of links which equals to its average degree.

Clustering coefficient is usually used to determine the small-world nature of social networks. It is the probability that two nodes with a common neighbor also link to each other. A small-world network usually has a significantly larger clustering coefficient (Watts et al. 1998) than its random model counterpart, indicating a high tendency for nodes to form clusters and communities. A small-world network also often has a relatively small average path length (i.e., average number of steps along the shortest paths for all possible pairs of network nodes) (Watts et al. 1998).

Degree distribution \( P(k) \) is the probability distribution of a node has exactly \( k \) links. Power-law degree distribution is used to characterize scale-free networks (Wasserman et al. 1994). In such networks, a small fraction of the nodes have a large number of links while a big fraction of nodes have just a few. This scale-free topology may be caused by the newly joined nodes’ preferential attachment to the existing nodes with high degrees (Albert et al. 2002).

In general, network topological analysis is good at describing the structure of complex network systems but lack the capabilities to explain the emergence of such topologies and analyze the determinants of various network processes such as link formation. As a result, statistical methods are used in SNA studies to complement the insufficiencies of topological analysis.

Statistical Analysis on Determinants of Social Network Links

Statistical analysis has been widely used to model topological changes of various networks (Albert et al. 2002). In such analysis, it is assumed that network structural changes are caused by certain stochastic processes of network effects such as reciprocity, transitivity, and balance. Thus several network topology models have been developed
based on these network effects. They were fitted to empirical data to identify which network effects account for the observed structural changes (Snijders 2001).

However, in this study, we focus on another type of statistical analysis that has been used to identify and examine the determinants of network link formation processes. Such analysis is widely adopted in many domains such as organizational studies, sociology, and network analysis. For instance, in organizational studies, Beckman et al. (2004) studied inter-organizational network change by statistically examining factors that affect the firms’ choices of partners. They analyzed data on alliance networks for the 300 largest U.S. firms from 1988 to 1993. The results showed that the stability of a firm’s alliance network structure depends on the type of uncertainty it experienced. The greater the uncertainty that a firm faces alone, the more likely this firm will expand its alliance network. Likewise, the greater the uncertainty that a firm’s market or industry faces, the more likely that firm will strengthen the ties it presently has with others.

In the sociology literature, Leenders (1996) used a continuous-time Markov model to study the determinants of link formation in a children’s friendship network. The results showed that the homophily in gender (i.e. being the same gender) significantly affects the link (friendship) formation among children. The Markov model assumes that only the state of the network at time t-1 affects the current state (at time t). However, this assumption may not be valid for most real-world networks. Not only limited to friendship, McPherson et al. (2001) argue that various other social relationships, including marriage, work, advice are also influenced by the homophily principle - similarity breeds connection. In addition, Snijders (2001) developed a class of actor-oriented models to examine if the nodes adjust their linking choices in the network based on certain parameters such as their degrees. However, these models assume that the nodes are aware of their positions with respect to the whole network which is often not true in large complex networks.

Another study done by Powell et al. (2005) examined the determinants of the partner selection process for biotechnology firms in 1990s. They examined several types of determinants such as preferential attachment and homophily (i.e. people tend to interact with others having similar characteristics) using McFadden’s (1980; 1974) discrete choice model, a variant of the conditional logistic model. This econometric model is usually used to statistically analyze the human behavior of making choices. In this model, a subject is presented with choice alternatives and asked to choose the best alternative. In addition, the explanatory variables are alternative-specific or subject-specific. One limitation of this model is that it requires detail personal information of the subjects and the alternatives. Such information is usually not provided in existing empirical data sources.

In addition, longitudinal network data were employed to study network determinants too. Kossinets and Watts (2006) used Cox survival analysis to identify determinants of the email link formation in a university campus over a year time period. They found that the mutual acquaintance (i.e. two individuals are acquainted with a common person) and shared class affiliations (i.e. attending the same class) significantly affect the future email link formation between two students. In addition, a similar survival analysis approach was also used by Nerkar and Paruchuri (2005) to determine that if network centrality of inventors had a statistically significant effect on the intra-firm citation of their patents. Survival analysis lends itself well to the longitudinal analysis of network data since it involves the modeling of time to event data. In the context of semantic analysis, an event is the formation of a link. However, sometimes the time information is not available or inaccurate in the network datasets, which makes survival analysis not suitable for identifying semantics.

