# Differences Detection Between Two Groups of Networks

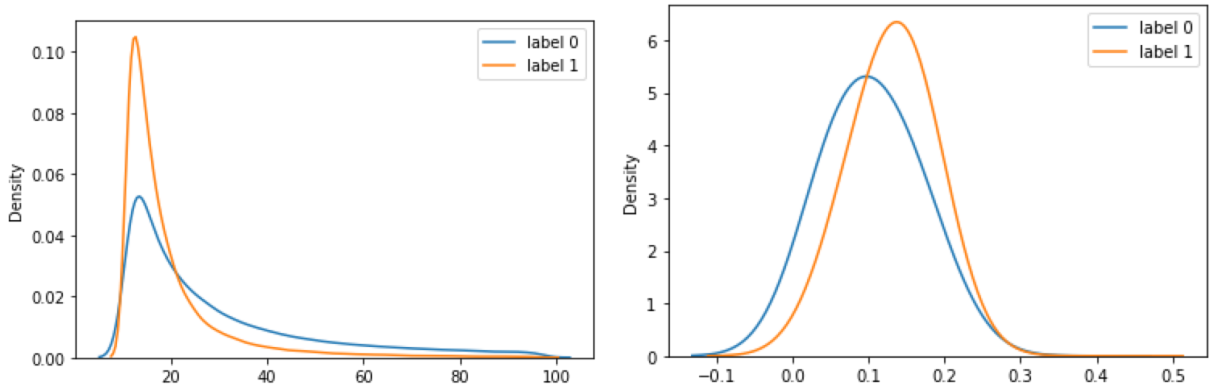
## 1 Dataset Introduction

Our dataset is a reddit thread networks dataset from Stanford Network Analysis Platform (SNAP). The dataset contains 203,088 undirected networks and each network is labeled 0 or 1 based on whether the corresponding thread is discussion based or non-discussion based. In table 1, we present the basic network statistics information of our datasets.

| Stats | Min | Max |
|---|---|---|
| #Nodes | 11 | 97 |
| Density | 0.021 | 0.382 |
| Diameter | 2 | 27 |

Table 1: Network statistics information

We are interested about what is the difference between these two types of networks. In figure 1, we present the number of nodes distribution and density distribution of networks with different labels. It is shown that there are no significant differences between networks with different labels. So we need to come up with another way to distinguish these networks. The past researches have shown that network subgraph density vector is a good way of representing the structure of network and also a good method to distinguish with different types of networks. So we turn to test whether these two types of networks are different in terms of the subgraph density vector.




(a) Number of nodes of networks

(b) Density of networks

Figure 1: Description of networks by labels

## 2 Question Setup

Consider two groups of networks

$$\mathcal{G} = \{G_1, \ldots, G_n\} \sim F_1, \quad \mathcal{H} = \{H_1, \ldots, H_m\} \sim F_2$$

We would like to test the null hypothesis of $H_0 : F_1 = F_2$ versus the alternative $H_1 : F_1 \neq F_2$. Let $\mathfrak{G}$ be the space of graphs. We consider the motif function $m : \mathfrak{G} \mapsto \mathbb{R}^d$, where for any graph $G \in \mathfrak{G}$, the vector $m(G) = (m_1(G), \ldots, m_d(G))$ denotes density of subgraphs of a given type.

We consider the following test statistic:

$$\rho(\mathcal{G}, \mathcal{H}) = \|\frac{1}{n}\sum_{i=1}^{n} m(G_i) - \frac{1}{m}\sum_{j=1}^{m} m(H_j)\|^2. \tag{1}$$

Let $\mu$ and $\Sigma$ be the mean and covariance matrix of $m(G)$ when $G \sim F_1$. Under $H_0 : F_1 = F_2$, $\mu$ and $\Sigma$ are also mean and covariance matrix of $m(H)$. Let

$$\widehat{\mu}_{\mathcal{G}} := \frac{1}{|\mathcal{G}|}\sum_{G\in\mathcal{G}} m(G), \quad \widehat{\mu}_{\mathcal{H}} := \frac{1}{|\mathcal{H}|}\sum_{H\in\mathcal{H}} m(H)$$

