Engineering Applications of Artificial Intelligence 55 (2016) 202–218

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Feature identification for predicting community evolution in dynamic social networks

Nagehan İlhan*, Şule Gündüz Öğüdücü

Faculty of Computer and Informatics, Istanbul Technical University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 12 December 2015
Received in revised form
25 May 2016
Accepted 6 June 2016

Keywords:
Dynamic networks
Community evolution
Feature selection

ABSTRACT

In parallel with the increasing popularity of commercial social-networking systems, the scales of such systems have grown notably, now with sizes ranging from hundreds of millions to more than a billion users. Besides being large, these systems also have a dynamic, temporal nature, with evolving structures. Thus, one of the main challenges is to understand and model the evolution of the meso-scale structures such as community structures within these networks. Most previous studies have concentrated on determining community events based on the community features extracted at different time points. However, both the huge volume of data and the dynamic structure of the networks hinder effective computation of these features. In this paper, we propose a novel framework that examines various structural features of the network and detects the most prominent subset of community features in order to predict the future direction of community evolution. Our approach is to extract the network structure and use it to determine the subset of community features that leads to accurate community event prediction. Unlike traditional approaches that harvest a large number of features at each time point, the proposed framework requires extraction of a minimal number of community features to effectively determine whether a community will remain stable or undergo certain events such as shrink, merge or split. Moreover, the extracted community features vary depending on the network structure, capturing network specific characteristics. Several experiments conducted on four publicly available datasets verified the effectiveness of the proposed framework.

1. Introduction

Social networks are made up of actors called nodes that are connected by various social familiarities or relationships represented by edges. A social network dynamically changes since the social ties between network actors change over time. The rapid advent in social networking systems has given rise to a growing need for Social Network Analysis (SNA) in order to investigate the relationships between network actors while being able to follow their evolution. In SNA, the main interest is to infer the structural characteristics of networks. Hence the connections between actors are key elements of the analysis that facilitates the mining of the important behavioral patterns among the actors.

Most of the networks involve parts that are more densely connected to each other than the rest of the network which are called groups, clusters or communities. The study of communities is a fundamental task in SNA since their structure and evolution is crucial in understanding the structure and dynamics of the networks as a whole. In the course of network evolution, different events may occur such that a node may join or leave a community, establish or drop a connection to another node, etc. Hence, the communities in these dynamic networks might grow or shrink and other events may occur over time (Newman and Park, 2003). By understanding the structural patterns of the network, we can derive the consequences of such events and predict the course of network evolution in different snapshots. Community evolution prediction is the prediction of possible events that a community might encounter during its lifetime such as growing, shrinking and merging with another community. Community evolution has a significant importance in numerous application domains. For instance, investigating the dynamics of communities that have genetic susceptibility to a disease is important in tracking the inception of an epidemic. A small alteration in a gene community may correspond to events such as gene fusion and may be crucial in foreseeing the genetic properties of the next viral strain. Similarly, the user comments and friendship ties within a social networking site can be used to follow emergence and development of new ideas and political views. Knowledge about the future of a community may also help to define the possible new members of the community.

* Corresponding author.
E-mail addresses: nilhan@itu.edu.tr (N. İlhan), sgunduz@itu.edu.tr (Ş.G. Öğüdücü).

http://dx.doi.org/10.1016/j.engappai.2016.06.003
0952-1976/

The study of social network evolution has received considerable attention from many researchers (Asur et al., 2007; Backstrom et al., 2006; Huang and Lee, 2011; Leskovec et al., 2005; Kumar et al., 2006; Palla et al., 2007; Takaffoli et al., 2014). The most widespread approach to modeling the structural changes in dynamic networks is to convert an evolving network into an ordered sequence of static network snapshots, each representing the state of the network at a given point in time (Backstrom et al., 2006; Leskovec et al., 2005; Berger-wolf, 2006; Tantipathananandh et al., 2007). A number of researchers characterize the evolution of a given community by describing its life-cycle, i.e. a series of critical events undergone by communities over time (Palla et al., 2007; Takaffoli et al., 2014; Bródka et al., 2012; Greene et al., 2010). Bródka et al. (2012) proposed a framework for modeling community evolution and event prediction by utilizing node numbers for the communities. Huang and Lee (2011) proposed an approach of incorporating activity features in measuring the influence of member activities to predict the network evolution. In our previous study (Ilhan and Öğüdücü, 2013), we proposed a framework for tracking the evolutionary dynamics of communities in social networks. The framework included a feature extraction component identifying the multiple structural features of communities. The experiments in that study have shown that the extracted features yield more accurate community event prediction. However, extracting a wide range of features is computationally expensive, especially when working with large datasets. In such cases, it is crucial to discard redundant and ineffective features.