In general, most existing statistical models for analyzing the determinants of network link formation are limited by specific assumptions or data completeness issues. There is a lack of statistical techniques which are general enough and can be applied on empirical network datasets with missing information.

**Expert Recommendation and Social Network Analysis**

Systems that help find individuals with expertise required by users are defined as expert recommendation systems. Recently the use of social network analysis in such system is becoming increasingly popular. Henry et al. (1997) developed a system called ReferralWeb which provides referrals via chains of named individuals. The users may choose to search for referrals to people who are closely linked with famous, trusted experts to help them. Another referral based system Yenta (Leonard 1997) was designed to find experts having similar interests with the users. In addition, Expertise Recommender (David et al. 2000) used the distance between the user and the expert in their social network to filter recommended experts. If the distance is less than a threshold, the recommended expert is added to the final recommendation.
However, David (2003) found that the experts recommended by SNA based systems often do not match users’ specific needs and the perceptions of their personal social networks. That is mainly because the semantics of social networks are largely ignored.

**Open Source Software Community and Social Network Analysis**

**Open Source Software Community**

Nowadays, OSS communities have emerged as a major place for software developers to seek help and share knowledge. Thus many researchers have begun to study the OSS community, aiming to find out how it is related to the success of OSS software development. Such studies mainly focus on two topics. The first topic is the composition of the OSS community. Koch et al. (2002) analyzed the logs of source code changes for an OSS project and identified a core set of developers who produce most of the source code output. Such core OSS community members are also found to have most intensive communications in a project (Robertsa et al. 2006).

Another set of OSS community composition research focuses on the participation process of members. For instance, von Krogh et al. (2003) found that new OSS community members gain benefits from specializing their initial contributions. Robertsa et al. (2006) have developed a theoretical model and evaluated it using empirical data from the Apache projects, trying to understand how participations, motivations and performance of OSS community members interrelate. The results showed that people with higher status motivations are more likely to contribute. Another empirical study (Bagozzi et al. 2006) surveyed 402 active members from 191 Linux User Groups (LUG) in 23 countries and found that the participation to LUG is positively related with the person’s experience level in Linux. In our study, the participation of an OSS project is modeled as joining the collaboration network of Ohloh community. Therefore, according to these prior studies, the link formation process of Ohloh collaboration network is likely to be influenced by members’ attributes such as community status and experiences.

The second research topic is studying various relationships among OSS community members. Most such studies focused on the collaboration relationship. Ducheneaut (2005) observed that successful OSS developers progressively enroll into a collaboration network of human and material allies to support each other. Another descriptive study (Yutaka et al. 2000) found that the communication in OSS development collaborations heavily relies on electronic media (e.g., forum, mailing lists) rather than face-to-face contact. In addition, Bergquist et al. (2001) found that OSS community members gain trust from others by actively giving out high quality source code and answering questions. However, the above research mainly focused on the relationships at a micro level. The overall network effects on OSS communities caused by the aggregation of multiple relationships are largely ignored. To address this problem, a stream of literature using social network analysis methods studied the topologies of social networks in OSS communities. We introduce these studies in the following section.

**Social Network Analysis on OSS Communities**

Social network analysis has been widely used to modeling and analyzing various relationships in OSS communities, especially the collaboration relationship. Madey (2002) uses SNA methods to study a collaboration network of OSS developers in SourceForge.net and found it displays the scale-free network features. The small fraction of the developers with a large number of collaboration links can be explained by people’s tendency to collaborate with high-profile, skillful members. A more recent empirical analysis (Jin et al. 2005) of SourceForge data has discovered similar scale-free features in the collaboration network. Moreover, small-world network features – large clustering coefficient and small average path length – were also found in those SourceForge networks. Crowston (2003) have studied the topology of OSS collaboration networks using data from bug reports of 122 projects. It was found that the network topologies of bigger projects are less centralized. This may be caused by the modularization process of large OSS projects. Another SNA study (Wagstrom et al. 2005) used empirical data from blog links and mailing lists to simulated OSS network evolution, aiming to develop and validate a model which can explain developers’ choices in their project participations. In addition, Grewal et al. (2006) examined OSS collaboration network embeddedness and discovered it has more influence on the technical success than the commercial success of OSS projects.
Dataset

