CS529 Data Mining
Assignment-II Phase-II

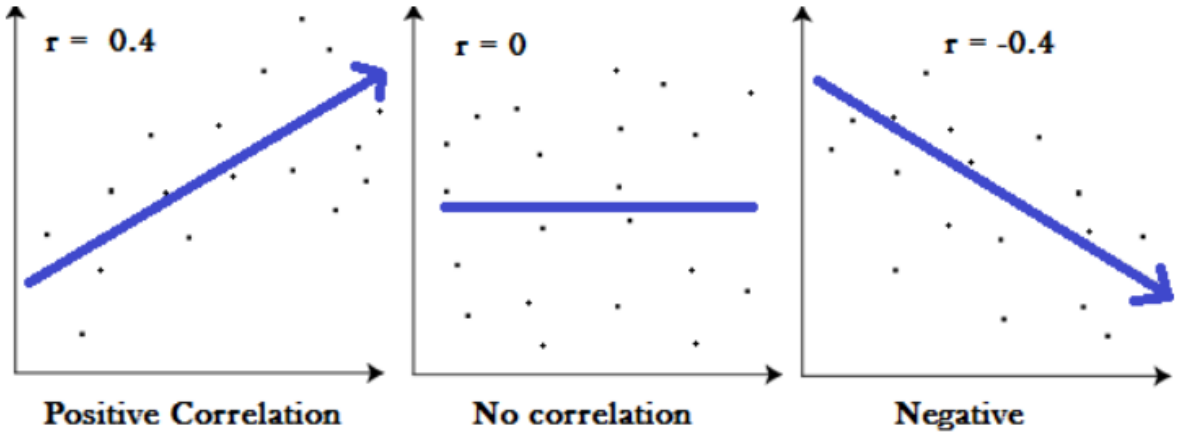
Group 11

Ayush Jain	150101012
Mukul Verma	150101038
Shubhanshu Verma	150101073

Introduction

This report shows coefficient analysis of the well-known link prediction metrics (common neighbour, jacard coefficient, adamic adar, resource allocation and preferential attachment) for network analysis studies on real-world network graphs representing diverse domains (ranging from 61 nodes to 10410 nodes).

-We use the Pearson correlation for correlation analysis which is a measure of the linear correlation between two variables X and Y. Owning to Cauchy-Schwarz inequality it has values between -1 and +1, where +1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

