Word Network

TERM PROJECT OF COMPLEX NETWORK THEORY

BY
DINESH PATHAK(02CS3010)
&
KUNDAN KUMAR(02CS3008)

ABSTRACT :

Several important characteristics of languages are based on complex network analysis, which provides tools for characterizing statistical properties of networks and explaining how they arise. This article examines some of the statistical patterns in syntax of Hindi, Bengali and English on the basis of a variant of "Word Co-occurrence Network". The various statistical properties of complex networks are used to find similarities and dissimilarities between these languages. Also a comparative study between the network and the properties of the languages has been done.

INTRODUCTION :

All languages share some universal tendencies at different level of organization: the phoneme inventories, the syntactic and semantic categories and structures, as well as the conceptualizations being expressed. At the same time there are also very deep differences between languages, and often universal trends are implicational. They are about the co-occurrence of features and not the features themselves. For example if a language has an inflected auxiliary preceding the verb then it typically has prepositions. There are also universal statistical trends in human languages such as Zipf's law, which is about the frequency with which common words appear in texts.

Language is clearly an example of a complex dynamical system which exhibits a number of universal patterns in their structure. Recently, important advances in Graph theory, and specifically the theory of complex networks, have given a number of ways for studying the statistical properties of networks and for formulating general laws that all complex networks abide by, independently of the nature of the elements or their interactions.

In this term project we aimed at finding some statistical patterns in syntax of Hindi, Bengali and English, and also similarities and dissimilarities between these languages. Comparison between languages was based mainly on network properties like degree distribution, clustering coefficients and assortativeness.

2. LANGUAGE NETWORKS

If network structure is a potential key for understanding universal statistical trends then the first step is clearly to define what kind of network is involved. It turns out that there are several viewpoints which can be basis of network analysis but in this term project we focused on network structure of the words which is one of the most important language elements. We narrowed our focus to take words as the fundamental interacting units partly because this is very common in linguistic theorizing but also because it is relatively straightforward to obtain sufficient corpus data to be statistically significant. Based on these we have built the following variant of co-occurrence networks:

A) Formation of Our Network

We formed the networks for the three different languages namely Hindi, Bengali and English where the distribution of the network is as follows:

Nodes : The nodes of the networks are words of the language.

An Edge between two nodes shows the co-occurrence of the respective words in a sentence. There is an edge between two words if they co-occur in a sentence.

Weight of edges: Weight of the edges is inversely proportional to the distance between the two words. The weight of all the edges is initialized with zero. If two words occur at a distance d (i.e. separated by d-1 words) in a sentence then the weight of the corresponding edge will be increased by 1/d. The final weight of the edge is the sum of all the weights.

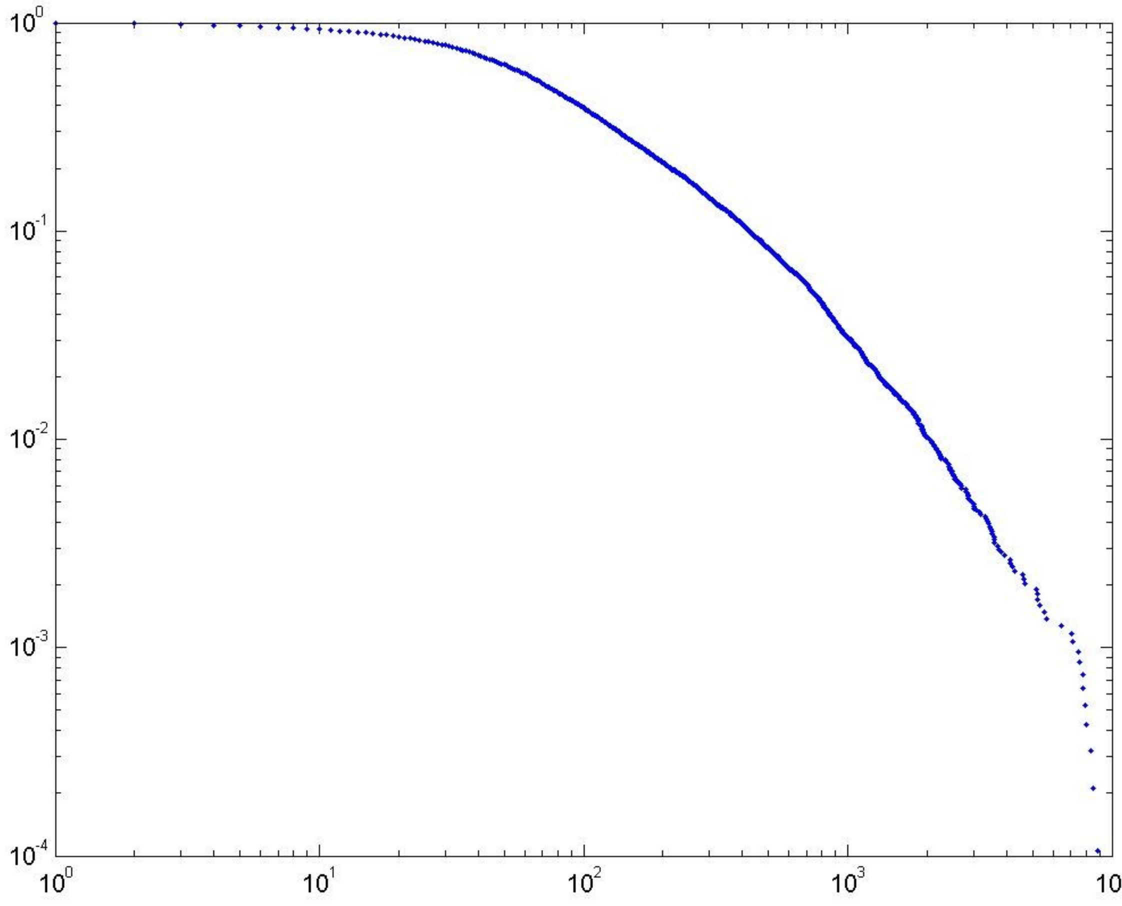
Weight of a node is sum of the weights of the edges having this node as the endpoint. (i.e. the weighted degree).