The dataset for this study was collected from a large online OSS community – Ohloh, which provides information about 11,800 OSS projects involving 94,330 developers. This data source is unique comparing with other major OSS communities such as SourceForge.net from two perspectives. Firstly, it provides evaluation information about OSS community members – the “Kudo” evaluation link. Each Ohloh member can send any other member a link called “Kudo” which is a simple gesture of thanks, praise, or endorsement. Sometimes a “Kudo” link can be given to a co-developer in the same OSS project as positive evaluation for his or her contribution. Sometimes people receive “Kudo” links from others as recognition of their programming skills or appreciation for their help. Therefore, the “Kudo” evaluation links may cover a lot of underlying social relationships among OSS community members. Moreover, Ohloh provides information about the individual attributes of registered developers while SourceForge.net does not. Such attributes includes nationalities, locations, and programming experiences. These attributes are crucial for the semantic analysis of social networks.

Secondly, Ohloh data set covers a more comprehensive list of major OSS projects than Sourceforge.net because of its data sources. It retrieves OSS related data from three major software revision control repositories – Subversion, CVS and Git while SourceForge.net only has data from Subversion.

In addition, Ohloh website provides several other types of information about OSS projects through its API. For example, the project activity information keeps track of every change made in OSS projects, including what was changed, when it was changed, and who made the change. Other global statistics such as programming language usage are also included. Such information coupled with the results from social network analysis may provide insights about the determinants of link formations in Ohloh networks.

Research Design

To address the first research question, we proposed to use a set of analytical methods include SNA topological analysis and conditional logistic analysis to discover semantics in Ohloh social networks. The research design is presented in Figure 1. It consists of two steps. The first step involves two components: network construction and semantic extraction. We constructed two social networks from Ohloh dataset – an evaluation network and a collaboration network – based on the “Kudo” evaluation links and past project participation information respectively. At the same time, the potential network semantics may be selected based on literature or theoretical conjectures on social networks in OSS communities. The second step, network analysis, contains both SNA topological analysis and semantic analysis. The details of the design are introduced in the following sub-sections.

Figure 1. Discovering Semantic of Social Networks in Ohloh Community
Network Construction

To construct social networks from Ohloh dataset, we first need to identify the network nodes and links. Among the 94,330 developers listed in Ohloh website, 14,075 of them registered with detail information such as location and nationality. The rest only have OSS development activity information retrieved from revision control repositories. Since only registered users are allowed to send and receive “Kudo” links in Ohloh community, the constructed evaluation network only contain 3,451 developers as nodes and 9,827 evaluation links among them.

In addition, the semantic analysis requires detail information about individual attributes and shared affiliations which is only contained in registered accounts of Ohloh community. Therefore, we also just include the registered developers in the Ohloh collaboration network. Each collaboration link indicates that the pair of developers has worked in the same OSS project before. The constructed collaboration network includes 3,798 registered developers with 77,513 collaboration links.

Semantic Extraction

In this study, the semantics are determinants of link formation in the Ohloh evaluation network and the collaboration network. These potential determinants were selected based on findings or conjectures from prior OSS studies. They include six individual attributes and three shared affiliations. The individual attributes selected were OSS experience, coding experience (Bagozzi et al. 2006), homophily (McPherson et al. 2001) in country, location, programming language, and community reputation, while the shared affiliations are participation in the same OSS project, mutual acquaintance for the collaboration relationship and the evaluation relationship. These potential determinants are explained in detail as follows.

Individual Attributes

- Coding experience: the total number of commits made by this developer for all OSS projects in Ohloh dataset (one commit refers to making changes for one time in the source code of an OSS project).
- OSS experience: the total number of months in which this developer made at least one commit to OSS projects.
- Developer degree: the number of links the developer has (incoming or outgoing) just prior to the link formation.