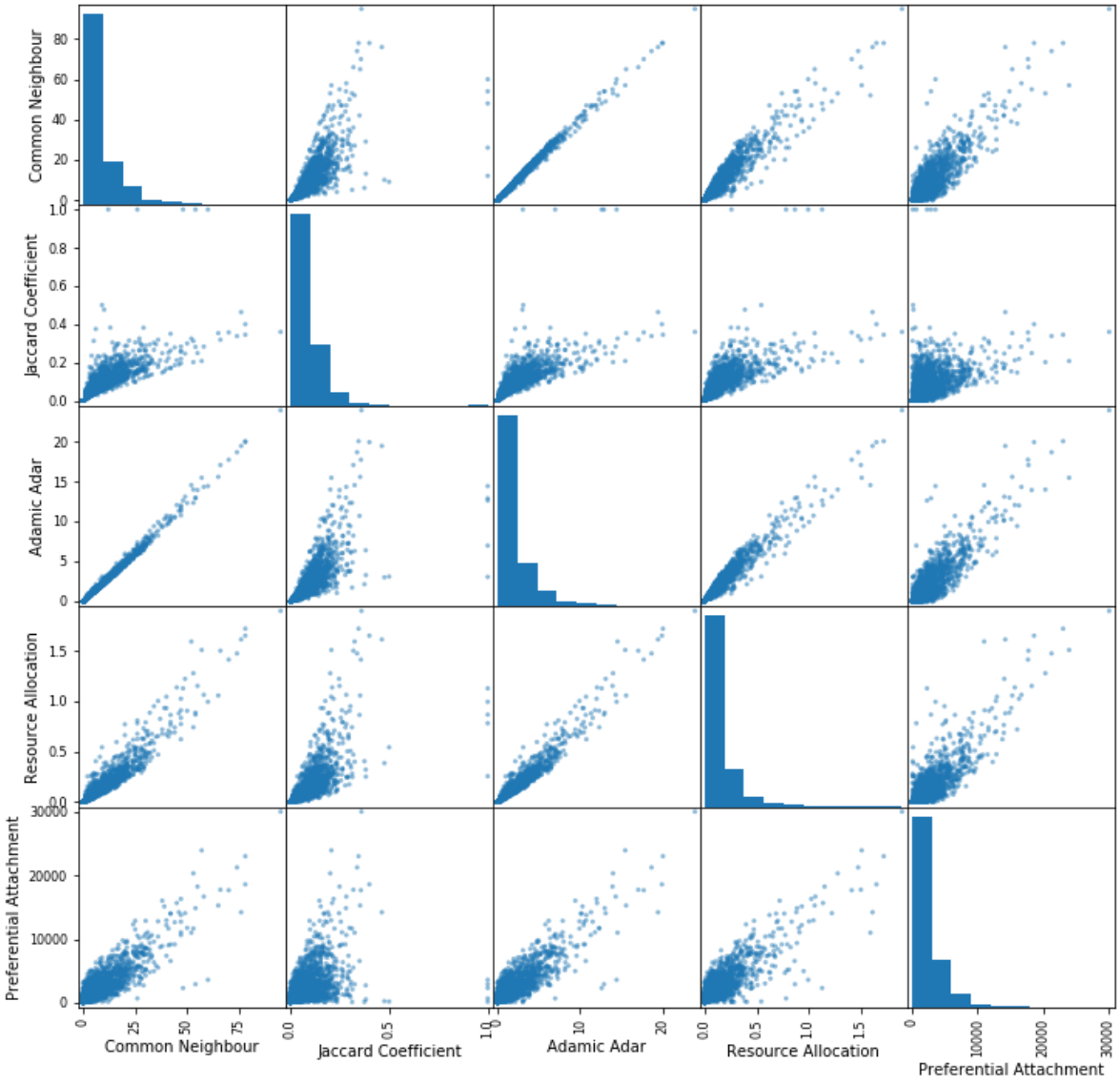


Scatterplot Matrix

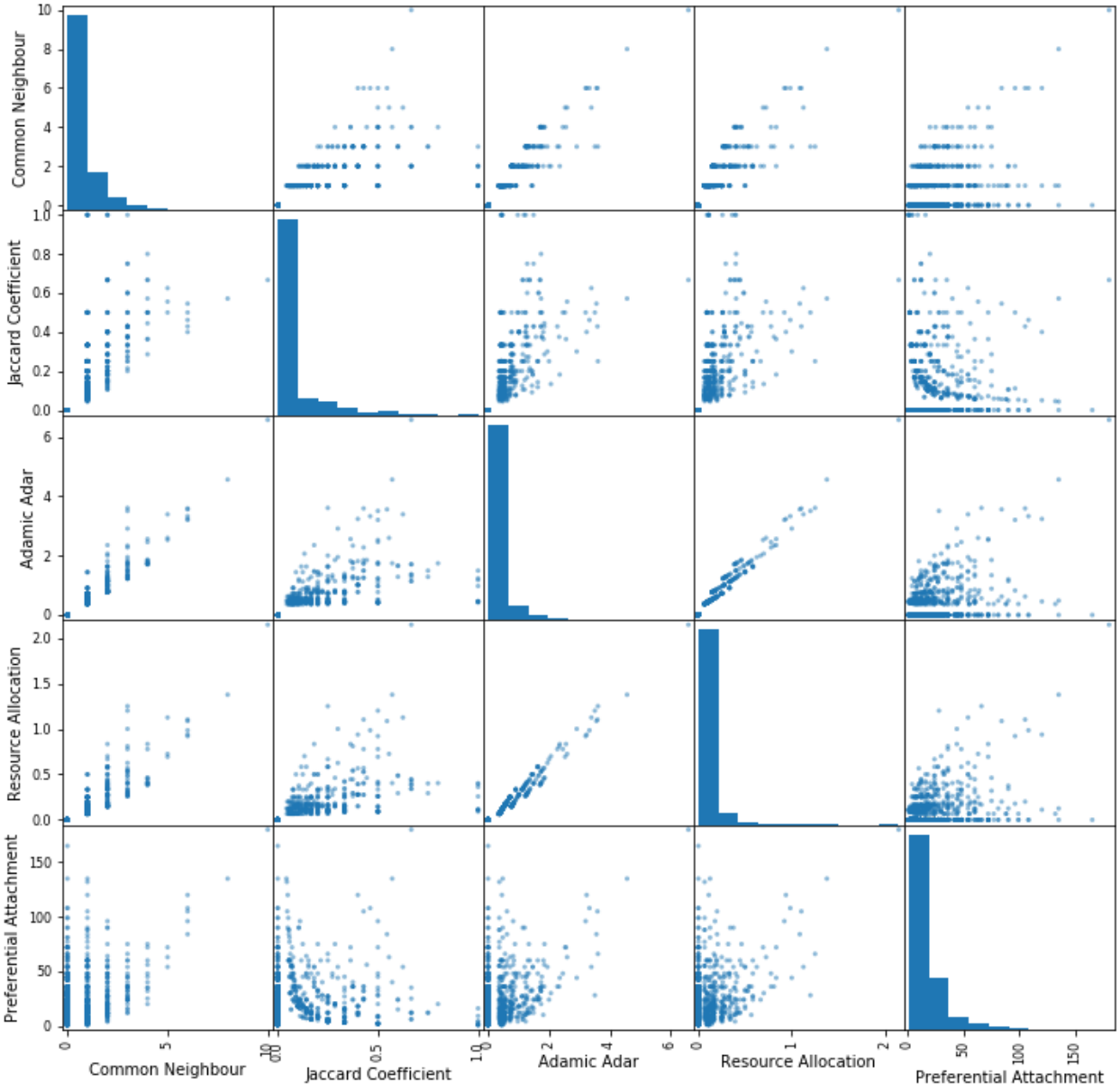
A scatterplot shows the relationship between two variables as dots in two dimensions, one axis for each attribute.

Scatter plots are useful for spotting structured relationships between variables, like whether you could summarize the relationship between two variables with a line. Attributes with structured relationships may also be correlated.