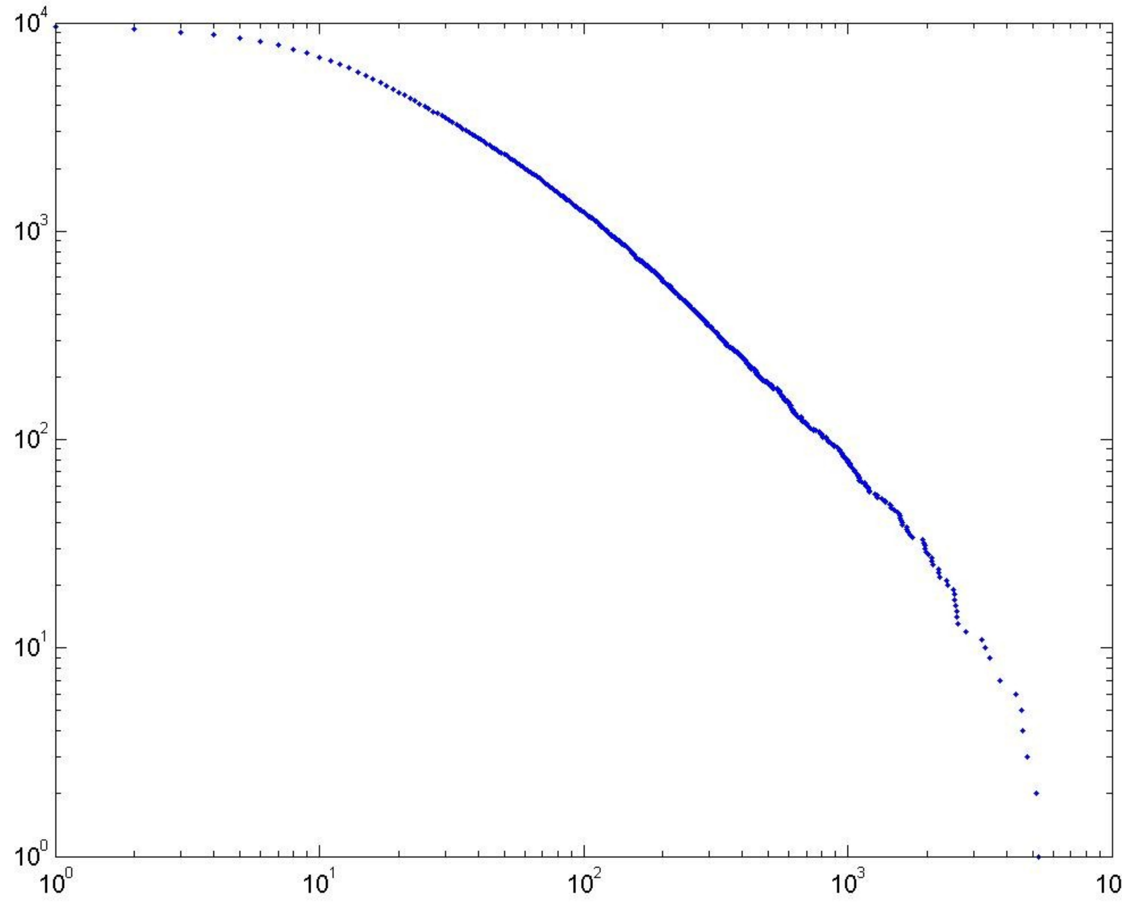
Size of the network: The number of nodes is 10000.The edges between them solely depends on the corpus.

