

Classifying Types of Network Communities Using Motifs

Kun Tu¹, Jian Li¹, Don Towsley¹, Dave Braines² and Liam Turner³
¹University of Massachusetts Amherst, ²IBM UK, ³Cardiff University

Abstract

One of the most significant tasks when studying network topology is the identification of network type of community. A network type is defined as the type of relation or interaction between nodes, including social ties in social networks, interactions between individuals in communication networks (e.g. email). Fundamentally, communities, which are groups of densely connected nodes in the network, enable us to discover clusters of interacting nodes and the relations between them. Communities from the same network type tend to have similar structures. Identifying their network types further allows us to study interactions between nodes, infer and predict unobserved network structures. Given a network topological structure, identifying the network type of a community can be viewed as a (sub)graph classification problem. Many existing methods use different graph embedding techniques to represent graphs in vector space and apply machine learning methods for classification. Yet, little work has applied network motifs, which capture the local structure of a network in terms of patterns that occur significantly more or less frequently, into network type identification. To the best of our knowledge, we are the first to use network motifs in classifying network type of community. In this paper, we find that there exists a strong relation between network type and motif distribution and subgraph ratio profile (SRP).

We first propose a generic framework to construct vectors for *feature representations* of static directed graphs from their topological structure using motif distribution and SRP. We argue that this fixed length feature representation can be used to classify and compare communities of varying sizes with high accuracy. We use three different null models, namely (i) **NM-1**: random graphs with the same number of nodes and edges; (ii) **NM-2**: random graphs with the same number of dyads; and (iii) **NM-3**: random graphs with the same in/out degree-pair sequence, to compute SRPs for all 16 triads in an empirical (sub)graph. For each null model, a 16 element vector containing SRPs of the corresponding 16 triads is computed to represent graph features. In order to evaluate our proposed graph feature representations, we collect 5108 communities from 955 real-world networks in 15 network types, including Google Plus and Twitter in social networks, high energy physics theory citation networks, Gnutella P2P networks and so on. We apply various well-known machine learning models along with our graph feature representation for network type of community classification. Our results achieve $90.73 \pm 2.88\%$ accuracy compared to $76.89 \pm 3.00\%$ from *struct2vec*, a state-of-the-art method that constructs graph representation using node attributes. In addition, we find that motif SRPs computed from **NM-1** achieves the best performance compared to the other two null models in network type of community classification.