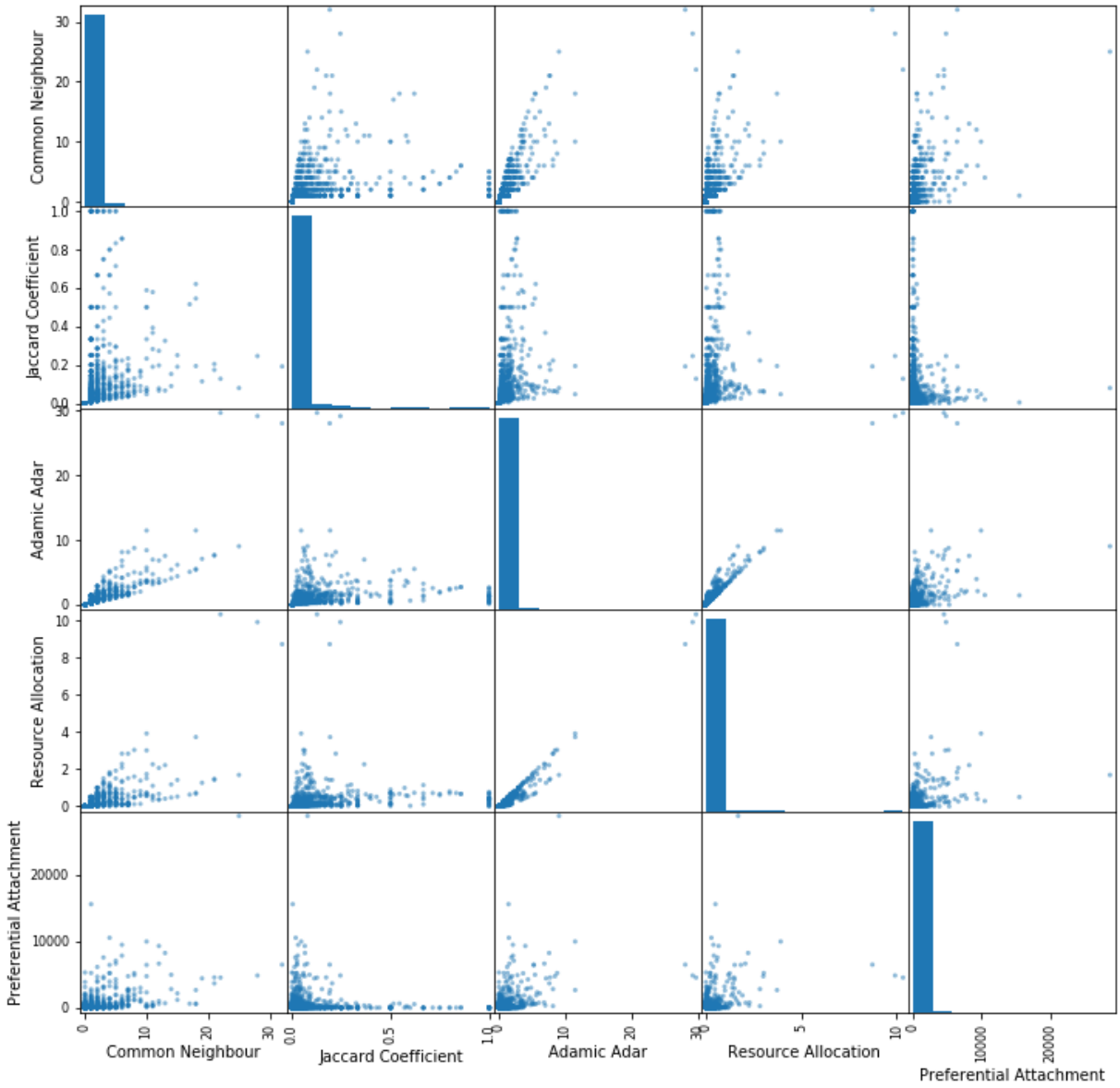
Higgs Titter Scatter Matrix



9-11 Hijackers Scatter Matrix



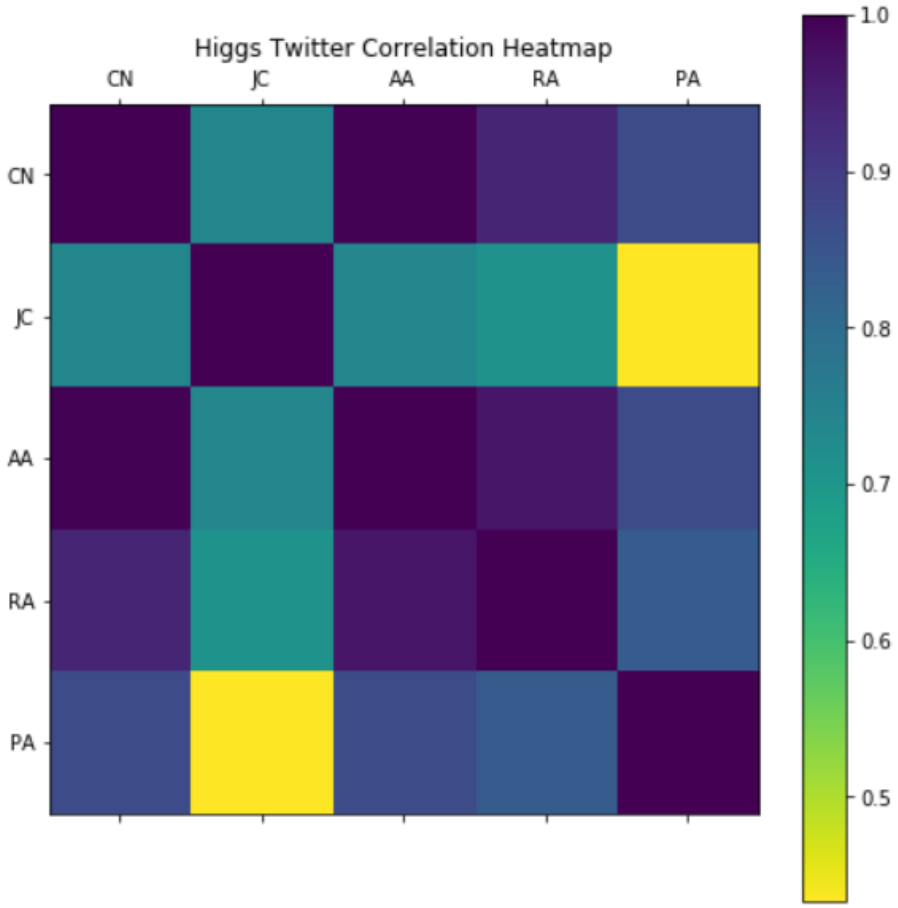
DBLP Co-authorship Scatter Matrix



Pearson Correlation Values

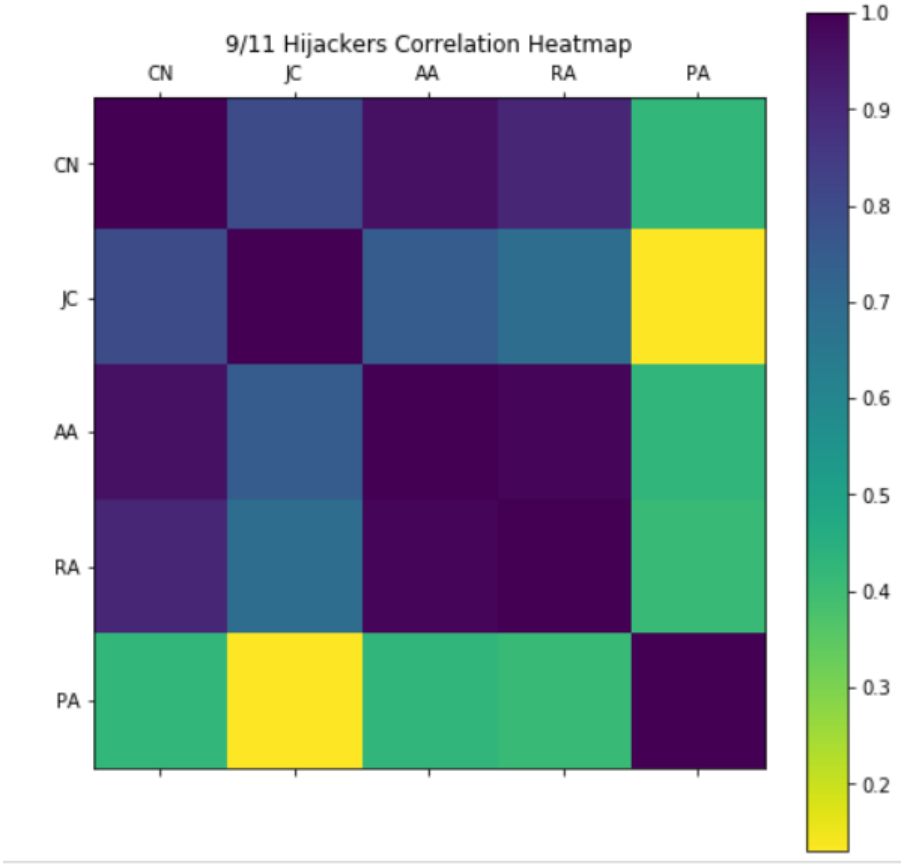
Higgs Twitter Dataset

	Common Neighbours	Jacard Coefficient	Adamic Adar	Resource Allocation	Preferential Attachment
Common Neighbours	1	0.766854	0.996709	0.946936	0.857708
Jacard Coefficient	0.766854	1	0.761071	0.720395	0.431916
Adamic Adar	0.996709	0.996709	1	0.969585	0.861190
Resource Allocation	0.946936	0.946936	0.969585	1	0.835564
Preferential Attachment	0.857708	0.857708	0.861190	0.835564	1



