Network Analysis of
Big Data Technologies

Shivam Singh
aba21shivam@iima.ac.in

Abstract

In this project we used Network Analysis to study the network of Big Data Technologies. Through Network Analysis we are trying to understand how one tool/technology relates with another. Technologies are depicted as nodes and their relation as edges.

Introduction

Big data define ways to analyse, extract information from data sets that are too large or complex to be dealt with by traditional data-processing applications. Big data analytics includes data ingestion, data storage, data analysis, search, transfer, visualization and querying.

Current usage of the term big data tends to refer to the use of predictive analytics, user behaviour analytics, or other advanced data analytics methods that extract value from big data.

Methodology

Big Data technology space is very dynamic in nature and requires a lot of technologies interaction to get insights from data. For analysis, we computed an xlsx format dataset then we computed an adjacency matrix to determine relationship between technology. Technologies were considered to be connected if :

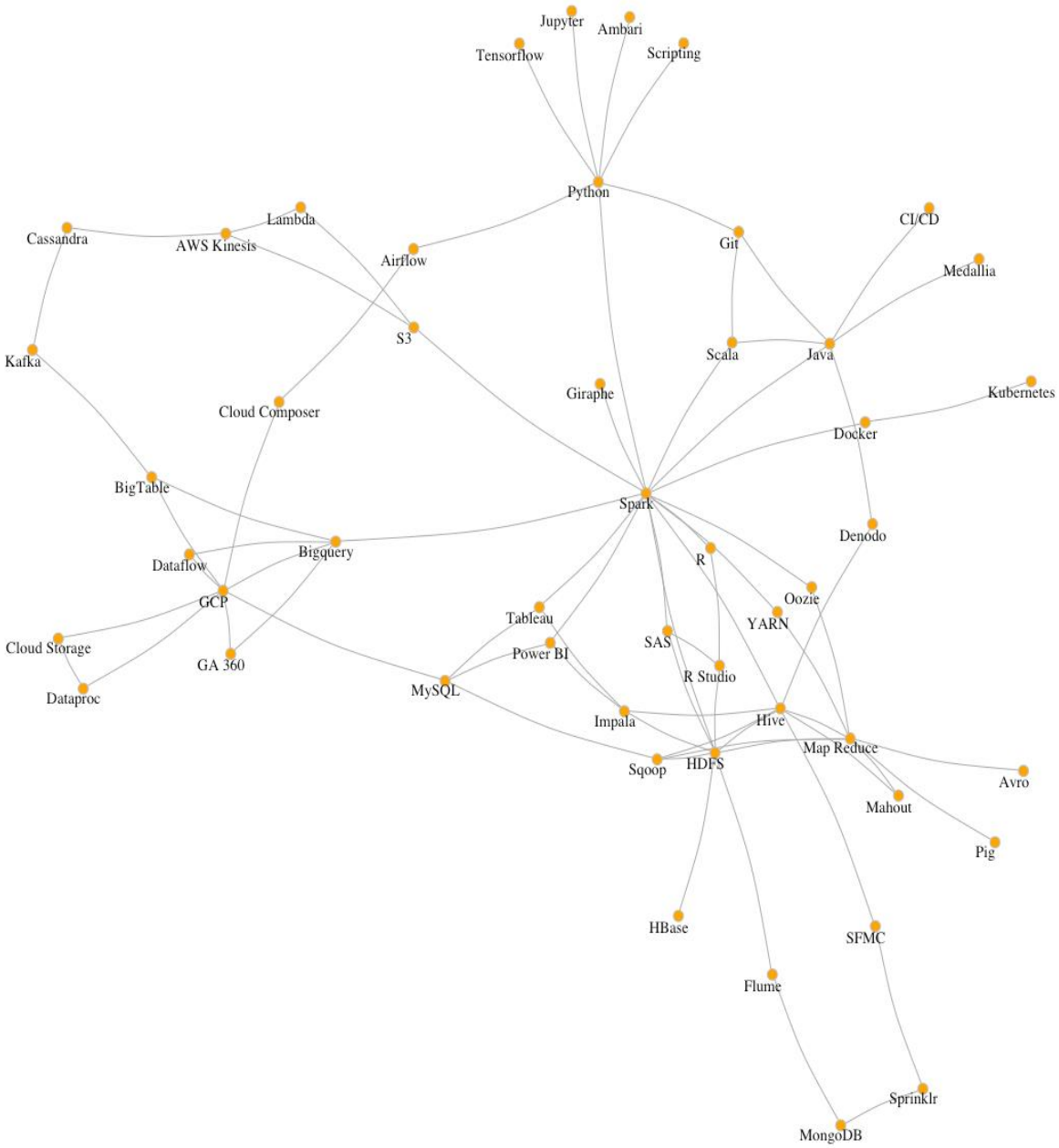
1. API call was available between them or
2. They are available under a suite (for example: Microsoft excel comes under office suite)

