CS529

Topic and Tools on Social Media Data Mining

(Assignment #2)

Link Prediction

Created by:-

Tech Pirates - **Kevin Savsani (160101063)**

- **Akul Agrawal (160101085)**

- **Aakash Agrawal (160122001)**

Datasets Introduction

Foursquare Restaurant Review Dataset

This dataset includes long-term (about 22 months from Apr. 2012 to Jan. 2014) global-scale check-in data collected from Foursquare, and also two snapshots of user social networks before and after the check-in data collection period. The check-in dataset contains 22,809,624 check-ins by 114,324 users on 3,820,891 venues. The social network data contains 363,704 (old) and 607,333 (new) friendships. Dataset contains four-column User ID (anonymized), Venue ID, UTC time, Timezone offset in minutes.

File dataset for friends contains Each row indicating a friendship between two users.

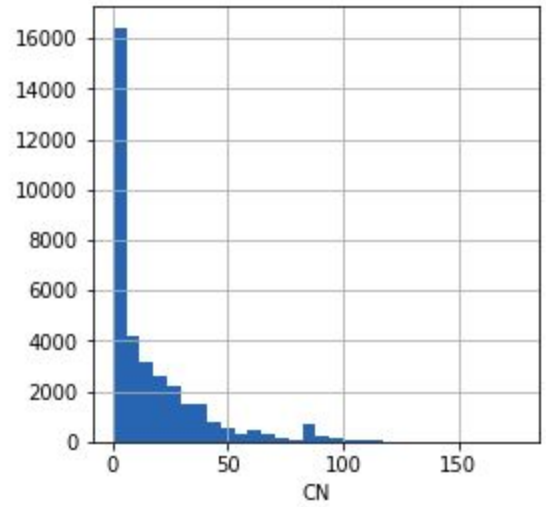
BlogCatalog data
This is the data set crawled from BlogCatalog (http://www.blogcatalog.com). BlogCatalog is a social blog directory website. This contains the friendship network crawled and group memberships. For easier understanding, all the contents are organized in CSV file format. It contains 10312 bloggers, 333,983 friendship pairs, 39 Groups.

DataSet Analysis

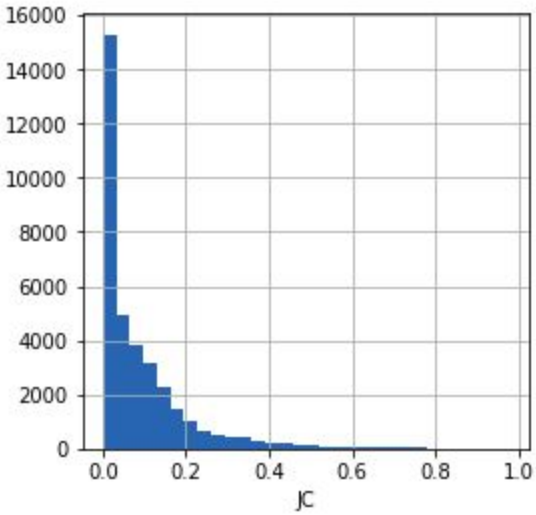
1. **Foursquare Restaurant Review Dataset**
   ● The graph is an undirected graph with 2060 nodes and 58810 edges and an average degree of nodes was 57.09.

   There were many Link measures that were calculated from this network.

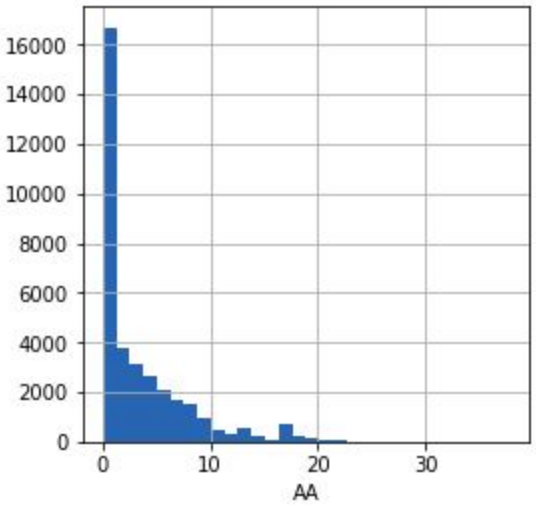
   Local Methods Include:
   Common Neighbour (CN)



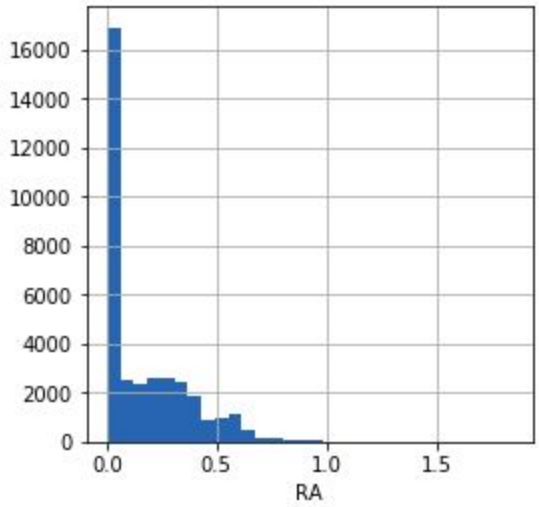
Jaccard Coefficient (JC)



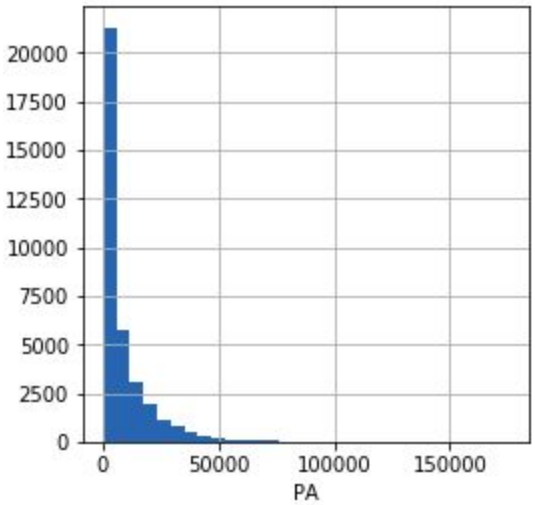
Adamic Adar (AA)