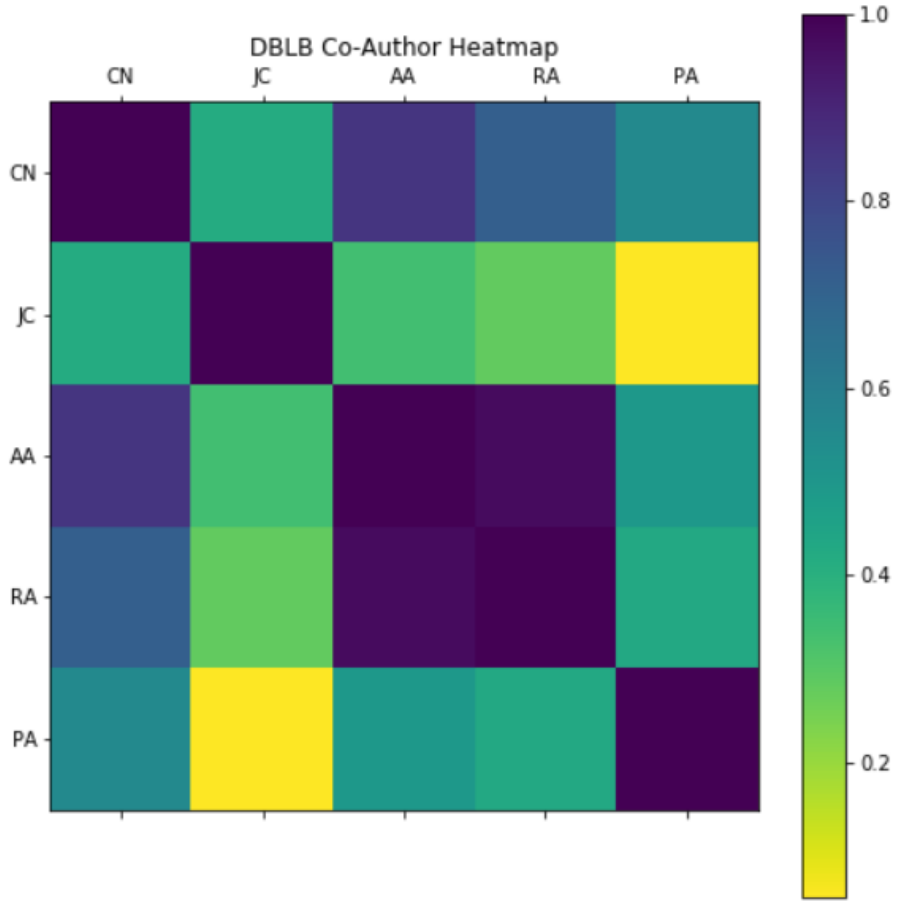
9/11 Hijackers Dataset

	Common Neighbours	Jacard Coefficient	Adamic Adar	Resource Allocation	Preferential Attachment
Common Neighbours	1	0.797406	0.959903	0.9054046	0.425382
Jacard Coefficient	0.797406	1	0.745208	0.688225	0.1298449
Adamic Adar	0.959903	0.745208	1	0.987026	0.425712
Resource Allocation	0.9054046	0.688225	0.987026	1	0.414662
Preferential Attachment	0.425382	0.1298449	0.425712	0.414662	1



DBLP Co-authorship Dataset

	Common Neighbours	Jacard Coefficient	Adamic Adar	Resource Allocation	Preferential Attachment
Common Neighbours	1	0.420160	0.855039	0.715211	0.554953
Jacard Coefficient	0.420160	1	0.339517	0.283474	0.055221
Adamic Adar	0.855039	0.339517	1	0.973004	0.495682
Resource Allocation	0.715211	0.283474	0.973004	1	0.428627
Preferential Attachment	0.554953	0.055221	0.495682	0.428627	1



Higgs Twitter Rank Correlations

	CN	JC	AA	RA	PA
CN	1.000000	0.928227	0.995567	0.963309	0.765110
JC	0.928227	1.000000	0.930608	0.895453	0.518982
AA	0.995567	0.930608	1.000000	0.979858	0.754554
RA	0.963309	0.895453	0.979858	1.000000	0.734886
PA	0.765110	0.518982	0.754554	0.734886	1.000000