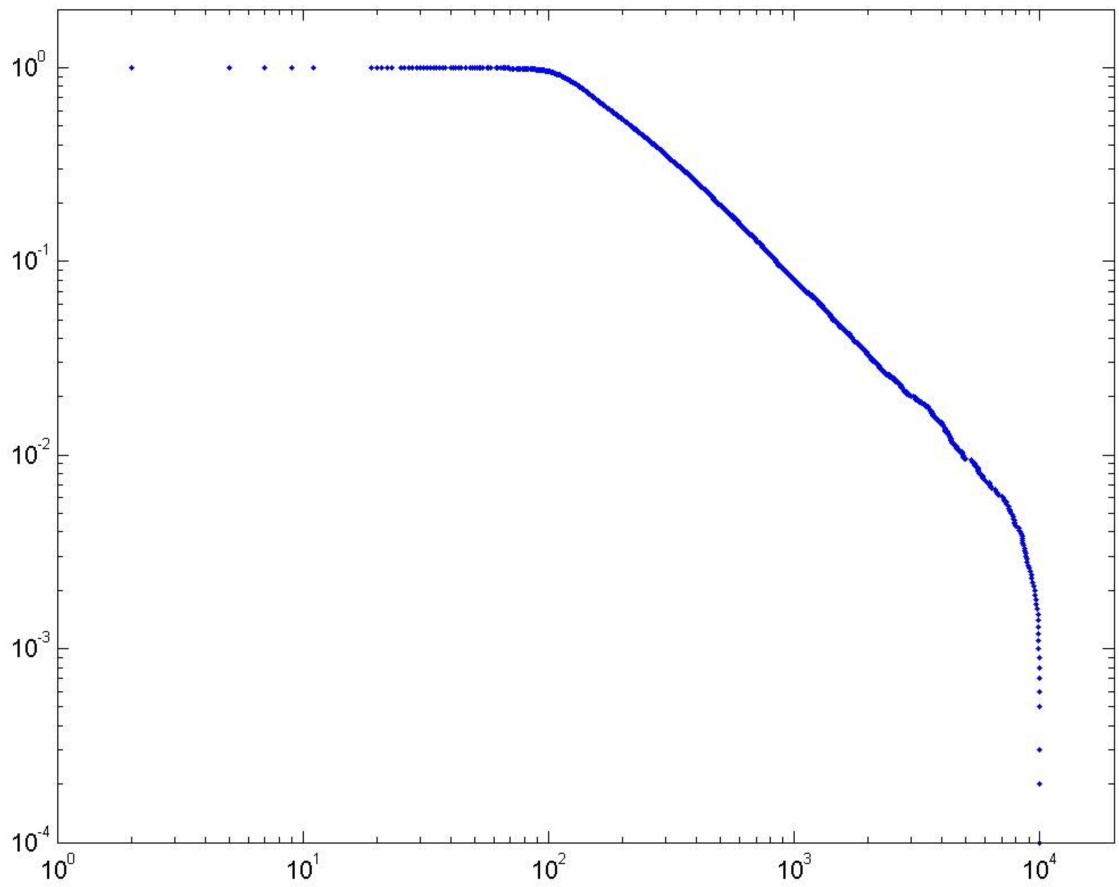
In addition to this we have plotted the graph again after removing the edges having value less than a threshold. The weight of all the edges remaining in the network was normalized and then the study if the network was done. Thus we have observed the changing behavior of the network.