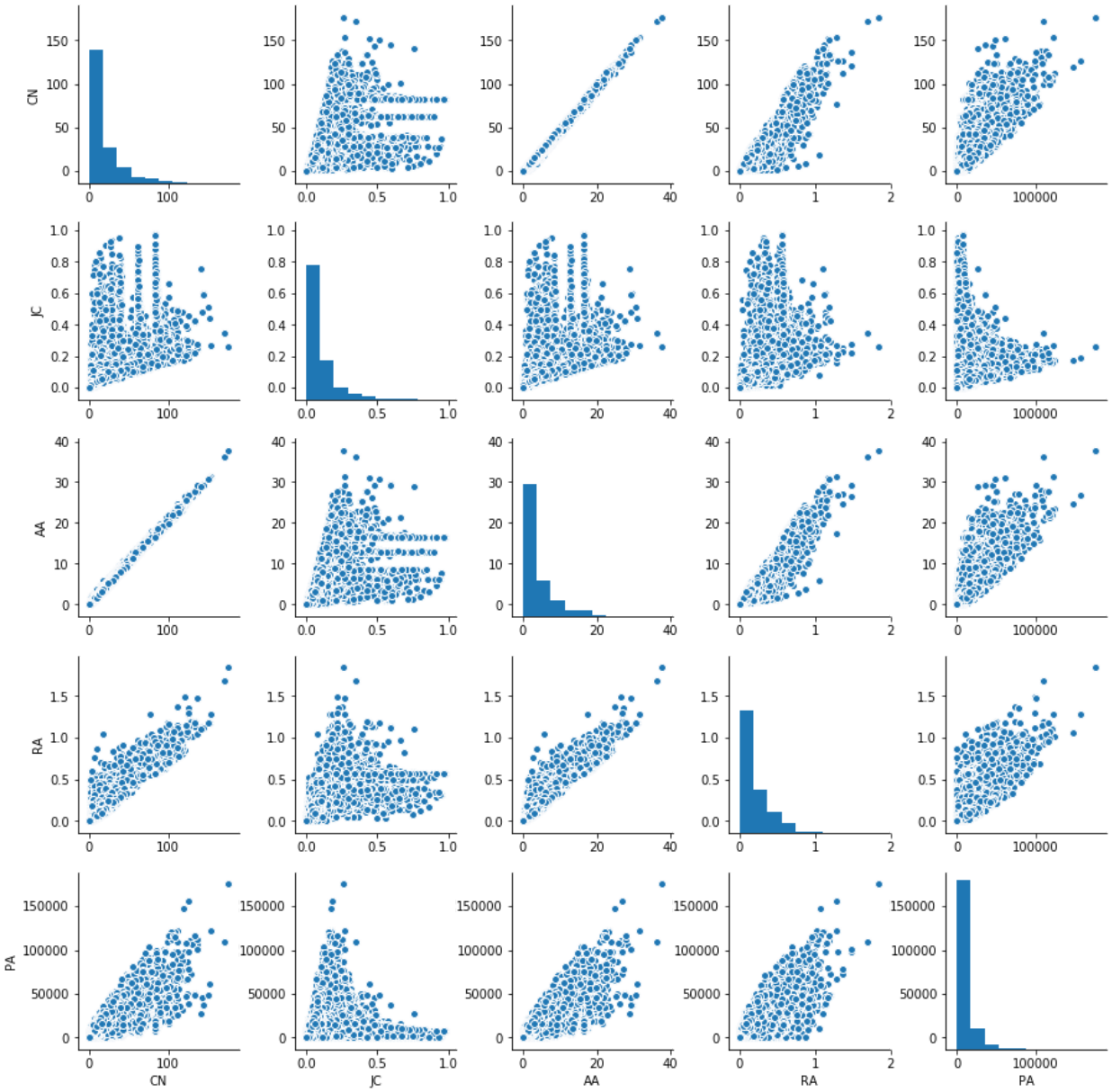


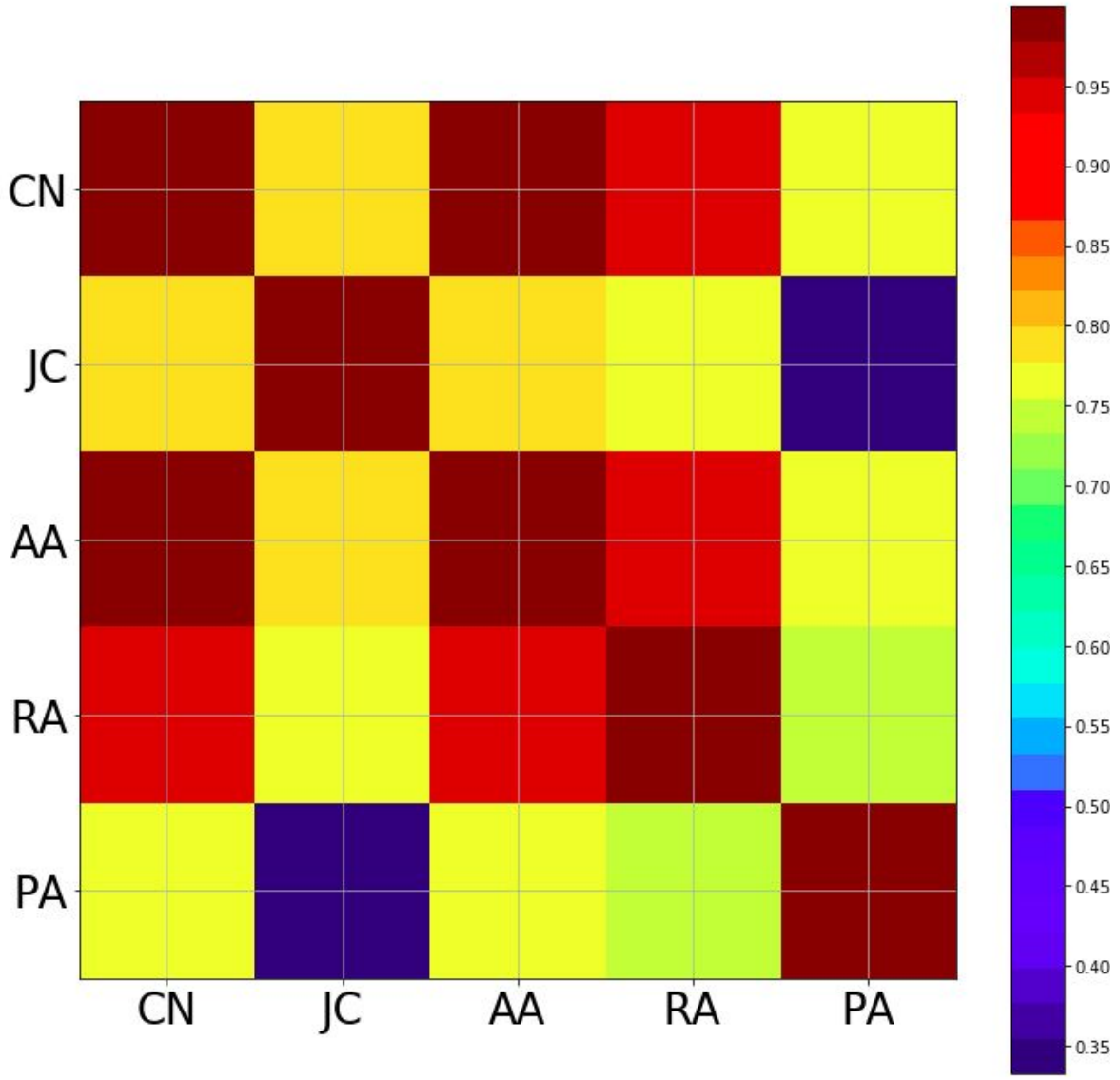
Resource Allocation(RA)



Preferential Attachment (PA)




CN
JC
AA
RA
PA
CN
JC
AA
RA
PA



	CN	JC	AA	RA	PA
CN	1.000000	0.783260	0.998514	0.937505	0.762457
JC	0.783260	1.000000	0.786932	0.759727	0.332490
AA	0.998514	0.786932	1.000000	0.954543	0.763387
RA	0.937505	0.759727	0.954543	1.000000	0.735900
PA	0.762457	0.332490	0.763387	0.735900	1.000000

['overall AUC', '-', 0.9648506426216747, 0.9773864096168762, 0.9715972588637749, 0.9841147381976568, 0.8469413901640535]
['PR@25%', '-', ['0.986', '0.488'], ['0.999', '0.500'], ['0.991', '0.495'], ['0.998', '0.499'], ['0.876', '0.438']]
['PR@50%', '-', ['0.893', '0.890'], ['0.919', '0.920'], ['0.913', '0.913'], ['0.939', '0.939'], ['0.766', '0.766']]
['PR@75%', '-', ['0.640', '0.993'], ['0.664', '0.996'], ['0.665', '0.997'], ['0.665', '0.997'], ['0.634', '0.950']]

MACHINE LEARNING PART

FOUR Square Dataset:

We used **5 fold Cross-Validation** to tune the parameters of the model. The results were as follows:

Naive Bayes

```
                GaussianNB

Accuracy:  0.9071975063757438

Classification Report
              precision    recall  f1-score   support

         0.0       0.86      0.97      0.91      3546
         1.0       0.97      0.84      0.90      3512

    accuracy                           0.91      7058
   macro avg       0.91      0.91      0.91      7058
weighted avg       0.91      0.91      0.91      7058


Confusion Matrix
[[3457   89]
 [ 566 2946]]

ROC AUC score
0.9068697830145166

                GaussianNB

Accuracy:  0.9105852345189174

Classification Report
              precision    recall  f1-score   support

         0.0       0.86      0.98      0.91      3463
         1.0       0.97      0.85      0.91      3594

    accuracy                           0.91      7057
   macro avg       0.92      0.91      0.91      7057
weighted avg       0.92      0.91      0.91      7057


Confusion Matrix
[[3383   80]
 [ 551 3043]]

ROC AUC score
0.9117937843915108
```

GaussianNB

Accuracy: 0.9074677625053139

Classification Report

	precision	recall	f1-score	support
0.0	0.86	0.98	0.91	3521
1.0	0.97	0.84	0.90	3536
accuracy			0.91	7057
macro avg	0.92	0.91	0.91	7057
weighted avg	0.92	0.91	0.91	7057

Confusion Matrix

```
[[3444   77]
 [ 576 2960]]
```

ROC AUC score

0.9076176425609241

GaussianNB

Accuracy: 0.9162533654527419

Classification Report

	precision	recall	f1-score	support
0.0	0.87	0.97	0.92	3552
1.0	0.97	0.86	0.91	3505
accuracy			0.92	7057
macro avg	0.92	0.92	0.92	7057
weighted avg	0.92	0.92	0.92	7057

Confusion Matrix

```
[[3456   96]
 [ 495 3010]]
```

ROC AUC score

0.9158730770713653

```
                GaussianNB

Accuracy:  0.9036417741249823

Classification Report
              precision    recall  f1-score   support

         0.0       0.86      0.97      0.91      3561
         1.0       0.97      0.83      0.90      3496

    accuracy                           0.90      7057
   macro avg       0.91      0.90      0.90      7057
weighted avg       0.91      0.90      0.90      7057


Confusion Matrix
[[3466   95]
 [ 585 2911]]

ROC AUC score
0.902994002211859
```

SVM

```
                SVM

Cross Validation Scores
[0.94786058 0.95041088 0.94573534 0.94316893 0.94756236]

Accuracy:  0.9436020972084456

Classification Report
              precision    recall  f1-score   support

         0.0       0.93      0.97      0.95      3561
         1.0       0.96      0.92      0.94      3496

    accuracy                           0.94      7057
   macro avg       0.94      0.94      0.94      7057
weighted avg       0.94      0.94      0.94      7057


Confusion Matrix
[[3442  119]
 [ 279 3217]]

ROC AUC score
0.9433884643387525
```

Decision Tree

```
                Decision Tree

Cross Validation Scores
[0.93525078 0.93893454 0.93610088 0.93253968 0.93579932]

Accuracy:  0.9326909451608332

Classification Report
              precision    recall  f1-score   support

         0.0       0.94      0.93      0.93      3599
         1.0       0.93      0.93      0.93      3458

    accuracy                           0.93      7057
   macro avg       0.93      0.93      0.93      7057
weighted avg       0.93      0.93      0.93      7057


Confusion Matrix
[[3354  245]
 [ 230 3228]]

ROC AUC score
0.9327065499686549
```