The dataset was prepared based on my experience in big data space.

Analysis

- ◊ The below graph represents the relationship between the technologies in big data space.
- ◊ The Graph is undirected
- ◊ Each node represents a technology and edge represents

Network of Big Data Technologies



Centrality Measures

Following centrality measures are calculated

Degree centrality here, demonstrates the technology which has maximum number of connectors available and is the most important tool in big data space. Spark has maximum degree centrality i.e. 9 and is fastest growing technology nowadays because we can use spark in a variety of ways such as extracting raw data, transforming data, pulling near real time data, building advanced machine learning models, connecting to a visualization tool and so on. Hence it is most important member of big data tech space.

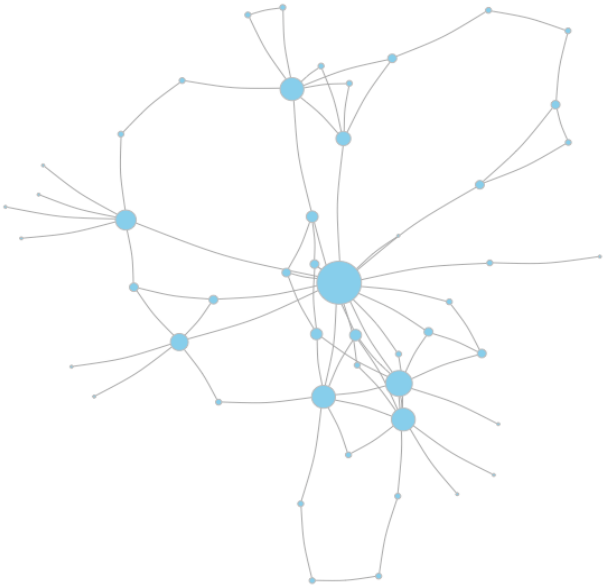
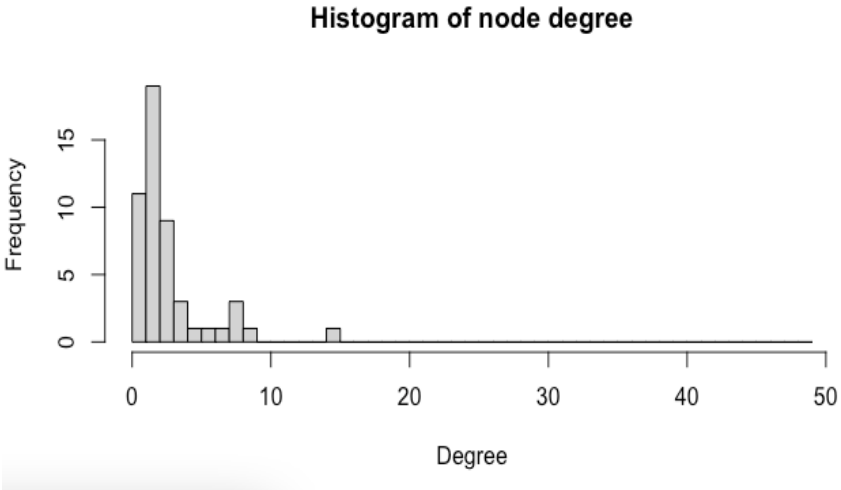


Figure 1 Degree Centrality Graph



Betweenness , Spark and Python have highest betweenness centrality here. As betweenness for each node is the number of shortest paths (edges) that pass through that node. As most coding specific work is done in python and spark code can also be written in python it is quite evident that they have high betweenness centrality.

