Directed Network Comparison

Jiaqian Yu
Depratment of Statistics
New York, NY, 10025
jy2880@columbia.edu

Abstract

Network comparison is of both theoretical and practical value. In this report, we try to establish methods to compare the directed networks based on different features and different linkage methods. We apply all the methods to the social networks data including google, amazon and twitter. We use Adjusted Rand Index to measure the performance of our methods and find out that methods with triad census features and complete linkage perform best in our dataset. In addtiion, we apply a Monte Carlo test to see whether our methods are affected by sizes and densities. The null hypothesis is that our method does not perform well on real networks than random networks is rejected under the significance level 0.05.

1 Introduction

Networks are representation of complex interaction dataset. Many realistic dataset could be considered as network, such as protein-protein interacions, world trade flow and the popular social networks like twitter and facebook. Analysis of networks try to gain important structures from great amount of networks and network comparison is one of the key questions to be addressed.

Network comparison is addressed in different ways. In machine learning, graph kernels are used to obtain classifiers to predict the class membership of networks.

Other computational algorithms are based on network alignment which can be quite computer-intensive. Instead algorithms which compare counts of small subgraphs have become popular. Network comparison based on small subgraphs is motivated by the observation that many real networks contain some characteristic small subgraphs, sometimes called motifs[1], which relate to the function of the network.

In this report, network comparison is used to cluster networks. Our idea is to look at induced subgraphs of different networks and count the number of different features in order to cluster networks according to their feature vectors.

Then we apply this to the dataset with a set of 31 sparse directed social networks from [4] using different distance measures, different features and different linkage methods to compare the results. Based on Adjusted Rand Index, Euclidean measures with triad census and complete linkage perform best to our dataset, clustering these networks correctly. Then we perfrom a Monte Carlo test and find out that our method capture information beyond sizes and densities.

This report is structured as follows. Section 2 provides the background for directed networks. Section 3 introduces our idea and introduces the best method in details. Section 4 gives the clustering results for the real-world dataset, results given by performance measure and results for Monte Carlo test. Discussion and conclusion are provided in Section 5 and 6.

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

2 Background

In this section, we introduce some background for the directed network which is useful in the following sections.

2.1 Directed Network

A directed network is an ordered pair $G = (V, E)$ with a set V of nodes and a set E of directed edges. Edges are ordered tuples of V. For $u, v \in V$, $(u, v) \in E$ indicates that the network contains a directed edge from u to v.

A simple directed network doesn't contain any self-edges or any multiple edges in the same direction. Moreover, the node set V is finite.

A directed network $G_1 = G_1(V_1, E_1)$ is called a sub-network of the directed network $G = (V, E)$ if and only if $V_1 \subset V$ and $E_1 \subset E$. Moreover, if E_1 contains all the edges in E that have both endpoints in V_1, then G_1 is called an induced sub-network of G.

2.2 In/Out Degree

The in-degree of a node v in a network is defined as the number of edges (u, v) directed to v and the out-degree of a node v is the number of edges (v, u) begins from v. The in- or out-degree distribution $P(k)$ of a network is defined to be the fraction of nodes in the network with in- or out- degree k. The dyad census counts the edges with different orientations in the network which caputers information from in- or out-degree distribution.

2.3 Triad

A triad is a directed network composed of 3 nodes[2]. According to the possibilities of different connections among the nodes, traids have 16 types. The concept of triad is widely used in social network analysis. It is shown that triad counts in social networks often capture information beyond the dyad census.

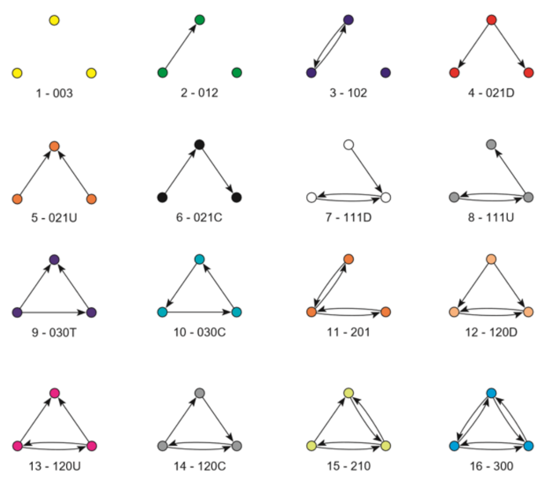


Figure 1: 16 Different Triads[3]

3 Methodology

3.1 Idea

For the purpose of clustering networks, the idea in this report is to combine the distance measures between feature vectors with hierarchical clustering. Based on this idea, we can establish different methods by choosing different distance measures, different featrues for graphs and different linkage methods.