These individual attributes are selected assuming that a person’s OSS experience level is positively related to the collaboration or evaluation links he or she will receive in OSS networks. In sociology, such phenomenon is referred as accumulative advantage (Powell et al. 2005) or preferential attachment. In our study, they are reflected in the nodes’ degrees and experience of the developers.

- Homophily in Country: the developer's claimed country in his or her registration file.
- Homophily in Location: the developer's claimed living location (city level) in his or her registration file.
- Homophily in Programming language: the programming language most often used by this developer, measured by the total number of his commits in that language.
- Homophily in Community reputation: a score called KudoRank ranging from 1 to 10 calculated based on the number and quality of the “Kudo” links this developer received. In other words, a high KudoRank comes from not only receiving a lot of Kudo links, but also receiving Kudo links from highly ranked people.

Homophily is assessed in several ways. We measure the country and location differences between two developers because the homophily in geographic location is found to be a determinant for forming various relationships in prior research. We also measure homophily in primary programming language since knowing the same programming language is a usually a prerequisite for developers to collaborate in the same OSS project.

Shared Affiliations

- Participation in the Same OSS project: a binary indicator if two developers have worked in the same OSS project before current evaluation/collaboration link forms.
- Mutual Acquaintance for Collaboration: a binary indicator if two developers both link to a common node in the Ohloh collaboration network, indicating they both collaborate with the same developer before.
• Mutual Acquaintance for Evaluation: a binary indicator if two developers both link to a common node in the Ohloh evaluation network, indicating they both evaluate the same developer before.

The reason to include past OSS project collaborations is because past positive collaboration experiences may facilitate future collaborations. Moreover, mutual acquaintance usually serves a bridge for two individuals to get to know each other and further form various relationships (Kossinets et al. 2006).

**Network Analysis**

After we construct the two networks and extract potential determinants, we conduct SNA topological analysis on these two networks and examine the semantics using conditional logistic analysis.

**Topological Analysis**

There are two goals for the topological analysis of Ohloh networks in our study. Firstly, it help us uncover the structure of Ohloh open source software community, and better understand the nature of OSS collaboration and evaluation relationships. Secondly, the determinants of link formation are found to significantly affect the network structural changes (Kossinets et al. 2006; Powell et al. 2005). For instance, experienced developers may attract more collaborators causing a scale-free network topology. Therefore, the results of network topology analysis can be used to verify network determinants examined by statistical analysis.

We use SNA centrality measures to describe the topology of the Ohloh evaluation network and identify its key members. High degrees usually indicate high levels of activity and wide social influence. Therefore, the OSS community members with high degrees are likely to be the leaders of their networks. The average degree of a network is also calculated to measure how dense a network is.

In addition, previous research (Jin et al. 2005; Madey 2002) found that OSS collaboration networks are scale-free networks and have small-world network properties. Thus we examine both Ohloh evaluation and collaboration networks to see if they have these features. Several SNA measures are examined, including the average path length, the clustering coefficient, link density, and the degree distribution. These properties are then, checked against the small world and scale-free models.

**Statistical Analysis to Examine Semantics**

Our choice of a statistical model for analyzing determinants of Ohloh network link formation is based on both theoretical and empirical considerations. Theoretically, our study intends to model human choice behavior in social networks. The research question asks what determinants account for differential (as opposed to random) patterns of the link formations in OSS evaluation and collaboration networks. Empirically, for the evaluation relationship, we need to model the choice behavior of Kudo senders to receivers. For the collaboration networks, the choices are modeled as bi-directional between two developers who collaborated in an OSS project. For these reasons, in our study, we choose to use conditional logistic model that takes each choice as a unit of analysis, which in our case are the formation of an evaluation or collaboration link between two developers.