Figure 2 Betweenness Graph

Closeness, Spark and Hive have more closeness score and are able to spread information very quickly and efficiently. As Spark is used for almost all purposes (extraction, transformation, ML) and hive works as a data warehouse and querying tool it is quite evident that they have high closeness scores

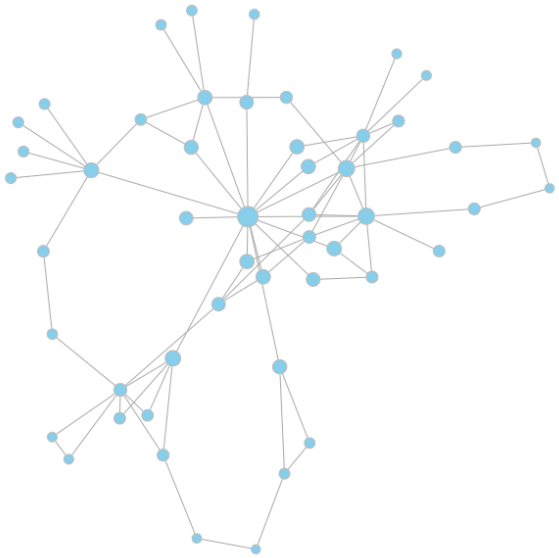


Figure 3 Closeness Graph

Eigenvector Centrality, Spark has the highest eigenvector centrality as expected since it is the most influential node in the entire network.

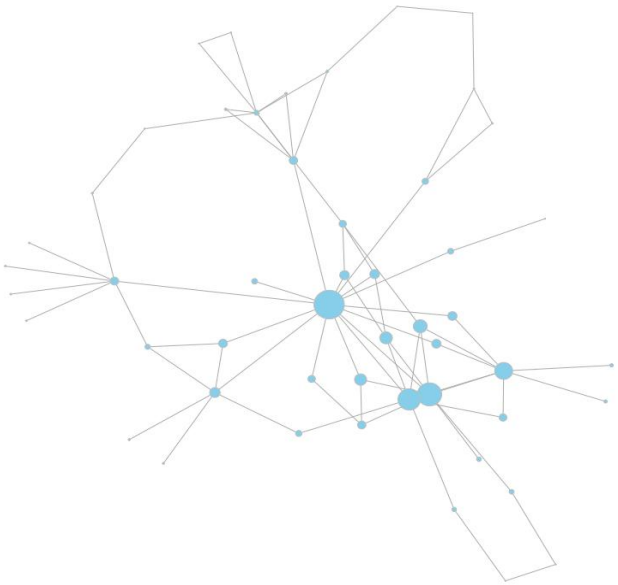


Figure 4 Eigenvector Centrality Graph

Summary of Centrality Measure

Degree Centrality		Betweenness		Closeness		Eigenvector	
Spark	15	Spark	786.90	Spark	0.0108	Spark	1.00
HDFS	9	Python	244.52	HDFS	0.0086	HDFS	0.80
Hive	8	Big query	218.27	Hive	0.0086	Hive	0.74
GCP	8	HDFS	204.06	Big query	0.0081	Map Reduce	0.59
Map Reduce	8	Hive	191.26	Python	0.0079	Sqoop	0.45

Correlation Matrix

	Degree	Betweenness	Closeness	Eigen Vector
Degree	1.0000000	0.8907434	0.7581356	0.8380747
Betweenness	0.8907434	1.0000000	0.7367934	0.7192906
Closeness	0.7581356	0.7367934	1.0000000	0.8732248
Eigen Vector	0.8380747	0.7192906	0.8732248	1.0000000