Then in order to improve our score, we removed some of the correlated features and also used Principal Components from **PCA** and **Singular Value Decomposition** as new features:
After PCA:

```
                GaussianNB

Accuracy:  0.8941626523094361

Classification Report
              precision    recall  f1-score   support

         0.0       0.83      0.98      0.90      3495
         1.0       0.98      0.81      0.89      3563

    accuracy                           0.89      7058
   macro avg       0.91      0.90      0.89      7058
weighted avg       0.91      0.89      0.89      7058


Confusion Matrix
[[3436   59]
 [ 688 2875]]

ROC AUC score
0.895011517596406
```

```
                GaussianNB

Accuracy:  0.8955646875442823

Classification Report
              precision    recall  f1-score   support

         0.0       0.84      0.98      0.91      3575
         1.0       0.98      0.81      0.88      3482

    accuracy                           0.90      7057
   macro avg       0.91      0.89      0.89      7057
weighted avg       0.91      0.90      0.89      7057


Confusion Matrix
[[3512   63]
 [ 674 2808]]

ROC AUC score
0.8944053534059277
```

GaussianNB

Accuracy: 0.888337820603656

Classification Report

	precision	recall	f1-score	support
0.0	0.82	0.98	0.90	3497
1.0	0.98	0.79	0.88	3560
accuracy			0.89	7057
macro avg	0.90	0.89	0.89	7057
weighted avg	0.90	0.89	0.89	7057

Confusion Matrix

```
[[3443   54]
 [ 734 2826]]
```

ROC AUC score

0.8891892087278662

```
                GaussianNB

Accuracy:  0.8893297435170753

Classification Report
              precision    recall  f1-score   support

         0.0       0.83      0.98      0.90      3571
         1.0       0.98      0.79      0.88      3486

    accuracy                           0.89      7057
   macro avg       0.90      0.89      0.89      7057
weighted avg       0.90      0.89      0.89      7057


Confusion Matrix
[[3508   63]
 [ 718 2768]]

ROC AUC score
0.8881955794534702

                GaussianNB

Accuracy:  0.9322658353407963

Classification Report
              precision    recall  f1-score   support

         0.0       0.91      0.96      0.93      3505
         1.0       0.95      0.91      0.93      3552

    accuracy                           0.93      7057
   macro avg       0.93      0.93      0.93      7057
weighted avg       0.93      0.93      0.93      7057


Confusion Matrix
[[3351  154]
 [ 324 3228]]

ROC AUC score
0.9324232756294097
```

Decision Tree

```
                Decision Tree

Cross Validation Scores
[0.93355058 0.9334089  0.93468405 0.93041383 0.93438209]

Accuracy:  0.9369420433612017

Classification Report
              precision    recall  f1-score   support

         0.0       0.94      0.94      0.94      3505
         1.0       0.94      0.94      0.94      3552

    accuracy                           0.94      7057
   macro avg       0.94      0.94      0.94      7057
weighted avg       0.94      0.94      0.94      7057


Confusion Matrix
[[3282  223]
 [ 222 3330]]

ROC AUC score
0.936938302425107
```

SVM

```
                    SVM

Cross Validation Scores
[0.95097761 0.94672712 0.94176821 0.93835034 0.94529478]

Accuracy:  0.9424684710216806

Classification Report
              precision    recall  f1-score   support

         0.0       0.93      0.96      0.94      3549
         1.0       0.96      0.92      0.94      3508

    accuracy                           0.94      7057
   macro avg       0.94      0.94      0.94      7057
weighted avg       0.94      0.94      0.94      7057


Confusion Matrix
[[3412  137]
 [ 269 3239]]

ROC AUC score
0.9423578533853949
```

Then we also went on to use some Boosting and Bagging techniques in order to improve our predictions. Our results were as follows:

```
                XGB

Cross Validation Scores
[0.95097761 0.94672712 0.94176821 0.93835034 0.94529478]

Accuracy:  0.9556468754428227

Classification Report
              precision    recall  f1-score   support

         0.0       0.94      0.97      0.96      3539
         1.0       0.97      0.94      0.95      3518

    accuracy                           0.96      7057
   macro avg       0.96      0.96      0.96      7057
weighted avg       0.96      0.96      0.96      7057


Confusion Matrix
[[3435  104]
 [ 209 3309]]

ROC AUC score
0.9556022062935203
```

Since the Validation score was best for the boosting model we made our final prediction on the test dataset using the XGboost.

Network Destruction

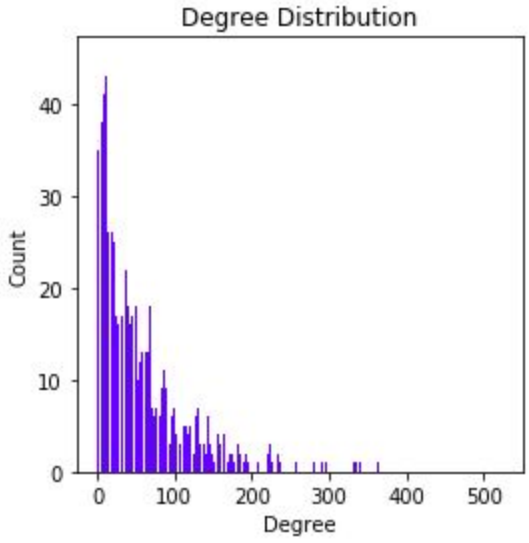
Orginal Properties:

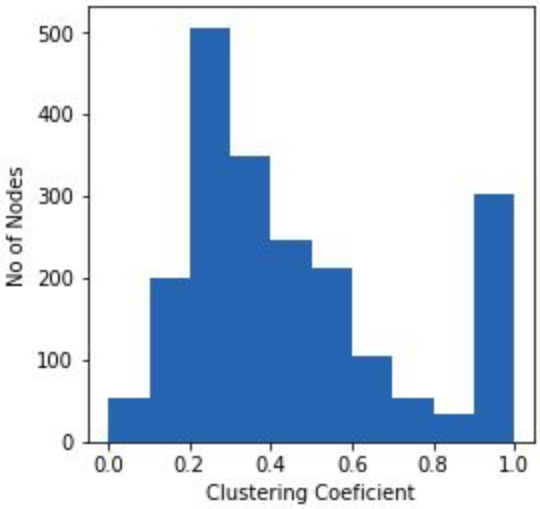
Nodes in Connected component = 16

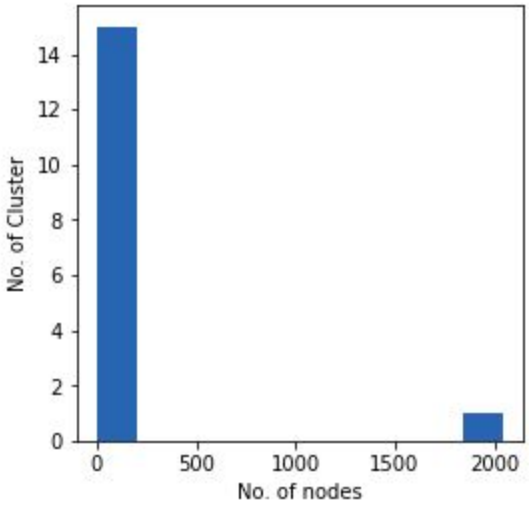
Nodes in giant cluster = 2045

Average clustering coefficient = 0.4578995252895736

Average shortest path length = 2.417796257398361







Here 30000 edges are removed to destroy the network.

Local Methods Include:

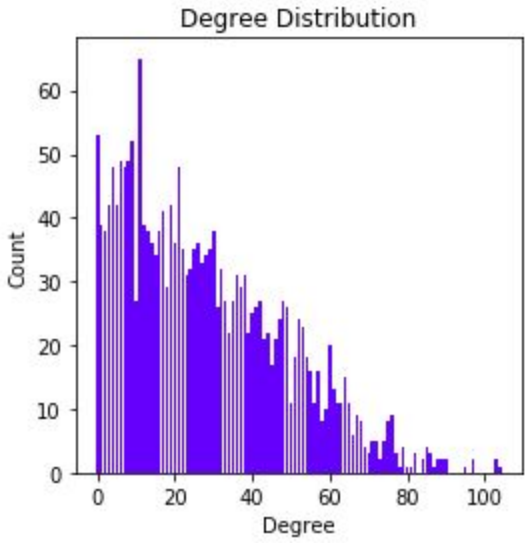
Common Neighbour (CN)