Conditional logistic model and its variations (McFadden 1980; McFadden et al. 1974; Powell et al. 2005) have been widely used to model human choice behavior and examine the determinants which affect the choices. In our study, for the Ohloh evaluation network, the probability of an OSS developer $i$ choose to send Kudo evaluation link to another developer $j$ from the alternative set $J_j$, is specified as follows:

$$\Pr(y_i = j) = \frac{\exp(X_j \beta)}{\sum_{j'} \exp(X_{j'} \beta)}$$

where $y_i$ is the observed choice for developer $i$ and $X_j$ is a vector of the characteristics of the developer $j$. The unknown coefficients $\beta$ are typically estimated by maximum likelihood methods.
We estimate the CLM for both the evaluation and collaboration network data using *clogistic* command in Stata 10/MP. The two dependent variables are binary indicators of the outcome for link formations in the two Ohloh networks. The independent variables are the selected potential determinants explained in the semantic extraction section. In addition, for the statistical analysis, these independent variables are operationalized as Table 1 shows.

<table>
<thead>
<tr>
<th>Table 1. Variables in Statistical Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
</tr>
<tr>
<td>Dependent Variables</td>
</tr>
<tr>
<td>Evaluation (Kudo) link</td>
</tr>
<tr>
<td>Collaboration link</td>
</tr>
<tr>
<td>Independent Variables</td>
</tr>
<tr>
<td>Coding experience</td>
</tr>
<tr>
<td>OSS experience</td>
</tr>
<tr>
<td>Developer degree</td>
</tr>
<tr>
<td>Same Country</td>
</tr>
<tr>
<td>Same Location</td>
</tr>
<tr>
<td>Same Programming Language</td>
</tr>
<tr>
<td>Same KudoRank</td>
</tr>
<tr>
<td>Past OSS project(s)</td>
</tr>
<tr>
<td>Mutual Acquaintance in the Evaluation Network</td>
</tr>
<tr>
<td>Mutual Acquaintance in the Collaboration Network</td>
</tr>
</tbody>
</table>

**Results**

*Topological Analysis*

We start from describing the basic statistics of the two networks in Table 2. Among the 9,827 evaluation relationships, about 93.1% of them have corresponding collaboration relationships before. This may imply that the past project collaborations may facilitate future evaluation relationships among OSS developers.

<table>
<thead>
<tr>
<th>Table 2. Key Statistics of Ohloh Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Network</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>No. of Nodes</td>
</tr>
<tr>
<td>No. of Links</td>
</tr>
</tbody>
</table>

Figure 2 shows a sample evaluation network and a sample collaboration network in Ohloh community. The sample collaboration network consists of dense, fully connected local clusters which represent OSS projects. In contrast, for
the evaluation network, it is less interconnected inside each cluster but there are more nodes and links serving as hubs and bridges among different clusters.

Figure 2. (a) A sample Ohloh evaluation network; (b) A sample Ohloh collaboration network

SNA centrality measures are used to describe the topologies of both Ohloh networks. Firstly, average degree was calculated for both networks. We found that the collaboration network has a much larger average degree than the evaluation network, indicating denser network structure. This structural difference may be caused by the different natures of these two types of links. Collaboration links indicate a developer’s work relationships with all the project members while evaluation links are more selective and personal relationships.

Prior studies (Jin et al. 2005; Madey 2002) found that OSS collaboration networks are scale-free networks and have small-world network properties. Table 3 shows that several SNA measures of the two Ohloh networks and a random network with similar link density. These measures were examined for all three networks, including the average path length, the clustering coefficient, and link density. The results then were checked against the small world and scale-free features. In addition, the degree distributions of the two networks were fit to power-law distribution using linear regression technique to test for scale-free topological features.

| Table 3. SNA Measure of Ohloh Networks and a Random Network |
|---------------------------------|-----------------|-----------------|-----------------|
|                                | Evaluation Network | Collaboration Network | Random Network |
| Average Degree                 | 4.44             | 27.60            | N/A             |
| Average Path Length            | **5.643**        | **4.187**        | 5.110           |
| Clustering Coefficient         | **0.455**        | **0.876**        | 0.002           |
| Link Density                   | 0.0017           | 0.0077           | 0.0017          |
| Degree Distribution            |                 |                 |                 |
| $R^2$                          | **0.91**         | **0.92**         | N/A             |
| $\gamma$                       | 1.95             | 1.86             | N/A             |

In our topological analysis, we focused on the largest connected clusters in the two Ohloh networks. Firstly, we found that both networks are small world networks. Their average path lengths are small with respect to their sizes. Thus, an Ohloh member can reach any other member in both networks through just 4 or 5 mediators. Another small-world property, high clustering coefficient (comparing with their random network counterpart), is also found for both networks. The clustering coefficients are significantly higher than their random graph counterpart in the fourth column.
In addition, the evaluation network is very sparse with a low link density (Wasserman et al. 1994) of 0.0017. This property has important implications for the cost of sharing codes and other resources in OSS communities. Since such cost increases as more people join in and their relationships become denser in one project (cluster), the small average path length and link sparseness can help lower costs and enhance communication efficiency for the overall network.