Correlogram

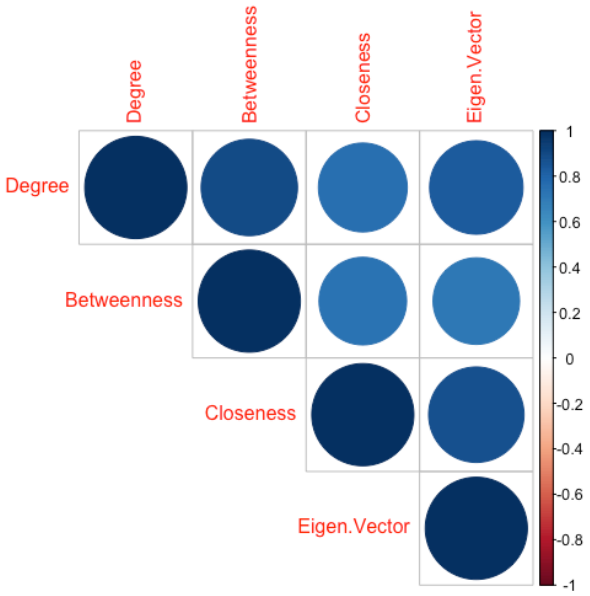
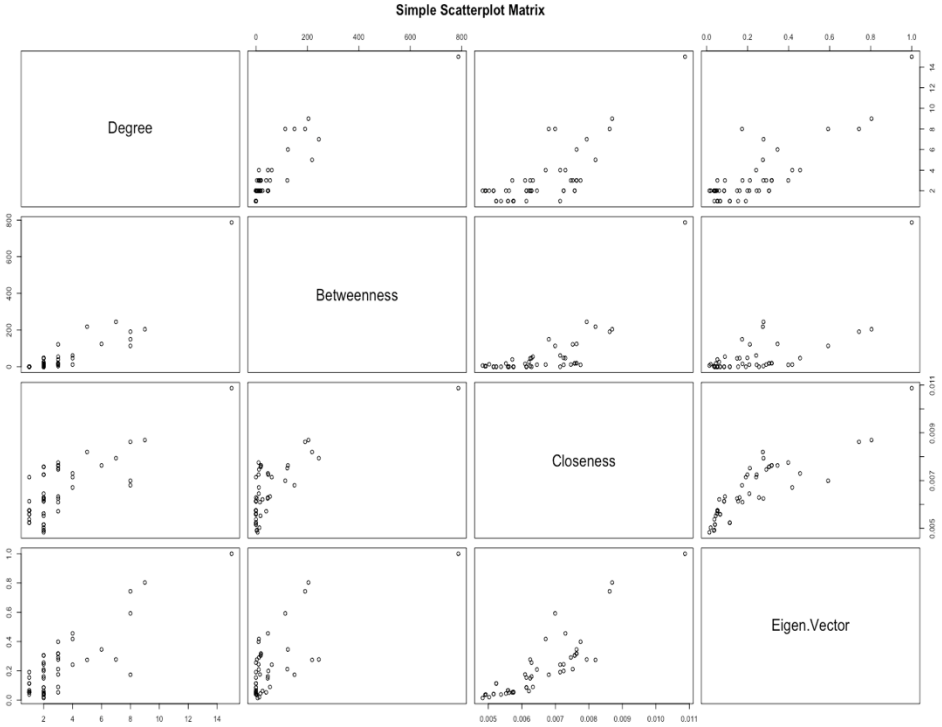


Figure 5 Correlogram of different centrality measures

Scatter Plot of Correlation



Comparison with Random, Small World and Scale Free Network

1. **Empirical Network:** Fruchterman and Reingold layout to the big data network