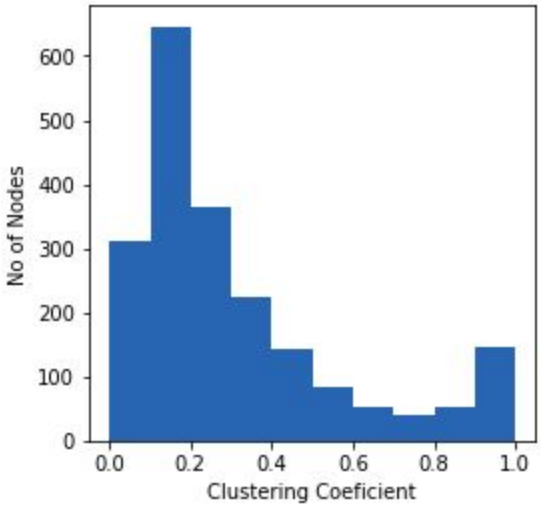
Nodes in Connected component = 54

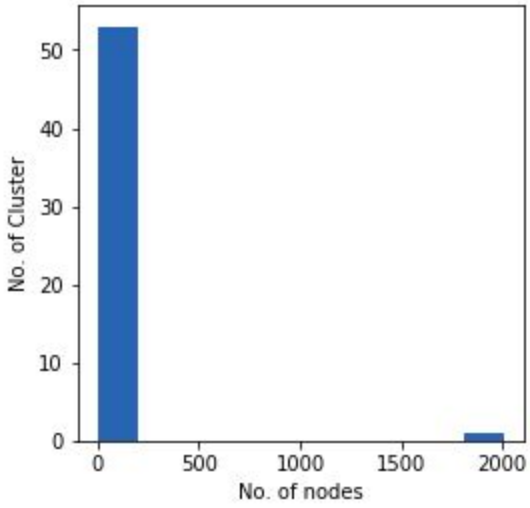
Nodes in giant cluster = 2007

Average clustering coefficient = 0.30763527220867737

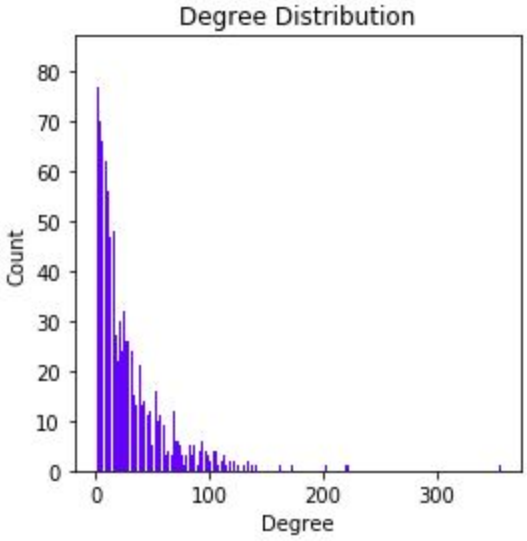
Average shortest path length = 2.7240028792546127

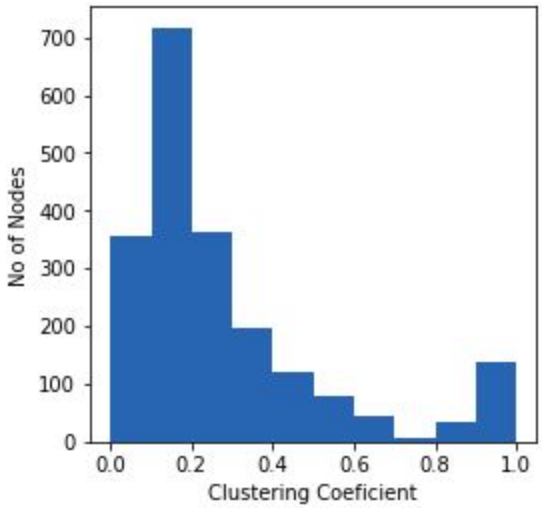


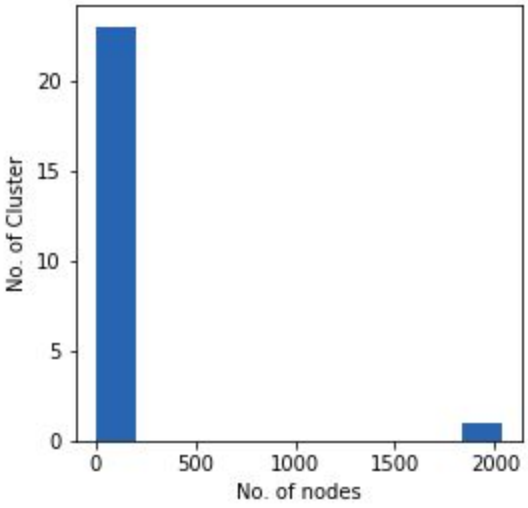




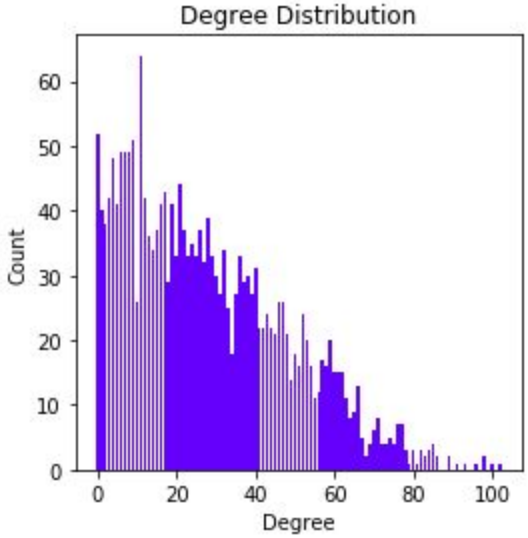
Jacard Coefficient (JC)
Nodes in Connected component = 24
Nodes in giant cluster = 2037
Average clustering coefficient = 2.651322344099773
Average shortest path length = 2.651322344099773

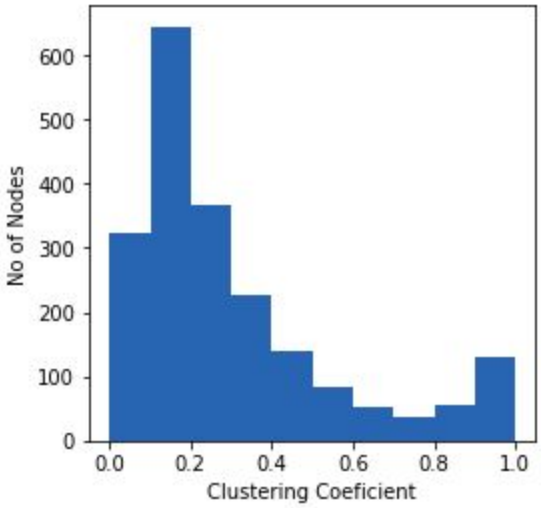


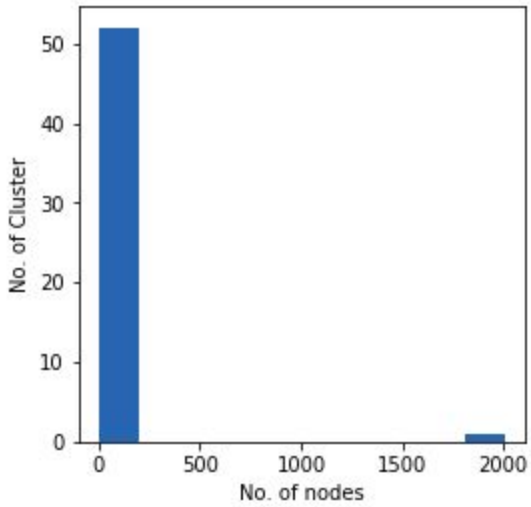




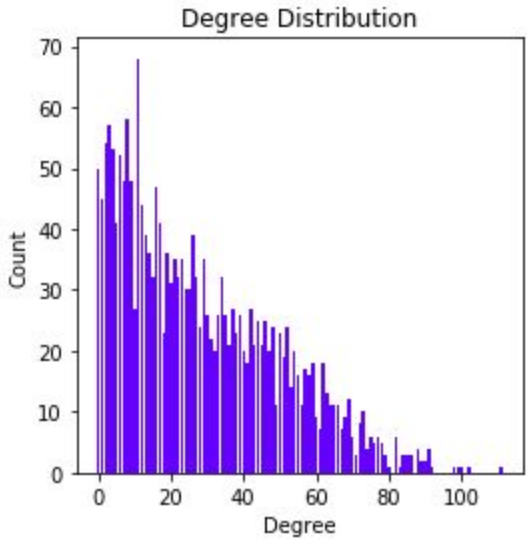
Adamic Adar (AA)
Nodes in Connected component = 53
Nodes in giant cluster = 2008
Average clustering coefficient = 0.3014681318237045
Average shortest path length = 2.723449500453592

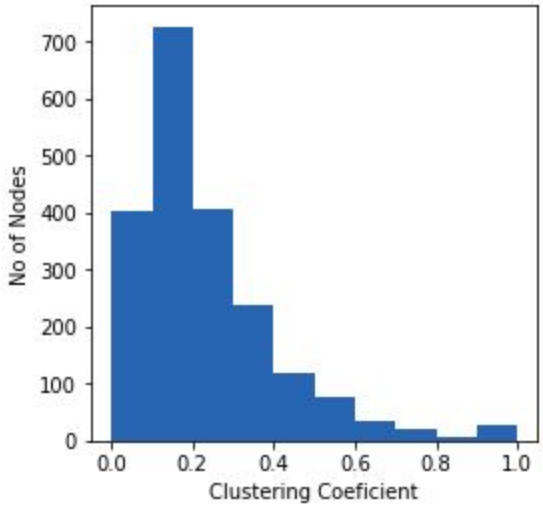


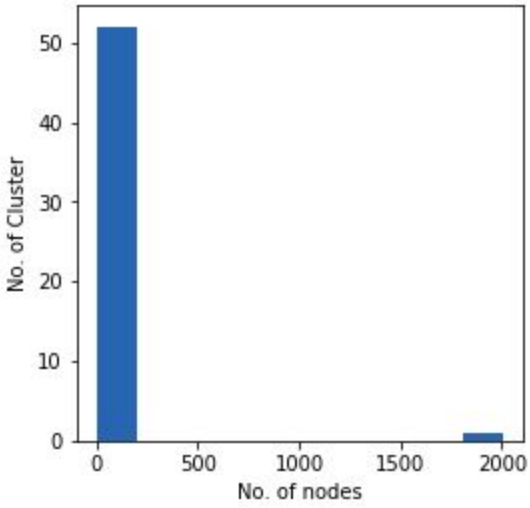




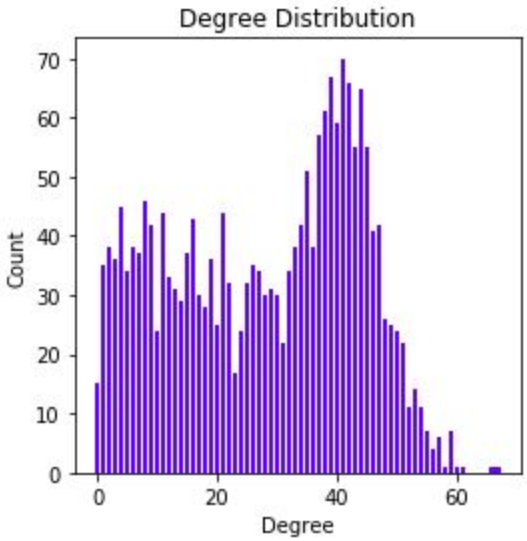
Resource Allocation(RA)
Nodes in Connected component = 53
Nodes in giant cluster = 2005
Average clustering coefficient = 0.2283178953326126
Average shortest path length = 2.7617373731340313