3. Experiments and Plots

A. Degree Distribution


10^0
10^-1
10^-2
10^-3
10^-4
10^0
10^1
10^2
10^3
10^4


10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2
10^3
10^4



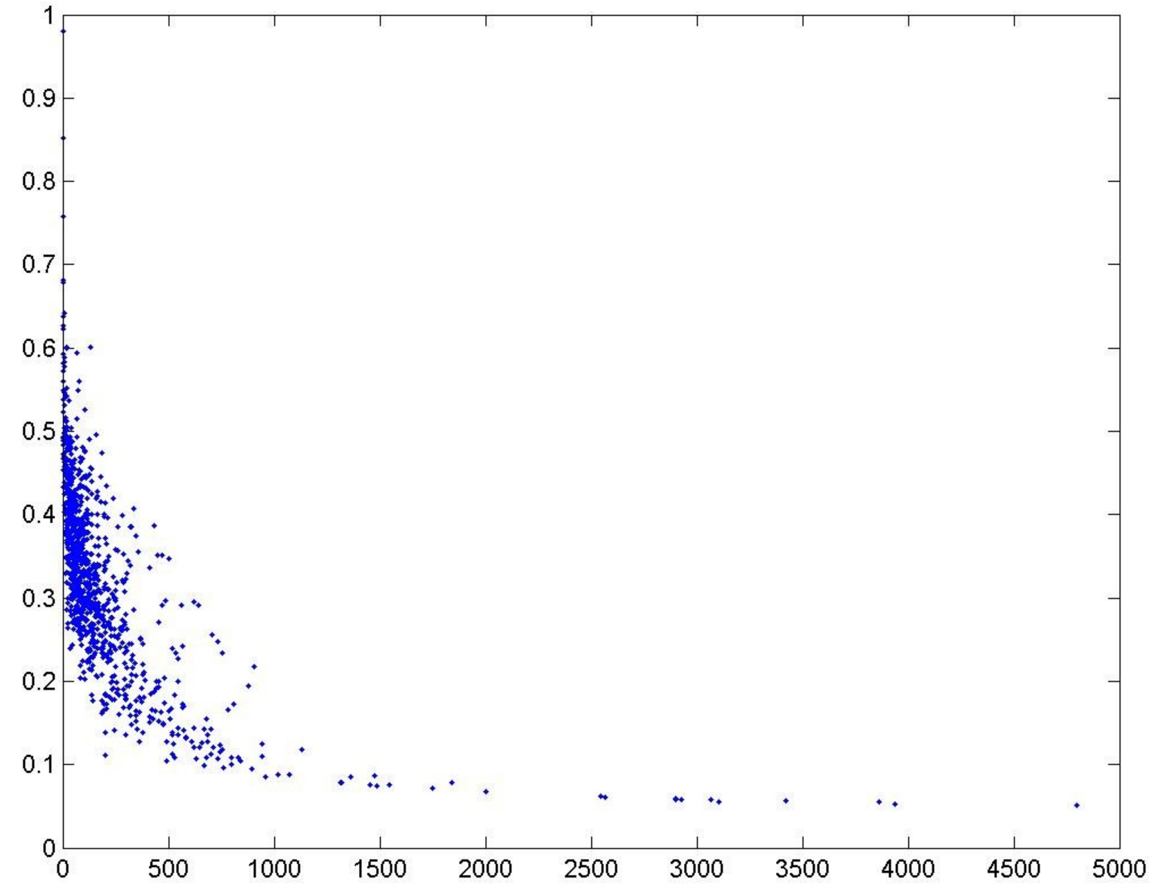
The above three figure show the log log graph of degree distribution for Hindi, Bengali and English Respectively. On X axis we have the degree and on the Y axis number of vertices.

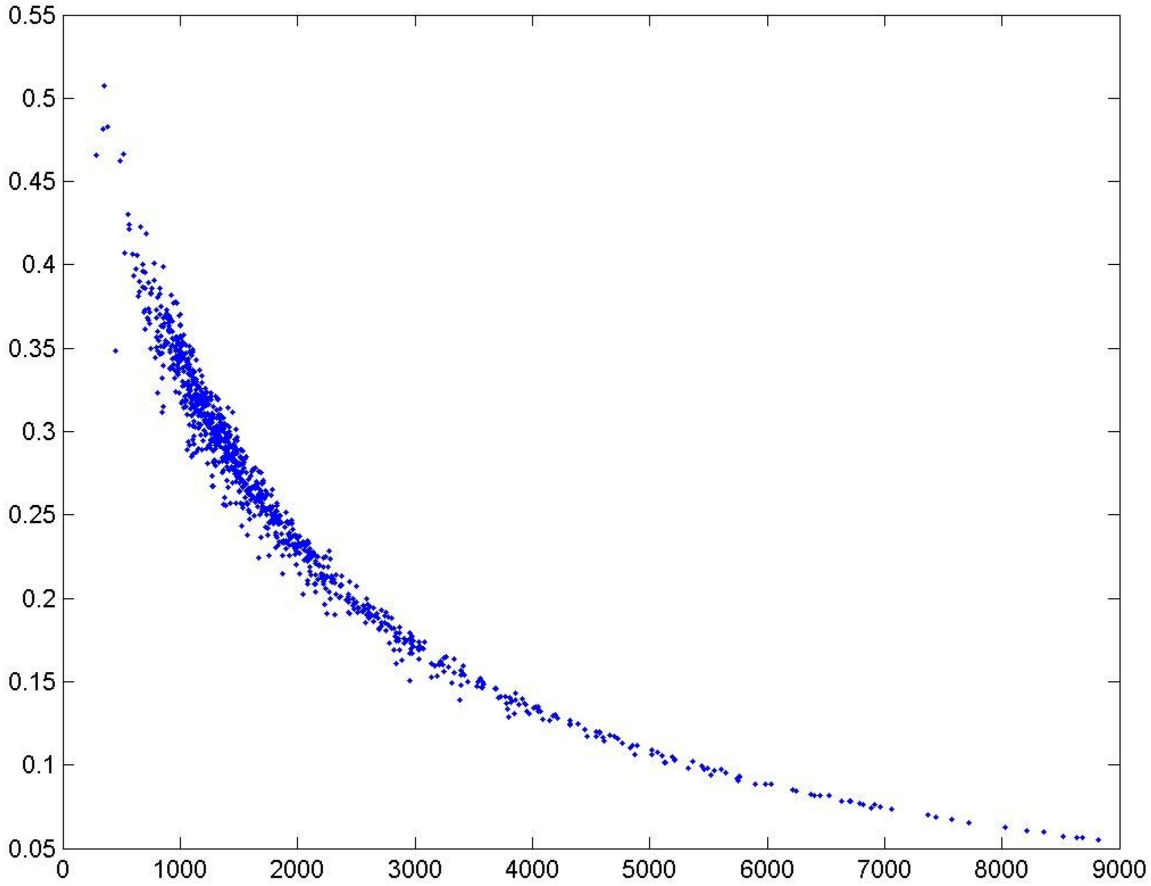
All degree distribution for all the three languages follow the power law.

