

**A Probabilistic Framework for Structural Analysis and Community Detection in Directed Networks**

**Abstract:**

There is growing interest in structural analysis of *directed* networks. Two major points that need to be addressed are: 1) a formal and precise definition of the graph clustering and community detection problem in directed networks and 2) algorithm design and evaluation of community detection algorithms in directed networks.Motivated by these, we develop a probabilistic framework for structural analysis and community detection in *directed* networks based on our previous work in *undirected* networks. By relaxing the assumption from *symmetric* bivariate distributions in our previous work to bivariate distributions that have the same marginal distributions in this paper, we can still formally define various notions for structural analysis in directed networks, including centrality, relative centrality, community, and modularity. We also extend three commonly used community detection algorithms in *undirected* networks to *directed* networks: the hierarchical agglomerative algorithm, the partitional algorithm, and the fast unfolding algorithm. These are made possible by two modularity preserving and sparsity preserving transformations. In conjunction with the probabilistic framework, we show these three algorithms converge in a finite number of steps. In particular, we show that the partitional algorithm is a linear time algorithm for large sparse graphs. Moreover, the outputs of the hierarchical agglomerative algorithm and the fast unfolding algorithm are guaranteed to be *communities*. These three algorithms can also be extended to general bivariate distributions with some minor modifications. We also conduct various experiments by using two sampling methods in directed networks: 1) PageRank and 2) random walks with self-loops and backward jumps.

**Existing System:**

The two marginal distributions of the bivariate distribution, denoted by *pV* (*・*) and *pW*(*・*), are the same and they represent the probability that a particular node is selected in the sampled graph. As such, the marginal distribution *pV* (*v*) can be used for defining the *centrality* of a node *v* as it represents the probability that node *v* is selected. The relative centrality of a set of nodes *S*1 with respect

to another set of nodes *S*2 is then defined as the *conditional* probability that one node of the selected pair of two nodes is in the set *S*1 given that the other node is in the set *S*2.

Based on the probabilistic definitions of centrality and relative centrality in the framework, the *community strength* for a set of nodes *S* is defined as the difference between its relative centrality with respect to itself and its centrality. Moreover, a set of nodes with a *nonnegative* community strength is called a *community*. Intuitively, if the bivariate distribution *p*(*v,w*) is the probability for *v* and *w* to appear respectively at the two ends of a randomly selected path, then a community is a set of nodes with the property that it is more likely to find the other end in the same community given one of the two ends in a randomly selected path is already in the community.

In the probabilistic framework, the *modularity* for a partition of a sampled graph is defined as the average community strength of the community. As such, a high modularity for a partition of a graph implies that there are communities with strong community strengths. It was further that the Newman modularity and the stability are special cases of the modularity for certain sampled graphs.

**Proposed System:**

We show two modularity preserving transformations: (i) the transformation from a sampled graph that has the same marginal distributions to another sampled graph that has a symmetric bivariate distribution, (ii) the transformation from a sampled graph with a larger number of nodes to another sampled graph with a smaller number of nodes by node aggregation. These two transformations not only preserve modularity but also preserve sparsity (of the sampled graph). As such, the computational complexity of these three algorithms for directed networks remains the same as that of their undirected counterparts. In conjunction with the probabilistic framework, we show these three algorithms converge in a finite number of steps. In particular, we show that the partitional algorithm is a linear time algorithm for large sparse graphs. Moreover, the outputs of the hierarchical agglomerative algorithm and the fast unfolding algorithm are guaranteed to be *communities*. Though there are several variants of fast unfolding algorithms in the literature, it seems (to the best our knowledge) that our fast unfolding algorithm is the first one that has provable guarantees and also is general enough to be applicable to directed networks.