Evaluation of Community Similarity based on Hierarchical Distance

Ricky Laishram Sucheta Soundarajan
Syracuse University, Syracuse, NY, {rlaishra,susounda}@syr.edu

Abstract

In network analysis, there are a number of techniques for calculating the similarity between two sets of communities, such as Jaccard Similarity, Mutual Information etc. are used. However these measures do not account for the "closeness" of the different communities, and as result, they can be misleading. In this paper, we examine this problem and propose a method of computing the community quality based on the distances in hierarchical community.

1 Introduction

A common task in network analysis is community comparison. For example, if two community detection algorithms identify two different sets of communities, how similar are those results? Common community similarity metrics include Jaccard Similarity, Mutual Information, etc. However, when considering hierarchical community structure, these measures do not account for the "closeness" of the different communities, and so can be misleading. In this paper, we examine this problem and propose a method of computing community quality based on the hierarchical distance.

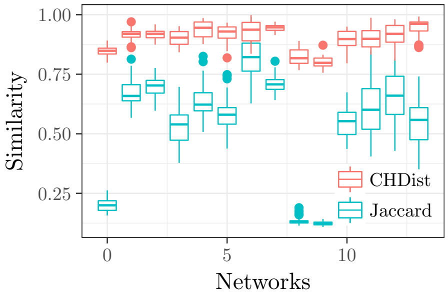

Figure 1: Similarity comparison for different real-world networks.

For example, consider the results shown in Figure 1, which depict the similarities between communities found by multiple runs of the Louvain modularity maximization algorithm on the same graph, across 14 real-world graphs. If we compare these communities using Jaccard similarity, we see that the community similarity may be very low. The Louvain method is non-deterministic, and so some variation in results is expected, but these results are shockingly low. Results for other standard measures are not shown because of space limitations, but are similar. We postulate that this occurs not because the detected communities are actually so dissimilar, but because the comparison metrics fail to take into account the hierarchical structure of the communities.

2 Community Hierarchical Distance

To address the problem described in Section 1, we introduce the *Community Hierarchical Distance (CHDist)*. Suppose that $\mathbb{C}$ and $\mathbb{C}'$ are two sets of communities in a graph G. The idea behind CHDist is that if a node u is in community $C \in \mathbb{C}$, but in a different community $C' \in \mathbb{C}'$, the penalty for this should be based on the change in modularity if C and C' are merged. (Other measures of community quality can also be used in place of modularity.)

Let $\mathcal{H}_{\mathbb{C},G}$ be the hierarchy of communities in G with the elements of $\mathbb{C}$ as the leaves. For $C_0, C_1 \in \mathbb{C}$, let $\eta\left(C_0, C_1, \mathcal{H}_{\mathbb{C},G}\right)$ be the normalized height of the smallest $C_\cup \in \mathcal{H}_{\mathbb{C},G}$ such that $C_0 \cup C_1 \subset C_\cup$. For a node u, let $\gamma(u, \mathbb{C})$ be the community $C \in \mathbb{C}$ such that $u \in C$. Let V_G denotes the set of all nodes in G.

For $C \in \mathbb{C}$, let us define, $\beta\left(C, \mathbb{C}'\right) = \underset{C' \in \mathbb{C}'}{argmax} |C \cap C'|$. Then,

$$\delta_H(\mathbb{C}, \mathbb{C}') = \frac{1}{|V_G|} \sum_{u \in V_{G_0}} \eta\left(\gamma\left(u, \mathbb{C}'\right), \beta\left(\gamma\left(u, \mathbb{C}\right), \mathbb{C}'\right), \mathbb{C}'\right)$$

Then we define the Community Hierarchical Distance between $\mathbb{C}$ and sample $\mathbb{C}'$ as the harmonic mean of $\delta_H(\mathbb{C}', \mathbb{C})$ and $\delta_H(\mathbb{C}, \mathbb{C}')$. As seen in Figure 1, the community hierarchical distance (denoted in red color) is much closer to the best theoretical value of 1.0 for all the networks considered.