For the distance measures, in this report we just choose the simplest Euclidean distance. But it is possible to choose distances measures between two probability distributions such as Kullback-Leibler Divergence and 1st Wasserstein Metric, etc. For the features for graph, our idea is to choose features based on dyad census data such as in- or out-degree distribution or triad census data. And for the linkage methods in hierarchical clustering, we just try the most common ones: complete linkage and average linkage.

So, in total we have 6 different methods and try to get the best one for our real-word dataset.

3.2 Network Comparion Method

Here, we will introduce the method using Euclidean distance, triad census based features and complete linkage in details.

For any directed netrwork, the triad census searches all of its induced sub-network with three nodes and counts the number of different triads.

Consider a directed network G, first we construct its feature vector based on triad census. There are in total 16 different types of triads and for any three-nodes induced sub-graph, there will be exactly one match in the triad set consists of 16 triads. However, as is shown in Figure 1, type 1,2 and 3 of triads are not connected. So we just exclude triads 1,2 and 3, leaving the other 13 triads under consideration. Denote the triad set with 13 triads as

$$\theta : \{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}.$$

Count the number of induced sub-graphs with 3 nodes in G that match triad i in set θ as $n_i(G)$. Set

$$N(G) = \textstyle\sum_{i\in\theta} n_i(G)$$

as the sum of $n_i(G)$ and calculate the fraction of matches for each triad i:

$$p_i(G) = \tfrac{n_i(G)}{N(G)}.$$

If we denote P as $P(G) = (p_i(G), i \in \theta)$, P will map the space of every finite simple directed networke with high dimensions to a feature space $([0, 1] \cap \mathbb{Q})^{13}$ with only 13 dimensions.

Now we generate a feature vector $p_i(G), i \in \theta$ for every network and we can measure the distance between different networks G_1 and G_2 with Euclidean distance:

$$Dis(G_1, G_2) = \sqrt{\textstyle\sum_{i\in\theta}(p_i(G_1) - p_i(G_2))^2}$$

Note this Dis is a pseudo distance on the space of finite newtorks, as $Dis(G_1, G_2) = 0$ does not imply $G_1 = G_2$, we can consider it as a measure for dissimilarity. For finite number of directed networks $G_i, i \in \phi$, we can generate the dissimilarity matrix DM where the entries are the pairwise Dis dissimilarities:

$$DM_{i,j} = Dis(G_i, G_j)$$

Then, we can use this dissimilarity matrix DM for hierarchical clustering. Hierarchical clustering is an unsupervised clustering method based on the dissimilarities between each pair of items. Initially, each item is viewed as one cluster. Then, the closet two clusters merge into one cluster. For the distance between two clusters, we choose the maximum distance between any pair, one in each cluster. Then every time the closest two clusters merge into one cluster until there is only one cluster left. In this report we cluster the networks $G_i, i \in \phi$ according to a pre-set number of clusters and stop the process when the number of clusters is reached. We use a dendrogram to show the process for clustering.

4 Case Study

In this section, we apply our methods to the real-word directed social networks.

4.1 Dataset

We selected 31 simple directed networks of 5 different clusters from SNAP Dataset(Stanford Large Network Dataset Collection)[4]. The 5 different clusters are Amazon networks, Google networks, p2p-Gnutella networks, soc-sign-Slashdot networks and Twitter networks.

An overview of this dataset is found in Table 1.

Table 1: Overview of the dataset

Type	Nodes range	Density range	Number
Amazon	26000+ 410000+	1.79e-05 2.08e-05	4
Google	326 2213	1.91e-02 9.72e-02	7
p2p-Gnutella	6301 26518	9.30e-05 5.23e-04	6
soc-sign-Slashdot	77000+ 82000+	8.14e-05 8.63e-05	3
Twitter	9 242	4.37e-02 7.5e-01	11

4.2 Results

We map the networks to feature space using the map P in section 3.2 to obtain the dissimilarity matrix. (Similarly, we also choose in- and out-degree distribution as features to obtain two more dissimilarity matrice). Set the number of clusters to 5 and use Hierarchical Clustering with 2 different linkage methods, we can get the following dendrograms.