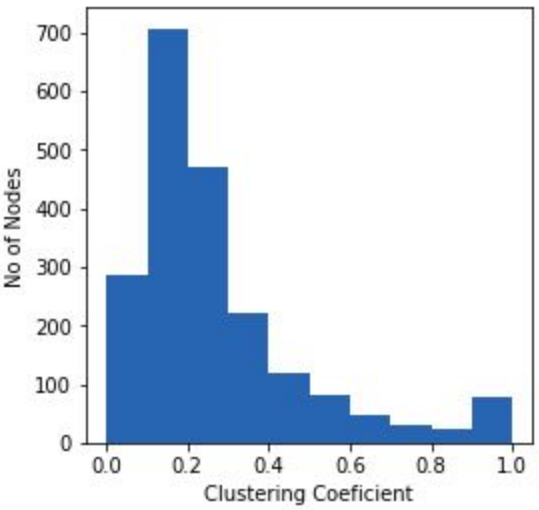


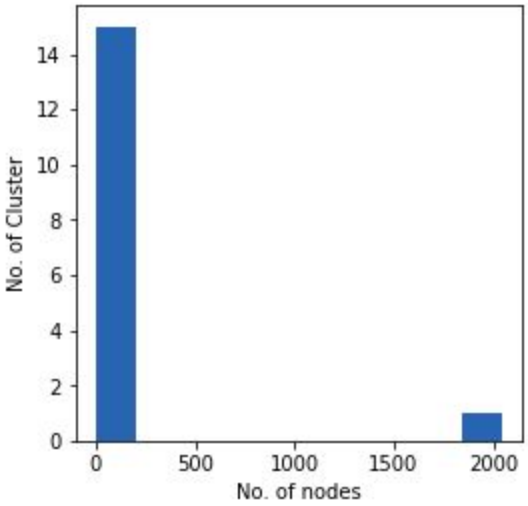




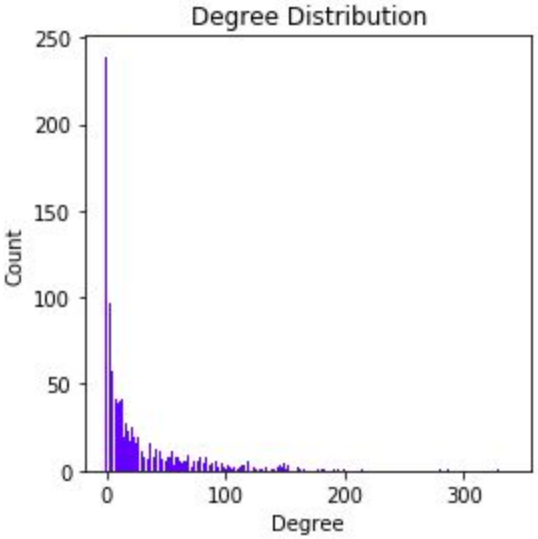
Preferential Attachment (PA)
Nodes in Connected component = 16
Nodes in giant cluster = 2045
Average clustering coefficient = 0.27037108651451036
Average shortest path length = 2.8005574189350186

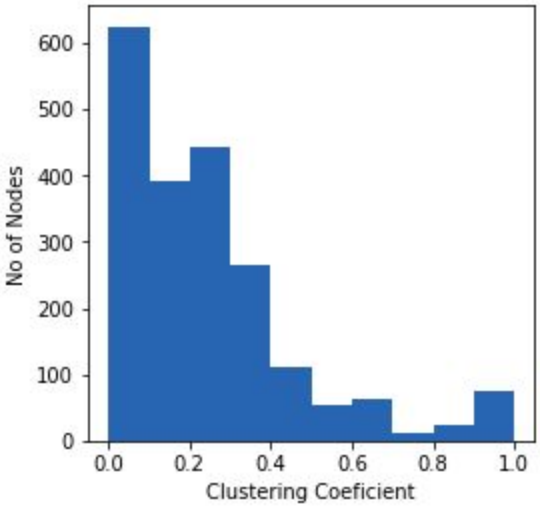


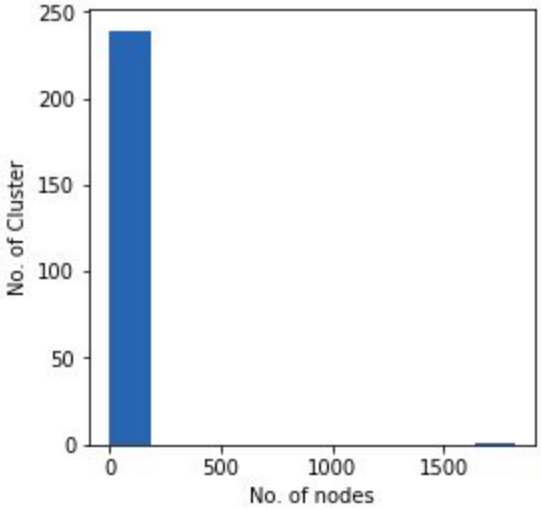




Global Methods Include:
Rooted PageRank (RP)
Nodes in Connected component = 240
Nodes in giant cluster = 1821
Average clustering coefficient = 0.23786037683610153
Average shortest path length = 2.6666666666666665


Degree Distribution
Count
Degree
0
50
100
150
200
250
0
100
200
300


No of Nodes
Clustering Coeficient
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0

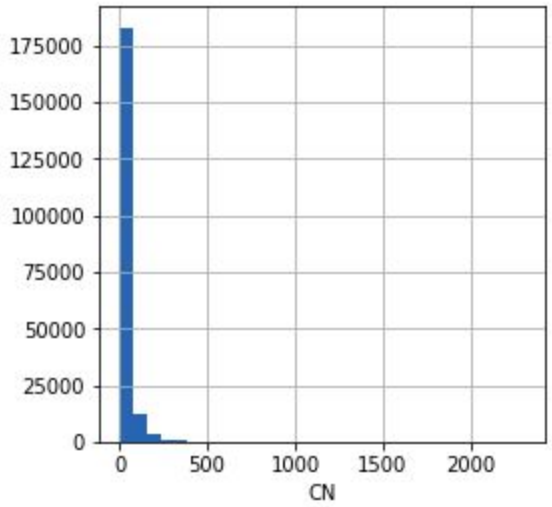

No. of Cluster
No. of nodes
0
50
100
150
200
250
0
500
1000
1500

2. BlogCatalog data

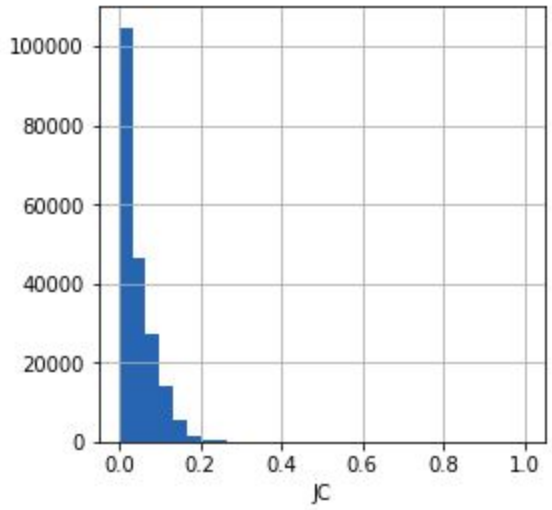
The graph is an undirected graph with 10312 nodes and 333983 edges and an average degree of nodes was 64.78.

There were many Link measures that were calculated from this network.