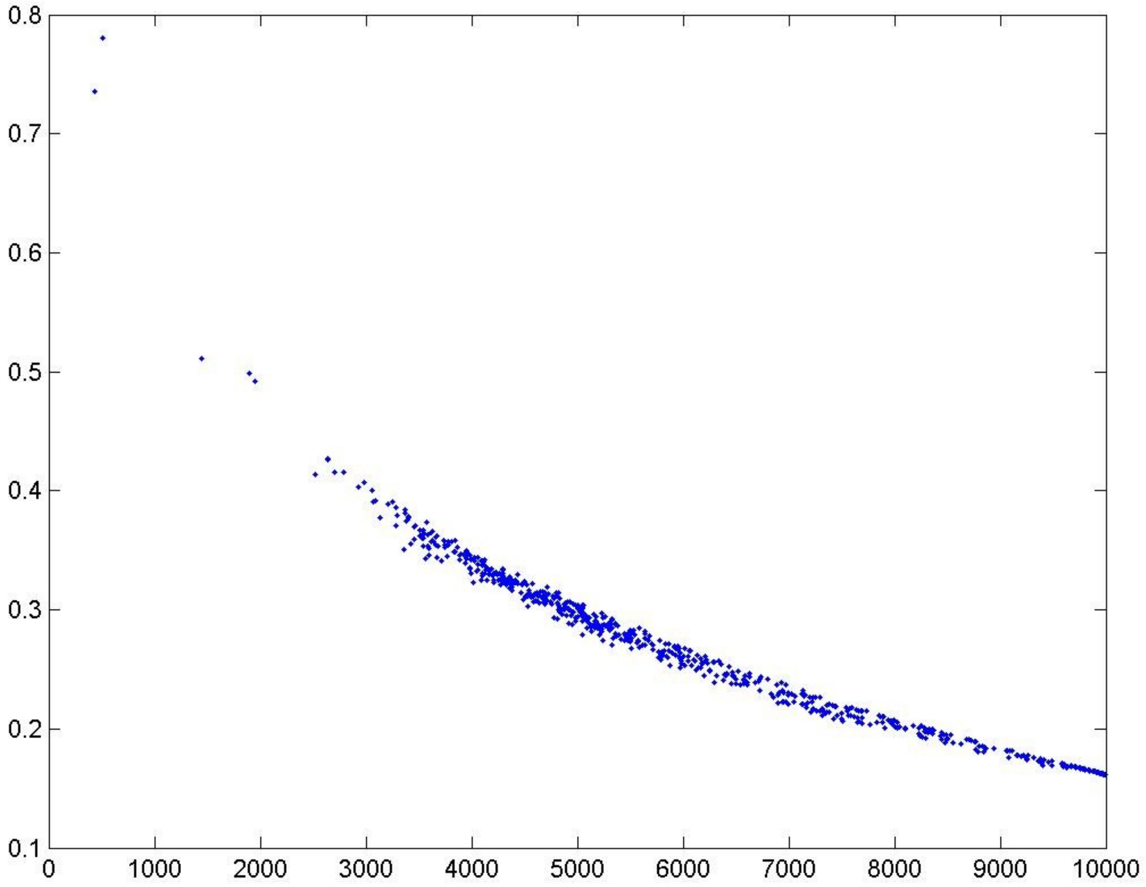
In case of English , the log-log plot of Degree Distribution has major kink.
In case of *Hindi*, the log-log plot of Degree Distribution has a minor kink towards the end.
In case of Bengali, the log-log plot of Degree Distribution does not have any significant kink.

The cause of the difference between Bengali and the rest is the use of linking words by joining it with the normal words to form a new word which results in the lesser use of independent linking Words in Bengali. Hindi also shows the same properties in some cases. But in case of English it is still rarer.