Under null, by the CLT, $\sqrt{n}(\widehat{\mu}_{\mathcal{G}} - \mu) \rightsquigarrow N(0, \Sigma)$ and similarly $\sqrt{m}(\widehat{\mu}_{\mathcal{H}} - \mu) \rightsquigarrow N(0, \Sigma)$ as $n, m \to \infty$. By independence of $\widehat{\mu}_{\mathcal{G}}$ and $\widehat{\mu}_{\mathcal{H}}$, we have

$$(\widehat{\mu}_{\mathcal{G}} - \widehat{\mu}_{\mathcal{H}}) \rightsquigarrow N(0, (\frac{1}{n} + \frac{1}{m})\Sigma)$$

Based on some calculation, we have a test statistic

$$\tilde{\rho}(\mathcal{G}, \mathcal{H}) := \frac{nm}{n+m}\|\widehat{\Sigma}^{-1/2}(\widehat{\mu}_{\mathcal{G}} - \widehat{\mu}_{\mathcal{H}})\|^2. \tag{2}$$

Under null $\tilde{\rho}(\mathcal{G}, \mathcal{H}) \rightsquigarrow \chi^2_d$. Let $\chi^2_d(1-\alpha)$ be the $(1-\alpha)$th quantile of $\chi^2_d$. Then, the test that rejects when

$$\tilde{\rho}(\mathcal{G}, \mathcal{H}) > \chi^2_d(1-\alpha)$$

will have asymptotic level $\alpha$.

## 3 Statistical Computation

In the following simulations, for the sake of simplicity, we only consider three node connecting subgraphs. There are only two types of three nodes subgraphs so $d = 2$ and we present them in fig 2.

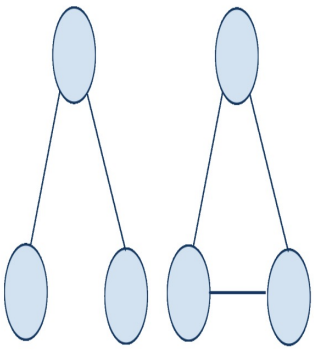


Figure 2: Two types of three node subgraph

So our statistical computation will be: for each network, we compute its number of above two types of subgraphs. We then calculate the density of each type of subgraph i.e. the number of one type of subgraph over the total number of two types of subgraphs. Here we give an example of how we calculate the subgraph number for one network. m3_1 and m3_2 are two types of subgraphs. We loop for all the 3 node combination in the network and see if they match our target subgraph.

```
m3_1=0;m3_2=0
for sub_nodes in itertools.combinations(G.nodes(),3):
# choose 3 from number of nodes n
    subg = G.subgraph(sub_nodes)
    if nx.is_connected(subg):
        if nx.is_isomorphic(subg, motifs['M3_1']):
            m3_1+=1
        if nx.is_isomorphic(subg, motifs['M3_2']):
            m3_2+=1
```

We find that on average, the time for completing above calculation for one network is about 0.3s. So with more than 200,000 networks, the total calculation time is more than 17 hours. So we decide to use CHTC to help us decrease the computation time.

We split 1000 parallel jobs and each job complete above calculation for about 200 networks. Now ideally, we could save the computation to only about 1 minute. In practice, it takes longer to wait for the computing resources. We set the required memory and disk space both 1GB for each job.

# 4 Conclusion

Based on the final calculation and above equation, we show that the p-value for our hypothesis is less than 0.0001. So that we reject the $H_0$ and conclude that two types of networks are different in structure in terms of the three node subgraph density.

However, even though two groups of networks are statistically different in terms of the 3-node subgraph density, they are not different in practice. In fact, for networks with label 0, the mean density vector is [0.9910 , 0.0090]. For network with label 1, the mean density vector is [0.9925 , 0.0075]. So it is hard to classify a network based only on the 3-node subgraph density. To solve this problem, we also explore the 4-node subgraph density. And we find a more practical significant results. There are in total 6 types of connected 4-node subgraph. The 4-node subgraph density for networks with label 0 is [0.798, 0.018, 0.101, 0.001, 0.001, 0] and for networks with label 1 is [0.676, 0.023, 0.267, 0.002, 0.001, 0]. We could find that there are more differences between 4-node subgraph density. And it will be easier for us to tell the differences between two groups of networks.

In this report, we only focus on the labels of the networks and the real meaning of the networks does not influence our analysis. We focus on telling the differences between two groups of networks.