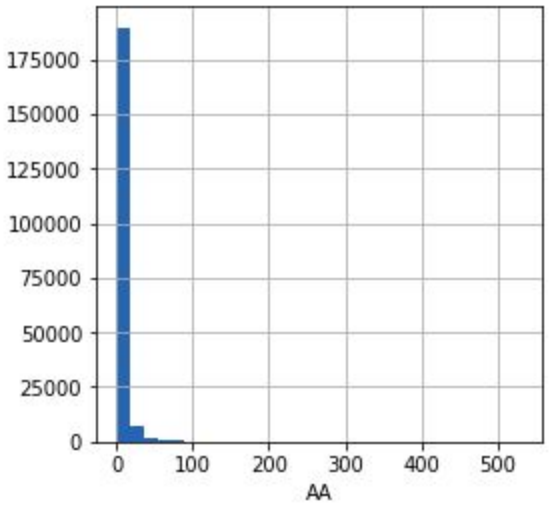
Local Methods Include:
Common Neighbour (CN)



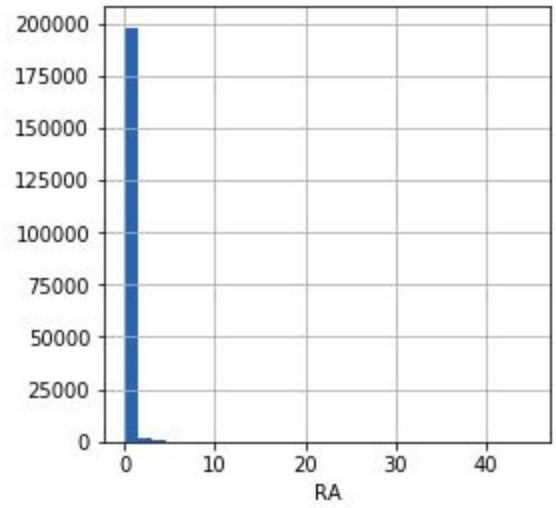
Jacard Coefficient (JC)



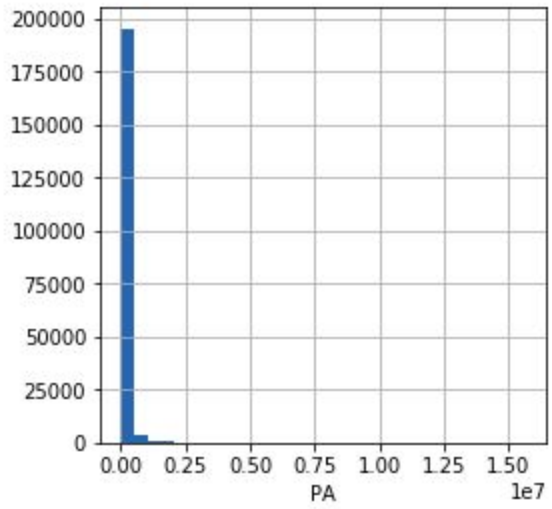
Adamic Adar (AA)

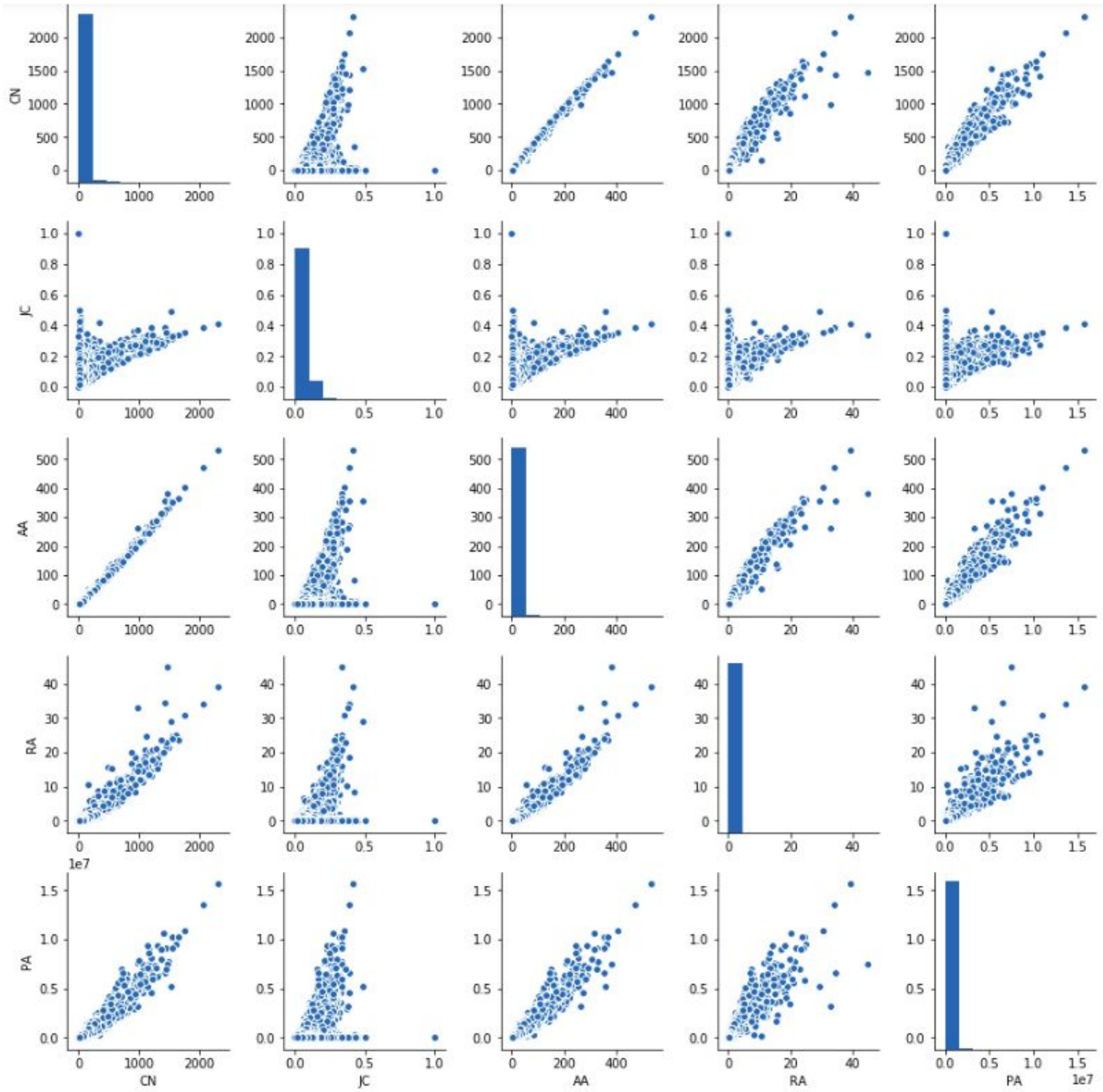


Resource Allocation(RA)

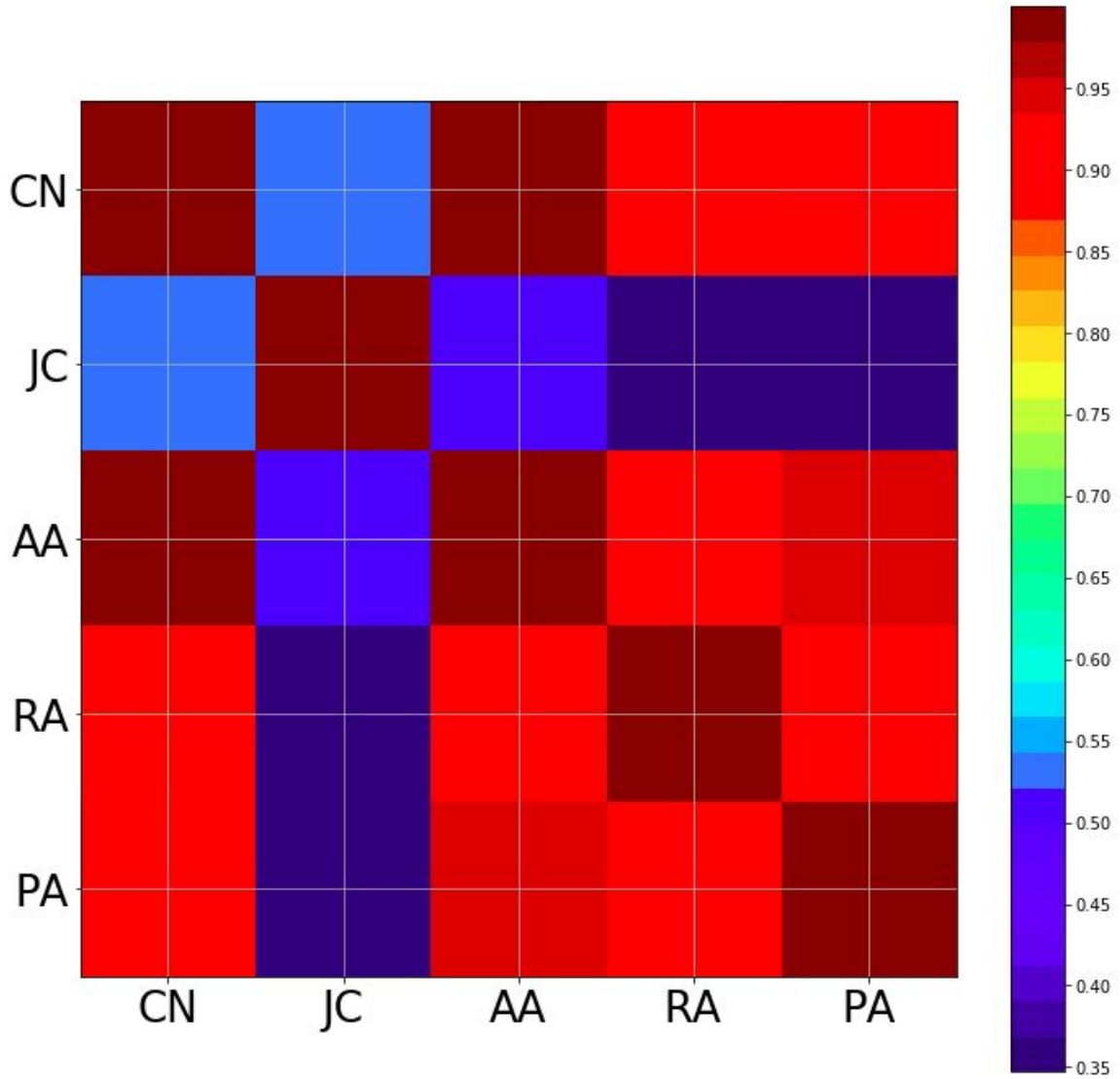


Preferential Attachment (PA)