B. **Cluster Co-efficients**


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000


0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1000
2000
3000
4000
5000
6000
7000
8000
9000



The above three figures show the graph of degree(X axis) vs clustering coefficient(Y axis) for Hindi, Bengali and English respectively.

The plot of Degree vs Cluster Co-efficients shows exponential inverse relation between them. This relation is clear in English and Bengali as the band of values is small. But in case of Hindi band of values for a degree is much higher.

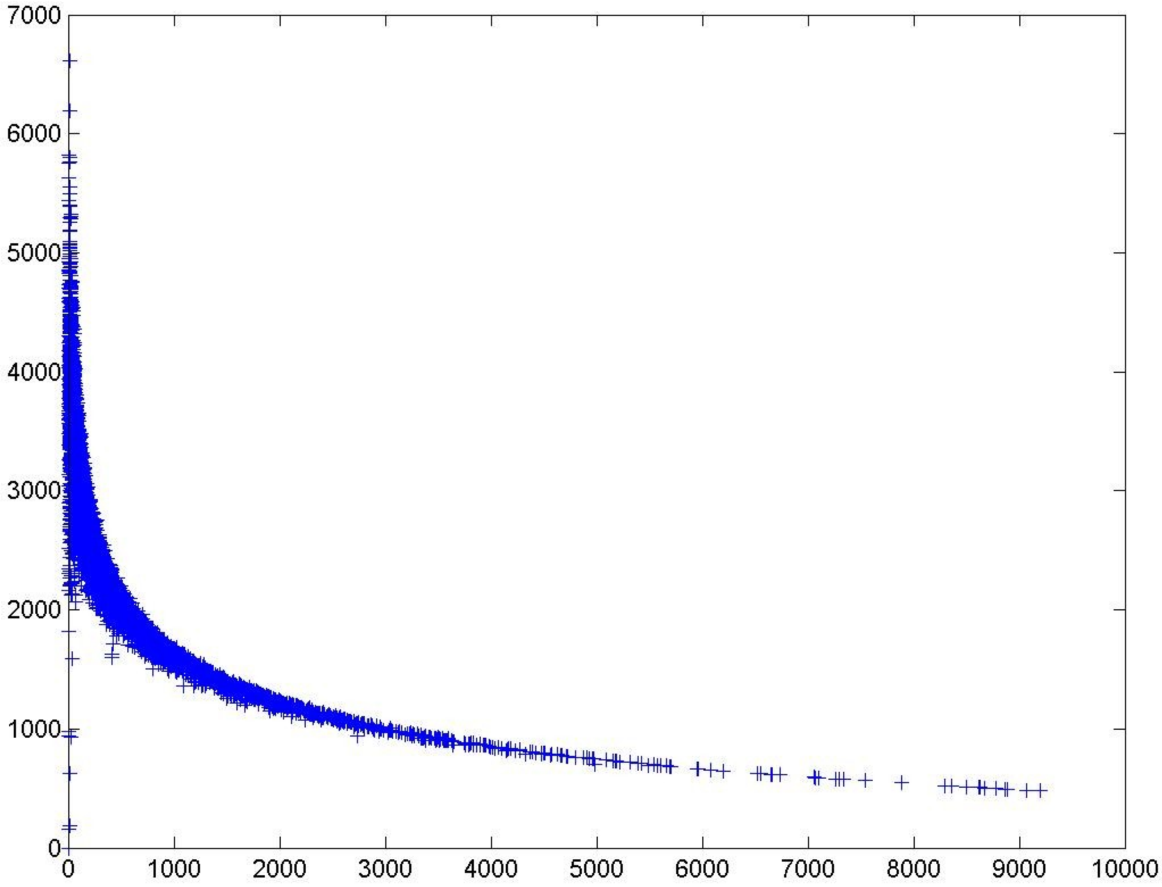
The reason of the above phenomenon must be the words like **jalstara** which has a relatively high frequency and low degree and hence are used with the words which are also used among them. **jalastara** is mainly used with the words like nadi, vridhi,etc. which are also used among themselves. Also some of the rarer words incase of Hindi are used with rarer words with relatively high frequency. This may not be in the case of Bengali and Hindi.