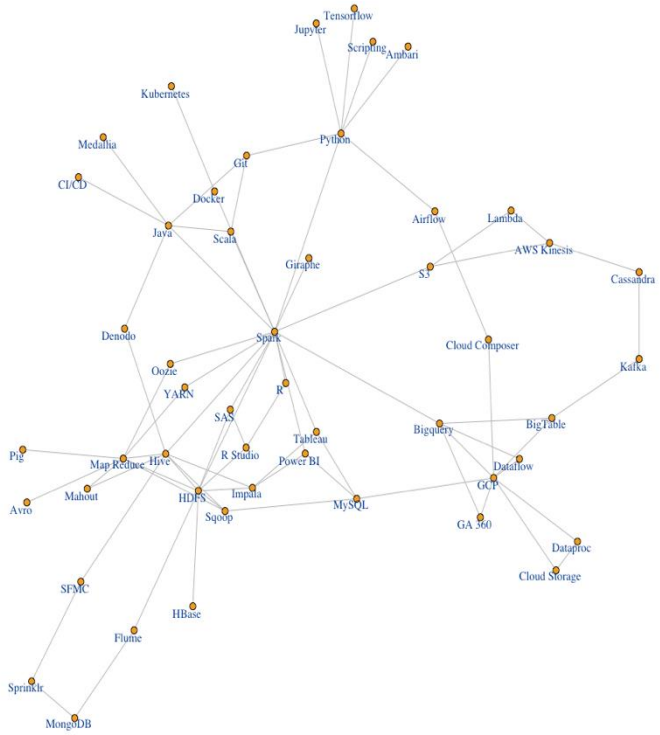


Figure 6 Big Data Network after applying Fruchterman and Reingold layout

2. Barabasi-Albert Network (a.k.a. Scale Free Network)

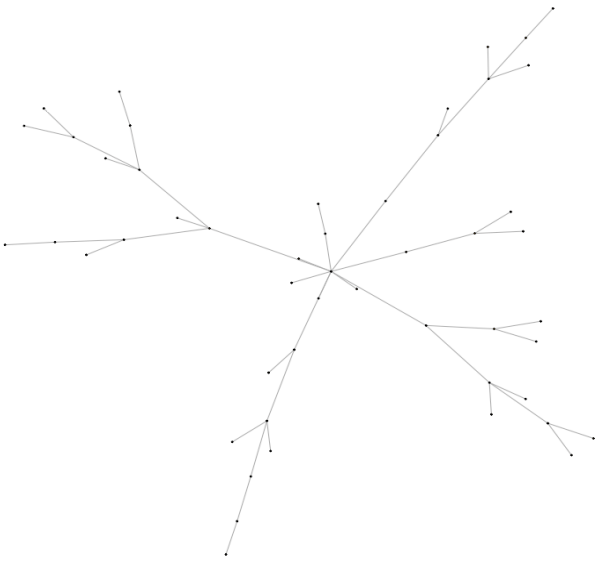


Figure 7 Big Data Network (Barabasi-Albert Network)

The Barabasi Albert Model is created with vertex = 50, which is same as our empirical network and with a power 1. This model depicts that vertex with varying degrees can exist together in same network, also there are hubs present in this network like our empirical network