	CN	JC	AA	RA	PA
CN	1.000000	0.538082	0.993660	0.880533	0.931319
JC	0.538082	1.000000	0.502424	0.364308	0.347139
AA	0.993660	0.502424	1.000000	0.926524	0.952675
RA	0.880533	0.364308	0.926524	1.000000	0.921784
PA	0.931319	0.347139	0.952675	0.921784	1.000000



['overall AUC', '-', 0.9487235201500089, 0.7643206754364498, 0.9528003024473359, 0.9607169304950027, 0.9550099640647768]
['PR@25%', '-', ['0.978', '0.489'], ['0.783', '0.392'], ['0.980', '0.490'], ['0.985', '0.493'], ['0.980', '0.490']]
['PR@50%', '-', ['0.872', '0.870'], ['0.674', '0.674'], ['0.885', '0.885'], ['0.900', '0.900'], ['0.887', '0.887']]
['PR@75%', '-', ['0.618', '0.984'], ['0.614', '0.922'], ['0.660', '0.990'], ['0.660', '0.990'], ['0.661', '0.992']]

MACHINE LEARNING PART

We used **5 fold Cross-Validation** to tune the parameters of the model. The results were:

```
               GaussianNB

Accuracy:  0.7325465342581965

Classification Report
              precision    recall  f1-score   support

         0.0       0.65      0.99      0.79     19983
         1.0       0.98      0.48      0.64     20095

    accuracy                           0.73     40078
   macro avg       0.82      0.73      0.71     40078
weighted avg       0.82      0.73      0.71     40078


Confusion Matrix
[[19792   191]
 [10528  9567]]

ROC AUC score
0.7332652274214123
```

```
               GaussianNB

Accuracy:  0.7381106841658766

Classification Report
              precision    recall  f1-score   support

         0.0       0.66      0.99      0.79     20129
         1.0       0.98      0.48      0.65     19949

    accuracy                           0.74     40078
   macro avg       0.82      0.74      0.72     40078
weighted avg       0.82      0.74      0.72     40078


Confusion Matrix
[[19933   196]
 [10300  9649]]

ROC AUC score
0.7369730987798009
```

```
               GaussianNB

Accuracy:  0.7543789610259993

Classification Report
              precision    recall  f1-score   support

         0.0       0.67      0.99      0.80     19861
         1.0       0.98      0.52      0.68     20217

    accuracy                           0.75     40078
   macro avg       0.82      0.76      0.74     40078
weighted avg       0.83      0.75      0.74     40078


Confusion Matrix
[[19622   239]
 [ 9605 10612]]

ROC AUC score
0.7564355746743673

               GaussianNB

Accuracy:  0.7407804780677678

Classification Report
              precision    recall  f1-score   support

         0.0       0.66      0.99      0.79     20003
         1.0       0.98      0.49      0.66     20075

    accuracy                           0.74     40078
   macro avg       0.82      0.74      0.72     40078
weighted avg       0.82      0.74      0.72     40078


Confusion Matrix
[[19804   199]
 [10190  9885]]

ROC AUC score
0.7412274896000968
```

```
                GaussianNB

Accuracy:  0.7390837866160986

Classification Report
              precision    recall  f1-score   support

         0.0       0.66      0.99      0.79     20219
         1.0       0.98      0.48      0.65     19859

    accuracy                           0.74     40078
   macro avg       0.82      0.74      0.72     40078
weighted avg       0.82      0.74      0.72     40078


Confusion Matrix
[[20012   207]
 [10250  9609]]

ROC AUC score
0.7368116632815781
```

SVM

```
                SVM

Cross Validation Scores
[0.84313089 0.84572583 0.84550127 0.84382953 0.84253206]

Accuracy:  0.8452018563800588

Classification Report
              precision    recall  f1-score   support

         0.0       0.78      0.96      0.86     20219
         1.0       0.95      0.73      0.82     19859

    accuracy                           0.85     40078
   macro avg       0.86      0.84      0.84     40078
weighted avg       0.86      0.85      0.84     40078


Confusion Matrix
[[19406   813]
 [ 5391 14468]]

ROC AUC score
0.8441632381627434
```

Decision Tree

```
                Decision Tree

Cross Validation Scores
[0.86496332 0.86798243 0.86666001 0.86773292 0.86601128]

Accuracy:  0.867508358700534

Classification Report
              precision    recall  f1-score   support

         0.0       0.87      0.87      0.87     20219
         1.0       0.87      0.86      0.87     19859

    accuracy                           0.87     40078
   macro avg       0.87      0.87      0.87     40078
weighted avg       0.87      0.87      0.87     40078


Confusion Matrix
[[17598  2621]
 [ 2689 17170]]

ROC AUC score
0.8674824260131309
```

After PCA

```
                GaussianNB

Accuracy:  0.6970407704975298

Classification Report
              precision    recall  f1-score   support

         0.0       0.62      0.99      0.77     20061
         1.0       0.98      0.40      0.57     20017

    accuracy                           0.70     40078
   macro avg       0.80      0.70      0.67     40078
weighted avg       0.80      0.70      0.67     40078


Confusion Matrix
[[19921   140]
 [12002  8015]]

ROC AUC score
0.696715468438239

                GaussianNB

Accuracy:  0.7929287888617197

Classification Report
              precision    recall  f1-score   support

         0.0       0.71      0.98      0.83     20139
         1.0       0.97      0.61      0.74     19939

    accuracy                           0.79     40078
   macro avg       0.84      0.79      0.79     40078
weighted avg       0.84      0.79      0.79     40078


Confusion Matrix
[[19714   425]
 [ 7874 12065]]

ROC AUC score
0.7919961047787938
```

GaussianNB

Accuracy: 0.6948949548380657

Classification Report

	precision	recall	f1-score	support
0.0	0.62	0.99	0.76	19960
1.0	0.99	0.40	0.57	20118
accuracy			0.69	40078
macro avg	0.80	0.70	0.67	40078
weighted avg	0.80	0.69	0.67	40078

Confusion Matrix

```
[[19845   115]
 [12113  8005]]
```

ROC AUC score

0.6960704264678078

GaussianNB

Accuracy: 0.6947951494585558

Classification Report

	precision	recall	f1-score	support
0.0	0.62	0.99	0.76	19966
1.0	0.98	0.40	0.57	20112
accuracy			0.69	40078
macro avg	0.80	0.70	0.67	40078
weighted avg	0.80	0.69	0.67	40078

Confusion Matrix

```
[[19843   123]
 [12109  8003]]
```

ROC AUC score

0.6958805830094135

```
                GaussianNB

Accuracy:  0.6978891162233645

Classification Report
              precision    recall  f1-score   support

         0.0       0.62      0.99      0.77     20069
         1.0       0.98      0.40      0.57     20009

    accuracy                           0.70     40078
   macro avg       0.80      0.70      0.67     40078
weighted avg       0.80      0.70      0.67     40078


Confusion Matrix
[[19926   143]
 [11965  8044]]

ROC AUC score
0.6974468370492932
```

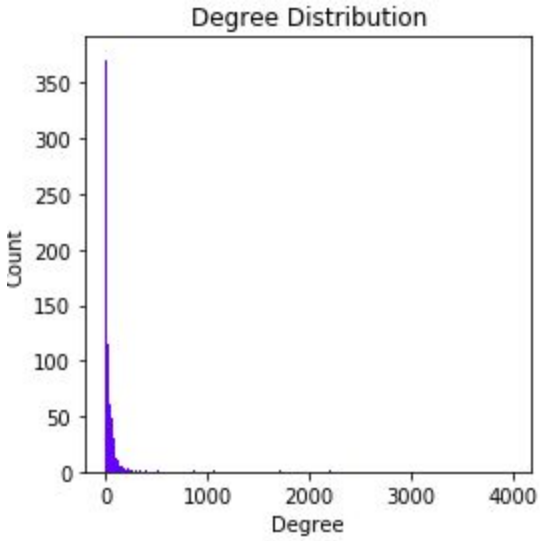
Decision Tree

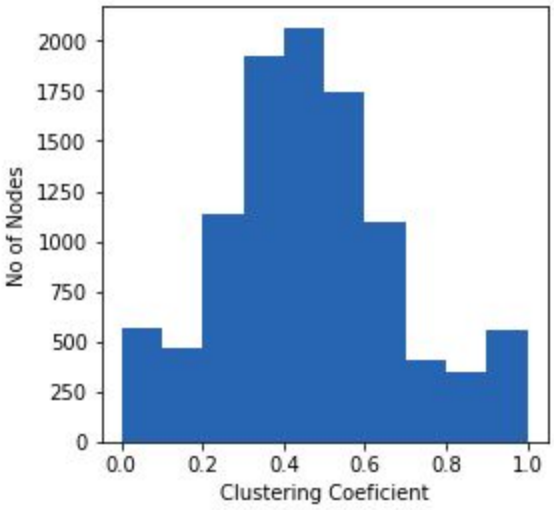
```
                Decision Tree

Cross Validation Scores
[0.86760816 0.86476371 0.8682569  0.8682569  0.8682569 ]

Accuracy:  0.863940316383053

Classification Report
              precision    recall  f1-score   support

         0.0       0.86      0.87      0.86     20069
         1.0       0.86      0.86      0.86     20009

    accuracy                           0.86     40078
   macro avg       0.86      0.86      0.86     40078
weighted avg       0.86      0.86      0.86     40078


Confusion Matrix
[[17360  2709]
 [ 2744 17265]]

ROC AUC score
0.8639387040394082
```