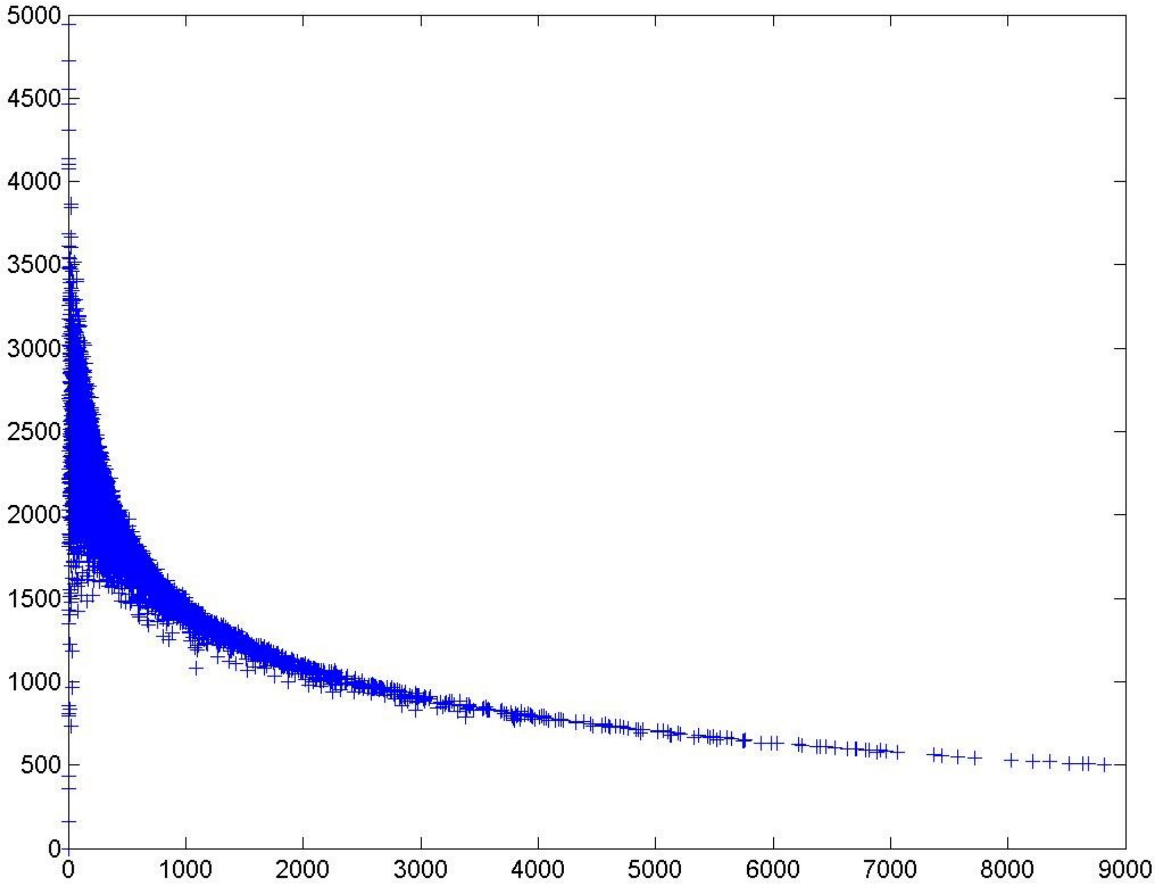
This plots also suggest more clustering in Hindi.