The last three rows of Table 3 show the results of linear regressions of the degree distributions for both evaluation network and collaboration network. It was found that the evaluation network follows the power-law degree distribution (Newman 2001a), $p(k) \sim k^{-\gamma}$, with exponent $\gamma = 1.95$. The coefficient of determination $R^2$ of the regression for evaluation network is extremely large at 0.91 (ranging from 0 to 1), indicating high fitness of the power-law degree distribution. The collaboration network also has similar results and shows scale-free features with $R^2$ at 0.92 and $\gamma = 1.86$.

**Statistical Analysis**

To facilitate interpretation of results from CLM analysis, we present the odds ratios and the coefficients as Table 4 shows. The odds ratios are obtained from the coefficients by using exponential function. That is $X_i = \exp(\beta_i)$ where $\beta_i$ is the coefficient. The odds ratio measures the change of the odds that a link is formed caused by each unit increase in an independent variable (odds ratios equal to one means no effects and less than one reflect negative effects). This means that the probability of the link formation would increase by a factor of odds ratio when the corresponding independent variable increases by one unit.

<table>
<thead>
<tr>
<th></th>
<th>Evaluation Network</th>
<th>Collaboration Network</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds Ratio</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Coding experience</td>
<td>1.000**</td>
<td>0.001*</td>
</tr>
<tr>
<td>OSS experience</td>
<td>1.000*</td>
<td>0.001*</td>
</tr>
<tr>
<td>Developer degree</td>
<td>1.049**</td>
<td>0.048**</td>
</tr>
<tr>
<td>Same Country</td>
<td>5.343**</td>
<td>1.676**</td>
</tr>
<tr>
<td>Same Location</td>
<td>9.017**</td>
<td>2.199**</td>
</tr>
<tr>
<td>Same Programming Language</td>
<td>3.345**</td>
<td>1.208**</td>
</tr>
<tr>
<td>Same KudoRank</td>
<td>1.488**</td>
<td>0.398**</td>
</tr>
<tr>
<td>Past OSS project(s)</td>
<td>3354.612**</td>
<td>8.118**</td>
</tr>
<tr>
<td>Mutual Acquaintance in the</td>
<td>26.288**</td>
<td>3.269**</td>
</tr>
<tr>
<td>Evaluation Network</td>
<td>2.74e-10</td>
<td>-22.01</td>
</tr>
</tbody>
</table>

* $p < 0.05$   ** $p < 0.01$   *** the independent variable is dropped because of collinearity

Table 4 shows the odds ratios and coefficients for each independent variables from the CLM analysis on both networks. For the evaluation network, the homophily in country, location, programming language, KudoRank score are found to be significant determinants with odds ratios larger than 1. In addition, two shared affiliations – past OSS projects and mutual acquaintance in the evaluation network – are also found to be determinants and have larger odds ratios than other determinants.

On the other hand, in the analysis of the collaboration network, mutual acquaintance for the evaluation network and collaboration network variables are dropped from the CLM due to collinearity. Seven independent variables are found to be statistically significant. Four of them – same location, same country, same programming language, and
past OSS projects – have odds ratios that are larger than 1.01. Therefore they are found to be significant determinants of link formation in the Ohloh collaboration network. The results for the evaluation network are discussed in the following section.

Discussion

Determinants of Link Formation

For both Ohloh evaluation and collaboration networks, the CLM regressions found that the homophily in location, programming language, KudoRank, and past OSS projects are significant determinates of link formation. Those determinants imply that two previously unconnected developers are likely to evaluate or collaborate with each other if they have lived in the same city, have mainly used the same programming language, have had similar community reputation, or have worked in the same OSS project before.