3. Small World Model

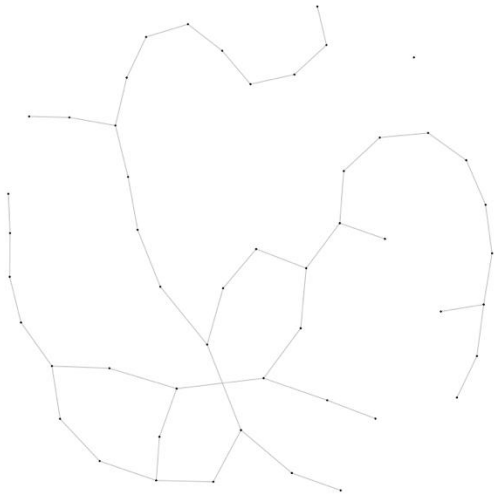


Figure 8 Small World Network

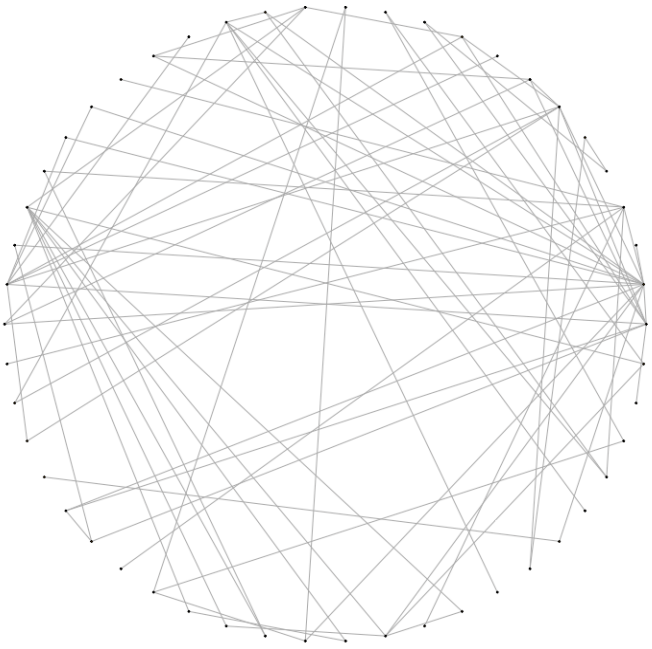


Figure 9 Circular Layout Big Data Small Network

4. Random Network

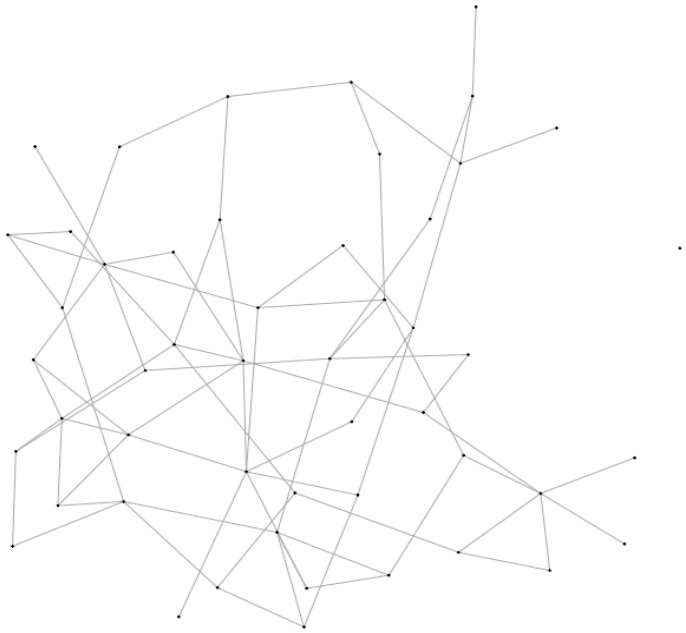


Figure 10 Random Network of Big Data Technologies

Conclusion

Measures	Empirical Network	Barabasi Albert Network	Small World Network	Random Network
Mean of Degree Centrality	3.08	1.96	2	3.08
Mean of Closeness Centrality	0.0065	0.0049	0.0007	0.0037
Mean of Betweenness Centrality	54.28	80.08	62.84	53.1414
Mean of Eigenvector Centrality	0.2150	0.1543	0.1932	0.3220
Centralization of Degree Centrality	0.2432	0.1640	0.0204	0.0800
Centralization of Closeness Centrality	0.4372	0.3201	0.0195	0.0968
Centralization of Betweenness Centrality	0.6356	0.7027	0.1401	0.1349
Centralization of Eigenvector Centrality	0.8176	0.8809	0.8404	0.6840
Average Path Length	3.2155	4.26	6.2984	3.4523

- To check small world phenomena in our empirical network, we compared average shortest path length 6.29 and log (number of nodes) 3.91, which are very distinct and far hence our empirical network doesn't resemble small world network.
- Comparing our empirical network with random network, there is a huge variation in centrality measures hence we can say our empirical network doesn't resemble Random Network.
- From centrality measures, it is evident that our empirical model resembles closely with Barabasi Albert Network Model.

GitHub Repository:

https://github.com/shivam-hadoop/Big-Data-Network-Analysis/

References:

1. https://en.wikipedia.org/wiki/Big_data
2. https://cloud.google.com/bigquery
3. https://cloud.google.com/solutions/smart-analytics
4. https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cdh_intro.html