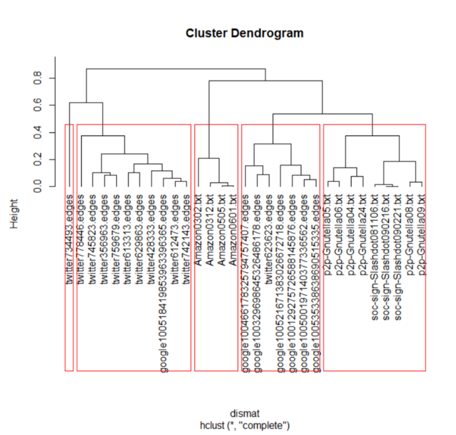


Figure 2: Triad census w. complete linkage

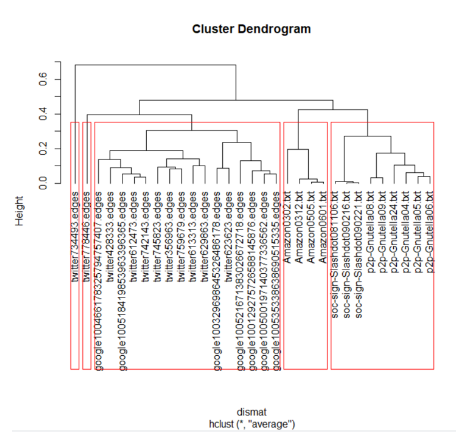


Figure 3: Triad census w. average linkage

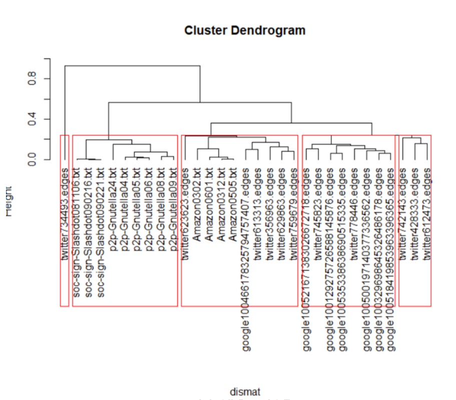


Figure 4: In-degree dist. w. complete linkage

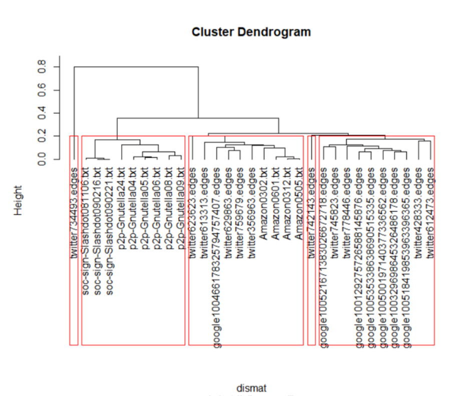
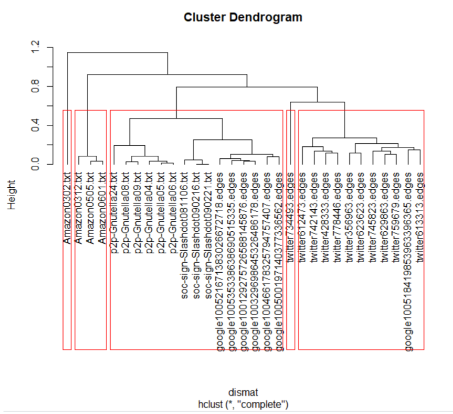


Figure 5: In-degree dist. w. average linkage



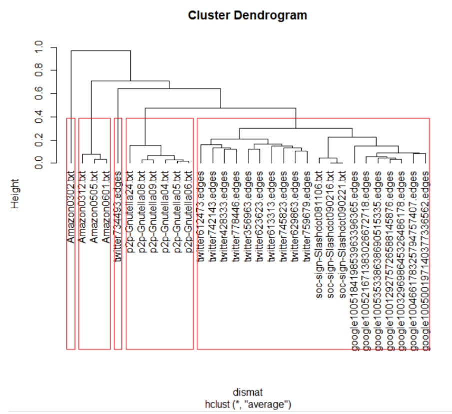


Figure 6: Out-degree dist. w. complete linkage Figure 7: Out-degree dist. w. average linkage

Intuitively, we can find out that Figure 2 seems to be the best which means triad census features with complege linkage is the best method for our dataset.