These findings may partly be explained by the following conjectures. Geographic propinquity indicates that such developers may have more opportunity to meet each other in person and form stronger personal relationships. Consequently this may increase the likelihood of future collaborations and then evaluations. Moreover, an OSS project usually requires a primary programming language. Therefore, developers may take that into their considerations for choosing collaborators. In addition, developers with similar OSS community reputation may be at the same stage of their OSS activities. They may have more common experiences which bring them together for collaboration or evaluation. At last, it is not surprising that past OSS collaboration experience is also a determinant, considering 93.1% of the evaluation links have corresponding past OSS collaboration links.

Homophily in the same country is found to be a significant determinant only for the evaluation network but not for the collaboration network. This may be caused by the personal nature of the evaluation network. A Kudo evaluation link means a developer explicitly select another one for positive evaluation. This indicates the Kudo sender must know the receiver in some depth. However, a collaboration link only indicates two developers have worked in the same OSS project but not necessarily know each other in person. In addition, global collaborations in OSS development are becoming more popular due to the emergence of Internet. Therefore, homophily in nationality as a personal attribute may have more weight in influencing link formation in the evaluation network than the collaboration network.

Another significant determinant – mutual acquaintance in evaluation network – implies that two previously unconnected individuals are likely to evaluate each other with one or more shared acquaintances. This determinant has been well studied in SNA research (Kossinets et al. 2006; McPherson et al. 2001). It was found that individuals tend to select new acquaintances who are friends of a friend. In the evaluation network under study, this suggests that OSS developers tend to have circles of trust that include close acquaintances. They are likely to form operational cliques which enhance communication within the network and increase the capacity to act. This is also in line with the social closure theory (Coleman 1990) which suggests that the greatest value is obtained from networks that are densely connected with a high level of trust among actors.

One surprising finding is that all accumulative advantage attributes of OSS developers – OSS and Coding experiences and OSS developer (Kudo Receiver) degree – have no effect on the link formations in both the Ohloh evaluation and the collaboration network. Therefore, the scale-free topologies of both networks found in topological analysis cannot be simply explained by the preferential attachment principle. More in-depth analysis is needed to discover the alternative mechanisms in Ohloh networks which account for their scale-free topologies.

Utilizing Network Semantics in the Design of Expert Recommendation Systems

User-based Link Prediction Mechanism

Our analysis presents a novel approach to discover semantics of social networks and quantify their impact on network link formation. Then the second research naturally arises – how to use the discovered semantics in the design of expert recommendation systems. To answer this question, we use a computational mechanism for expert recommendation system to predict users’ positive evaluation choices based on our semantic analysis.
As Figure 3 shows, the proposed mechanism consists of three steps: data processing, conditional logistic analysis, and expert ranking. Firstly, the semantic information – homophily and shared affiliations - is extracted by matching users’ profiles with candidate experts’ information. Secondly, the conditional logistic model is used to calculate the probability for a user to positively evaluate a candidate expert based on the discovered network semantics. For example, as Table 4 shows, in the Ohloh evaluation network, the coefficient for homophily in location (city) is 2.199 and coefficient for using the same programming language is 1.208. If two developers a and b both live in New York City and use Java as their primary programming language (without any other homophily and shared affiliations), the probability for a to positively evaluate b from an alternative set can be calculated as:

\[ \Pr(y_a = b) = \exp(1 \times 2.199 + 1 \times 1.208) / \sum_i \exp(X_i \beta) \]

This calculation can be applied to any pair of members in Ohloh community. Then a list of all candidate experts can be ranked based on this link prediction probability for each user. The more likely a candidate expert is positively evaluated by the user, the higher this expert is ranked in the recommendation list.

This mechanism can quantitatively reflect individual users’ evaluative opinions on each recommended expert. By embedding this user-based mechanism into the design of expert recommendation systems, the recommended experts should be more acceptable for users.