Network Destruction

Orginal Properties:

Average clustering coefficient = 0.4631956780330237





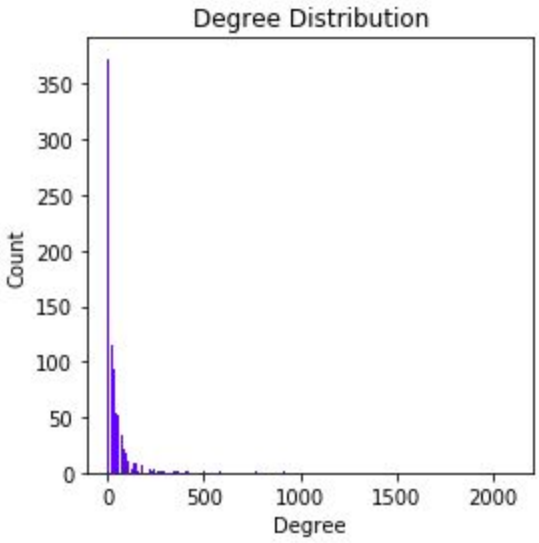
Here 100000 edges are removed to destroy the network.

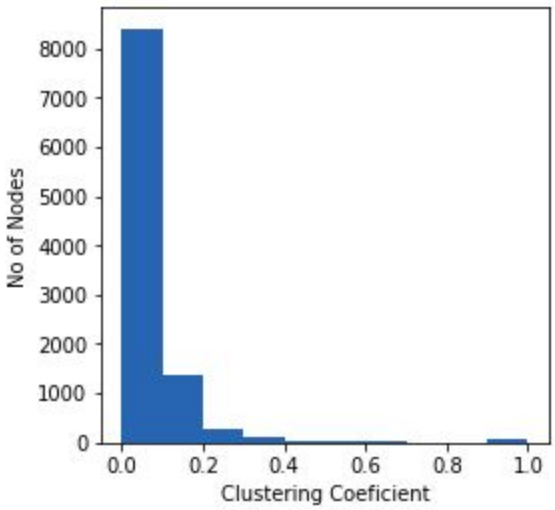
Local Methods Include:

Common Neighbour (CN)

Average clustering coefficient = 0.06885971728334217

Average shortest path length = 2.6046527471960284

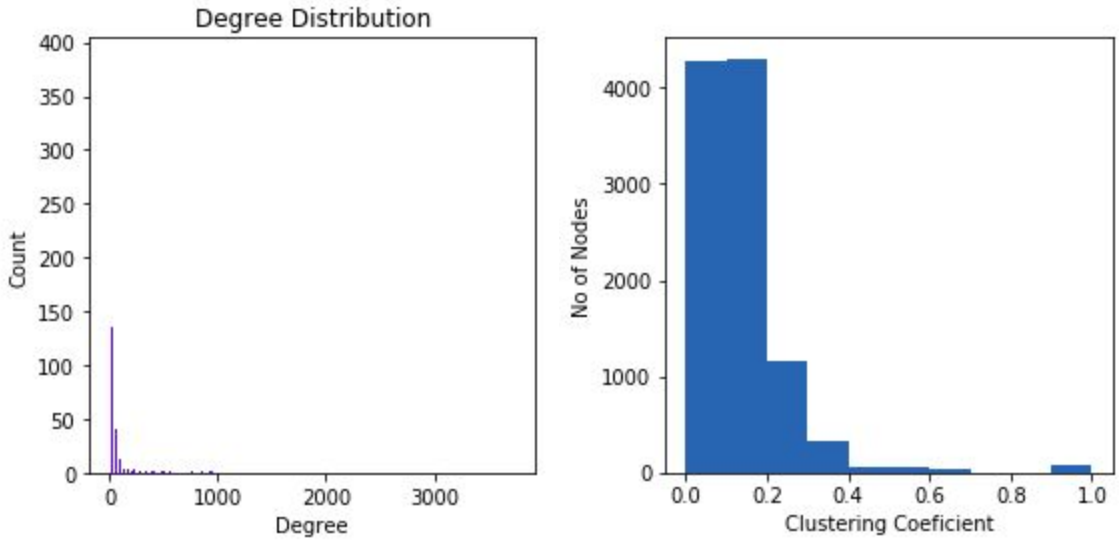




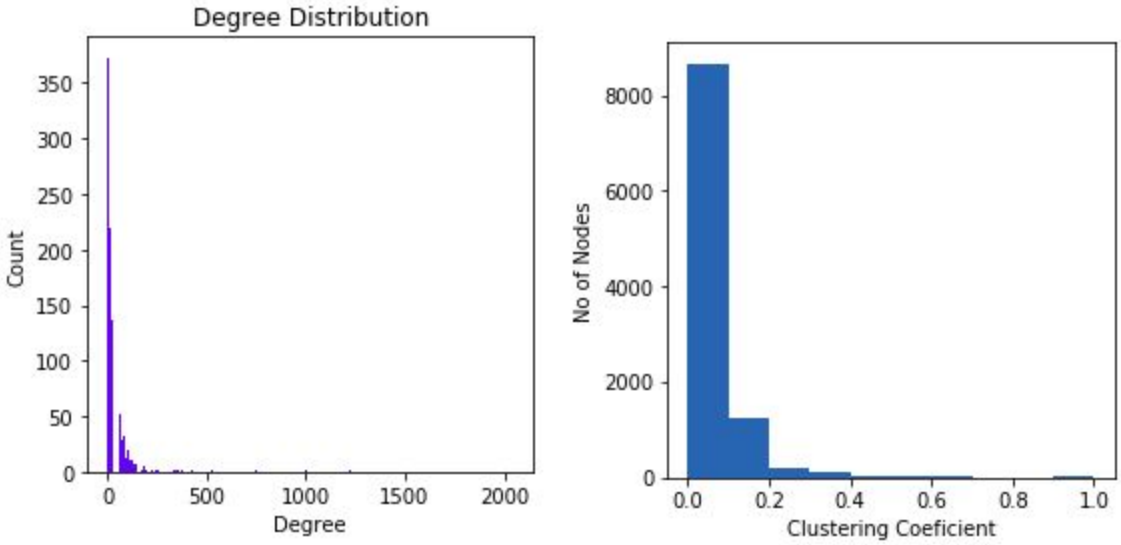
Jacard Coefficient (JC)

Average clustering coefficient = 0.13110953980015427

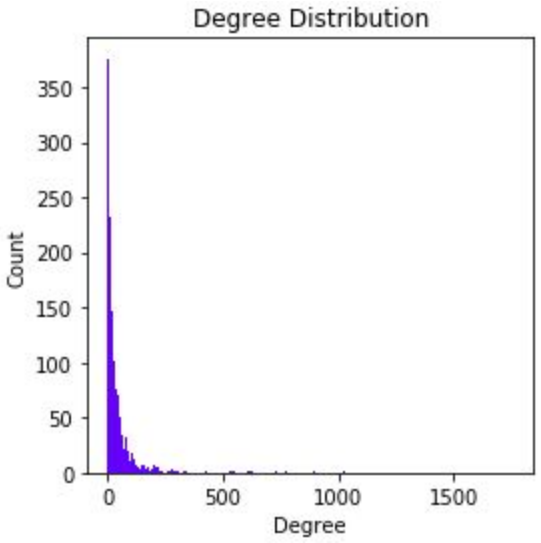
Average shortest path length = 2.4272320325841505

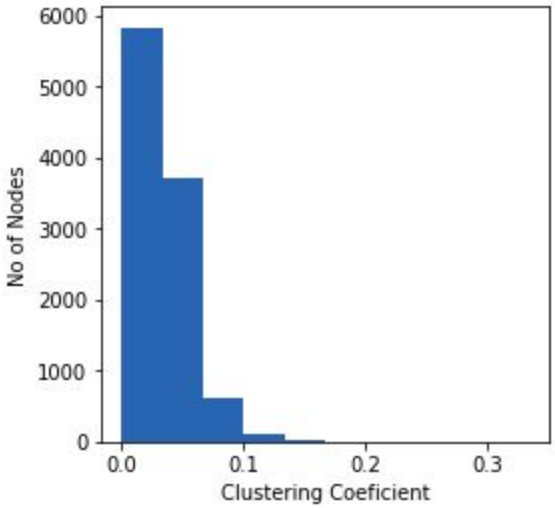


Adamic Adar (AA)
Average clustering coefficient = 0.0619709796228059
Average shortest path length = 2.611849581205276



Resource Allocation(RA)
Average clustering coefficient = 0.028196369929399157
Average shortest path length = 2.6476208232728626





Preferential Attachment (PA)

Average clustering coefficient = 0.07343092848462593

Average shortest path length = 2.7082501842052733