4.3 Performance Measure

We also want to assess our results quantitatively. In order to assess the outcome of our clustering method, we use one performance measure called Adjusted Rand Index(ARI)[5]. The Adjusted Rand Index is defined as:

$$AdjustedRandIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

The Adjusted Rand Index can compare two partitions and it has a value between -1 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the two clusters are exactly the same, and it yields negative values if the index is less than the expected index.

For our datset, we assume that newtorks from the same sourse belong to the same cluster and pre-assign the clusters for every networks.

Then we can use ARI to assess the similarity between the partition obtained through our clustering method and another partition given by our pre assignment.

Table 2 shows the ARI result for our different methods. We can see that triad census features with complege linkage is still the best method for our dataset based on ARI result. Its ARI is 0.64, indicating that this method can identify different types of networks .

Table 2: ARI result for different methods

Triad w. Complete linkage	0.64
Triad w. Average linkage	0.48
In-degree dist. w. Complete linkage	0.35
In-degree dist. w. Average linkage	0.32
Out-degree dist. w. Complete linkage	0.47
Out-degree dist. w. Average linkage	0.39

The detailed clustering result for this best method is shown below in Table 3.

4.4 Monte Carlo Test

While we find out the result for method with triad census features and complete linkage performs well based on ARI, as we can see from Table 1, these networks actually have very different sizes and densities. Here our question is whether the method can capture information beyond sizes and densities.

Table 3: Pre-assign and clustering result

Newtork	Pre-assign	Clustering
Amazon0302	1	1
Amazon0312	1	1
Amazon0505	1	1
Amazon0601	1	1
google10012	2	2
google10032	2	2
google10046	2	1
google10050	2	2
google10051	2	2
google10052	2	2
google10053	2	2
p2p-Gnutella04	3	3
p2p-Gnutella05	3	3
p2p-Gnutella06	3	3
p2p-Gnutella08	3	3
p2p-Gnutella09	3	3
p2p-Gnutella24	3	3
soc-sign-Slashdot081106	4	3
soc-sign-Slashdot090216	4	3
soc-sign-Slashdot090221	4	3
twitter356963	5	1
twitter428333	5	4
twitter612473	5	4
twitter613313	5	1
twitter623623	5	1
twitter629863	5	1
twitter734493	5	5
twitter742143	5	4
twitter745823	5	2
twitter759679	5	1
twitter778446	5	2

In our analysis, we take Adjusted Rand Index as the test statistics and apply a Monte Carlo Test[6]. Here the null hypothesis is the method preforms equally or better on random graphs with same densities and sizes than on our real dataset. The test procedure is as follows: We simulate N groups of independent random networks under the null hypothesis which means in each group, the networks share the same densities and sizes with our real dataset. Compute the ARI result for each group and get $Score_1, Score_2, ..., Score_N$. Denote the ARI result for our real dataset as $Score_0$. Since the higher ARI result indicates a better performance of our method, under significance level α, we reject the null hypothesis $\frac{m}{N+1} < \alpha$ where m denote the number of groups with higher ARI result than $Score_0$.

Here we generate 30 groups of random networks with same densities and sizes with our real dataset using ER model. Conduct the test described above, we find out the p-value is 0.03. So the null hypothesis is rejected under the significance leverl 0.05, indicating that our method captures information beyond sizes and densities of the networks.

5 Discussion

Further research may focus on more distance measures to see if the methods with other distance measures can perform better.

Also, there are lots of networks which are not simple or are signed networks from the dataset SNAP. So how to deal with those networks could also be addressed in the future.

6 Conclusion

In this report, we focus on comparison of directed networks. We try different methods for clustering the networks based on different features and different linkage methods. Then we apply our methods to the real dataset consisting of 31 social networks. We use ARI to measure the performance of our methods and find out that method with triad census features and complete linkage perform well on the data. We also apply a Monte Carlo test and find out that our methods can capture information beyond sizes and densities.

References

[1] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon. *Network motifs: simple building blocks of complex networks.* Science.2002,298:824-827

[2] S. Wasserman, K. Faust. *Social network analysis: Methods and applications.*

[3] http://mrvar.fdv.uni-lj.si/sola/info4/uvod/part4b.pdf

[4] http://snap.stanford.edu/data

[5] L. Hubert, P. Arabie. *Comparing partitions.* Journal of Classification. 1985,2 (1): 193–218

[6] M. Dwass. *Modified randomization tests for nonparametric hypotheses.* The Annals of Mathematical Statistics,1957:181-187

[7] X. Xu, G. Reinert. *Triad-based comparison*