9/11 Hijackers Rank Correlations

	CN	JC	AA	RA	PA
CN	1.000000	0.992824	0.997579	0.996689	0.312144
JC	0.992824	1.000000	0.993528	0.992840	0.262872
AA	0.997579	0.993528	1.000000	0.999756	0.306673
RA	0.996689	0.992840	0.999756	1.000000	0.305372
PA	0.312144	0.262872	0.306673	0.305372	1.000000

DBLP Co-Authorship Rank Correlations

	CN	JC	AA	RA	PA
CN	1.000000	0.998148	0.999156	0.998653	0.465608
JC	0.998148	1.000000	0.998640	0.998611	0.458946
AA	0.999156	0.998640	1.000000	0.999852	0.464158
RA	0.998653	0.998611	0.999852	1.000000	0.463163
PA	0.465608	0.458946	0.464158	0.463163	1.000000

Critical Analysis

- We observe that the correlation values between Adamic/Adar and Resource Allocation to be very high which is quite evident from their formulas. In Adamic/Adar, we take the summation of inverse of log values will in Resource Allocation we take the summation without taking log:

Higgs Twitter:	0.969585
DBLP Co-authorship network:	0.973004
9/11 Hijackers:	0.987026

- We observe that even though Common Neighbours and Jacard Coefficient have $|\zeta(A) \cap \zeta(B)|$ in the numerator, the term $|\zeta(A) \cup \zeta(B)|$ in the denominator of Jacard coefficient in inversely proportional hence we get a moderate value, not a high value.

Higgs Twitter:	0.766854
DBLP Co-authorship network:	0.420160
9/11 Hijackers:	0.797406

- We observe very high correlation values between Common Neighbours and Adamic/Adar , Resource Allocation. In Adamic/Adar & Resource Allocation, as the Common Neighbours increase, the no. of terms in summation increase.

	CN-AA	CN-RA
Higgs Twitter:	0.996709	0.946936
DBLP Co-authorship network:	0.855039	0.715211
9/11 Hijackers:	0.959903	0.9054046

- We observe very low correlation values for Jacard Coefficient and Preferntial Attachment. This can be seen as if we increase the neighbours of the nodes, the preferential attachment increases but the denominator of Jacard Coefficient is more likely to increase then the numerator, hence it will decrease.

DBLP Co-authorship network:	0.1298449
9/11 Hijackers:	0.055221
Higgs Twitter:	0.431916

- We observe a moderate correlation value between Preferential Attachment and Adamic/Adar & Resource Allocation as there are some chances that as neighbours of A or B increase, the common neighbours of A and B will also increase

	PA-AA	PA-RA
DBLP Co-authorship network:	0.495682	0.428627
9/11 Hijackers:	0.425712	0.495682
Higgs Twitter:	0.861190	0.835564

Classification model Analysis

Decision Tree

	Precision	Recall	Accuracy	AUC
Higgs Twitter	0.635	0.654	0.65	0.650
9/11 Hijackers	0.875	0.84	0.86	0.76
DBLP Co-authorship	0.885	0.948	0.915	0.915

SVM

	Precision	Recall	Accuracy	AUC
Higgs Twitter	0.621	0.83	0.662	0.662
9/11 Hijackers	0.696	0.92	0.76	0.76
DBLP Co-authorship	0.885	0.969	0.923	0.924

Naïve Bayes

	Precision	Recall	Accuracy	AUC
Higgs Twitter	0.798	0.443	0.657	0.656
9/11 Hijackers	0.928	0.52	0.74	0.73
DBLP Co-authorship	0.976	0.348	0.675	0.670

- As we can see from the calculated values, the order of AUC scores were SVM > Decision Tree > Naïve Bayes

- We observe high accuracy in DBLP Co-authorship network due to the nature of subgraph chosen and training dataset. The subgraph contains around 60000 edges. To preserve the properties of this subgraph, <5% edges are removed(3000). The training dataset is of 4800 data points.

- We observe very low recall in Naïve Bayes value across all three datasets.