Feature selection is one of the most frequently used techniques to choose an appropriate subset of features in order to reduce the computational complexity and increase the prediction accuracy. In the literature, there exist several studies that apply feature selection as a pre-processing step of classification, however this is merely for removing unnecessary attributes and preventing disturbances (Huang and Lee, 2011; Takaffoli et al., 2014; Saganowski et al., 2015). Even when employing feature selection methods, the entire set of features must first be the calculation cost that remains unchanged. In order to overcome this problem, we focus on determining the subset of features that can deliver an optimal performance for community event prediction without need for calculating all features beforehand. We found that the subset of community features yielding better predictions is correlated with the structural characteristics of the network. Each network has its own characteristics and topological properties. Thus, good results should not be expected with the same subset of community features in all networks. Instead, we propose that a distinct set of community features is prominent in different networks. The question comes to mind, which characteristics of the networks have most effect upon the selected feature subsets?

In this study, we propose a novel framework named Feature Identification for Event Prediction (FIEP) to identify a proper subset of features for a given network that achieves good prediction results without the need for calculating all features at the beginning. The suggested framework utilizes various structural network measures including clustering coefficient, average path length, embeddedness and betweenness in order to determine the accurate subset of features. The contribution of this paper is threefold. First, the proposed generic methodology for predicting the community evolution facilitates the identification of the predictive community features based on the structure of networks. The community event prediction is then modeled as a classification problem. Second, our methodology is capable of determining a useful subset of community features at the first observation moment of the network without observing the dynamic behavior of the network at different time periods. Third, we have empirically tested different factors related to the network structure and community features that may contribute positively to the community event prediction performance.

The rest of the paper is organized as follows: we survey the related work in Section 2. The FIEP framework appears in Section 3. Section 4 gives the experimental setup and results. Finally, Section 5 concludes the paper.

2. Related work

In this section, we present the related work on tracking community evolution, community event prediction, general feature selection and our contribution within the context of network analysis.

2.1. Tracking community evolution

Recently, many researchers have been interested in mining the temporal evolution of social networks (Asur et al., 2007; Backstrom et al., 2006; Leskovec et al., 2005; Kumar et al., 2006; Palla et al., 2007; Takaffoli et al., 2014; Berger-wolf, 2006; Tantipathananandh et al., 2007; Fagnan et al., 2014). A common way to study temporal network behavior is taking static snapshots and analyzing the structural characteristics of the static networks within each snapshot.

Several studies have been carried out on constructing an event prediction framework that characterizes the evolution of communities in dynamic networks. Palla et al. proposed an extension of the Clique Percolation Method (CPM) (Palla et al., 2005) to identify events such as birth, growth and merging in the evolution of dynamic graphs (Palla et al., 2007). This extension involved applying CPM on a graph formed by the communities discovered at pairs of consecutive snapshots. The resulting clique based communities were subsequently matched to communities and events pertaining to the communities specified. Asur et al. (2009) define critical events between detected communities at two consecutive snapshots which are implemented in the form of bit operations. However, these events do not cover all of the transitions that may occur for a particular community. Wang proposed an intuitive method to compare two communities of the consecutive timestamps with rules based on tracking specific core nodes that are more representative of their community than others (Wang, 2008). Greene et al. (2010) allowed for tracking of similar communities in different snapshots. They proposed a model for tracking the evolution of communities over time in a dynamic network, where each community is characterized by a series of significant evolutionary events. Their model introduces an effective community-matching strategy for efficiently identifying and tracking dynamic communities in multiple snapshots of a dynamic network. Chen et al. (2012) presented an approach to discover all possible types of community-based anomalies in evolutionary networks characterized by overlapping communities. Tajeuna et al. proposed a novel approach for modeling and detecting the evolution of communities. Their model comprises a new similarity measure, named mutual transition, for tracking the communities and rules for capturing significant transition events a community can undergo (Tajeuna et al., 2015). Leskovec et al. (2005) studied the patterns of graph evolution based on the various properties of the large social networks such as the degree distribution and the small-world phenomena. They also propose Forest Fire model to produce networks satisfying the discovered patterns. Ahn et al. (2007) analyzed different behavior scaling in degree distribution on online social networks, extracting the main characteristics of online social networks and performing an analysis of the evolution of Cyworld network. However, in these studies the influence of structural properties is examined at individual level and the prediction is lacking. A prediction which discards structural properties of communities may be insufficient when predicting several different events.