**Support Top-N most recognized mechanism with Semantic (Contextual) Information**

While the link prediction mechanism has reflected users’ perspectives, our analysis can also enhance expert recommendation by providing semantic (contextual) information from the network perspective. A very common mechanism in SNA based expert recommendation is to measure the number of links (i.e. degree) an expert receives in the network. The underlying assumption is that the more links an expert receives, the more recognized and popular this person is. This mechanism reflects an aggregated opinion on the candidate expert by other members in this community. In our paper, we refer it as the *Top-N most recognized* mechanism.

However, the experts identified by Top-N most recognized mechanism do not always meet users’ needs in terms of required expertise. This is mainly because that social network analysis on such datasets mainly relies on collaboration and communication links (e.g., emails), and largely ignore network semantics. For instance, a software developer is looking for an expert to answer his questions about java programming. However, the topological analysis of his email network only identifies the individual who receives most emails but cannot tell why.

Our analysis can address this problem by integrating the discovered network semantics with corresponding links to construct a semantic social network. Figure 4 shows an example of using semantic social networks for expert recommendation. It uses Ohloh community and the results from our semantic analysis as the setting. We assume that there are only five developers with ID number ranging from 1 to 5 in the dataset. They have two types of relationships – collaborations and evaluations.
In this example, a user Josh is starting an OSS project using XML languages. He cannot run his first XML-based Java application and needs some help from others. He wants to use expert recommendation systems to find an expert who is most recognized by others for his 1) XML-related knowledge and 2) mutual acquaintance connections. The first criterion proves the recommended individual’s expertise in related technical area. The second one increases the probability of reaching such expert through mutual acquaintances and guarantees the expert himself is well connected to reach out for help.

There are three steps in this example. The first step shown on the left panel of Figure 4 presents the semantics identified from our analysis for both Ohloh collaboration network links and evaluation network links. For example, the identified semantic of the collaboration links is homophily in the programming language. Therefore, in the collaboration network, the semantic of the link between developer 1 and 5 is noted as [Java, XML], indicating they both use Java and XML as their programming languages. The second step shown in the middle involves aggregating the two types of links into one composite link and integrating all its identified semantics to this link. This composite semantic social network uncovers the structure of all types of relationships among the developers with their corresponding semantics.

At the third step, a sub semantic social network is extracted from the composite one and based on the semantics queried by the user Josh - homophily in using XML and mutual acquaintances. The recommended developer from the sub semantic network is developer 1 since it has the largest number of links with these two semantics.

Without the identified semantics of social networks, the structural analysis based expert recommendations will yield distinct results from the recommended developer. In the collaboration network, developer 5 is recommended since s/he has the largest number of collaboration links. For the evaluation network, developer 2 with the most evaluation links is selected. However, those recommendation results are far less accurate than the one from the sub semantic social network since they simply count the number of links without considering the semantic information required by the users. Therefore, semantic social networks may help provide more accurate expert recommendations to match users’ specific needs.

**Conclusion**

In this paper, we use both SNA topological analysis and conditional logistic analysis to examine the semantics (determinants) of link formation in two real-world OSS social networks. The results indicated that both networks have features of scale-free and small world topologies. We also found that the homophily in country, location, programming language, and KudoRank score were significant determinants for both networks under study.
Homophily in the same country and mutual acquaintance were found to be significant only for the evaluation network. We also explored the possible social causes and implications for the significance of various determinants. The set of methods for the semantic analysis of SNA used in our study may be applied to other types of networks.

In addition, from the design science perspective, we discussed how to embed network semantics into the design of expert recommendation systems through two mechanisms. The user-based link prediction mechanism is based on the conditional logistic model. It can quantitatively calculate the likelihood for a user to positively evaluate a recommended expert based on their homophily and shared affiliations. For the Top-N most recognized mechanism, we showed an example of constructing and utilizing semantic social networks to provide contextual information to meet users’ specific needs. Our analysis may help the researchers and practitioners in the design science community to better understand the semantics of social networks and devise various business applications.

Our future work consists of two directions including (1) integrating the user-based link prediction mechanism and semantic enhanced Top-N most recognized mechanism for better expert recommendation, and (2) implementing and evaluating a semantic social network based expert recommendation system. Our efforts will open a new venue of research in the expert recommendation systems and semantic social network analysis.

References