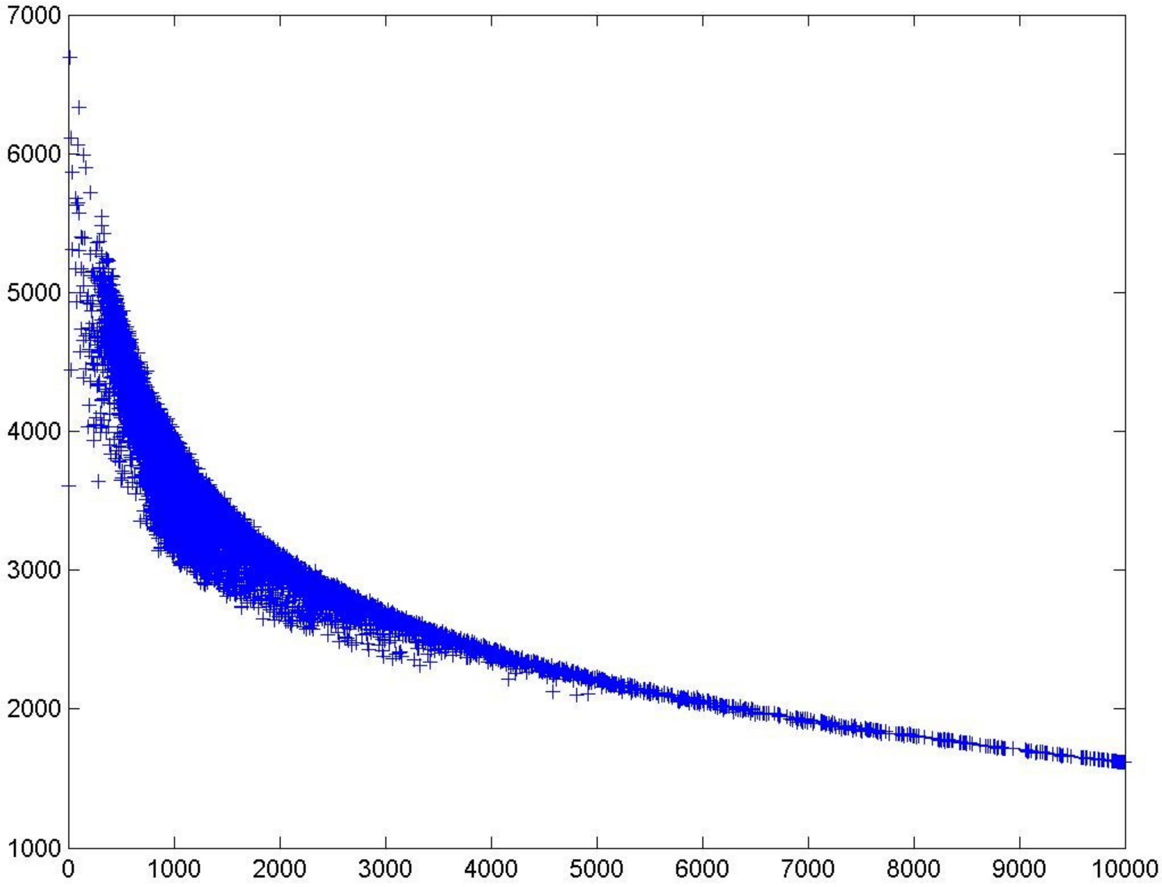
C. Assortativeness

Due to the lack of enough computational power and complexity of the standard formula, it could not be used. Therefore, we made two kind of plots:

(a) Degree vs mean of degree of adjacent nodes.

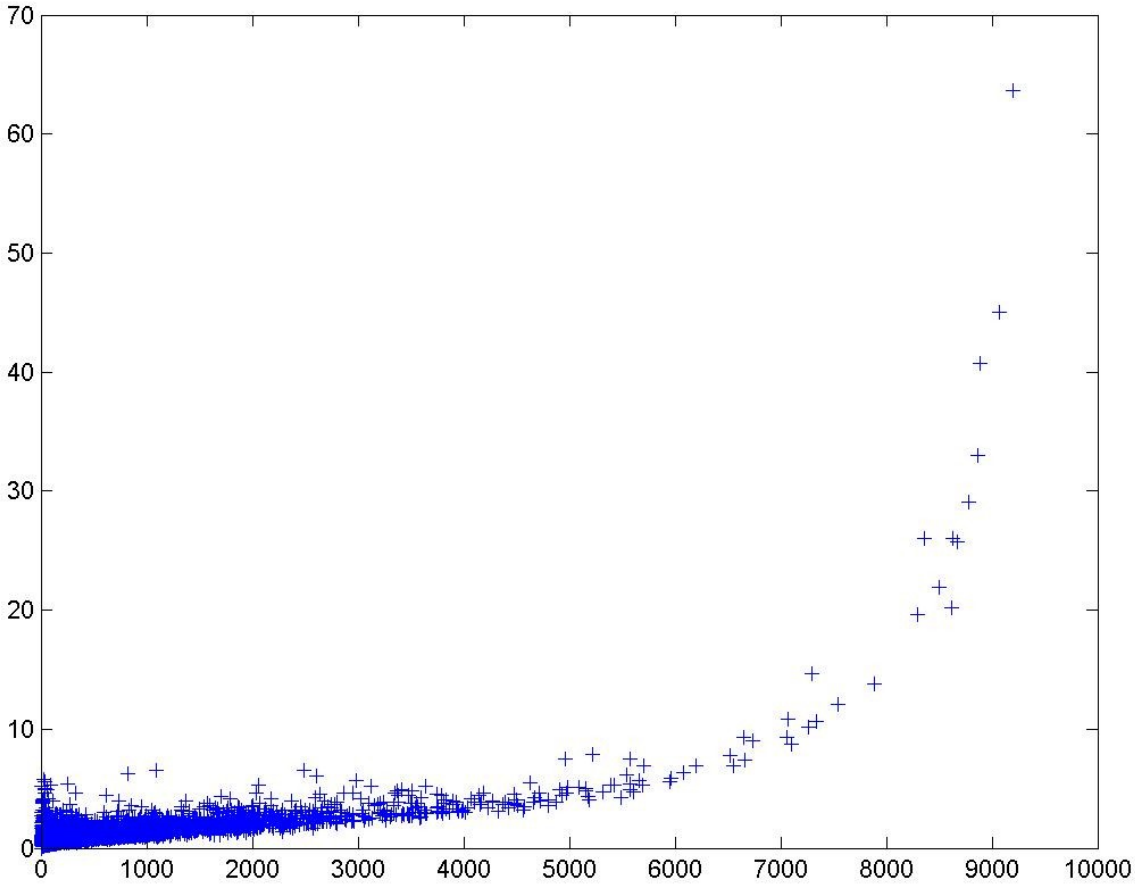

7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

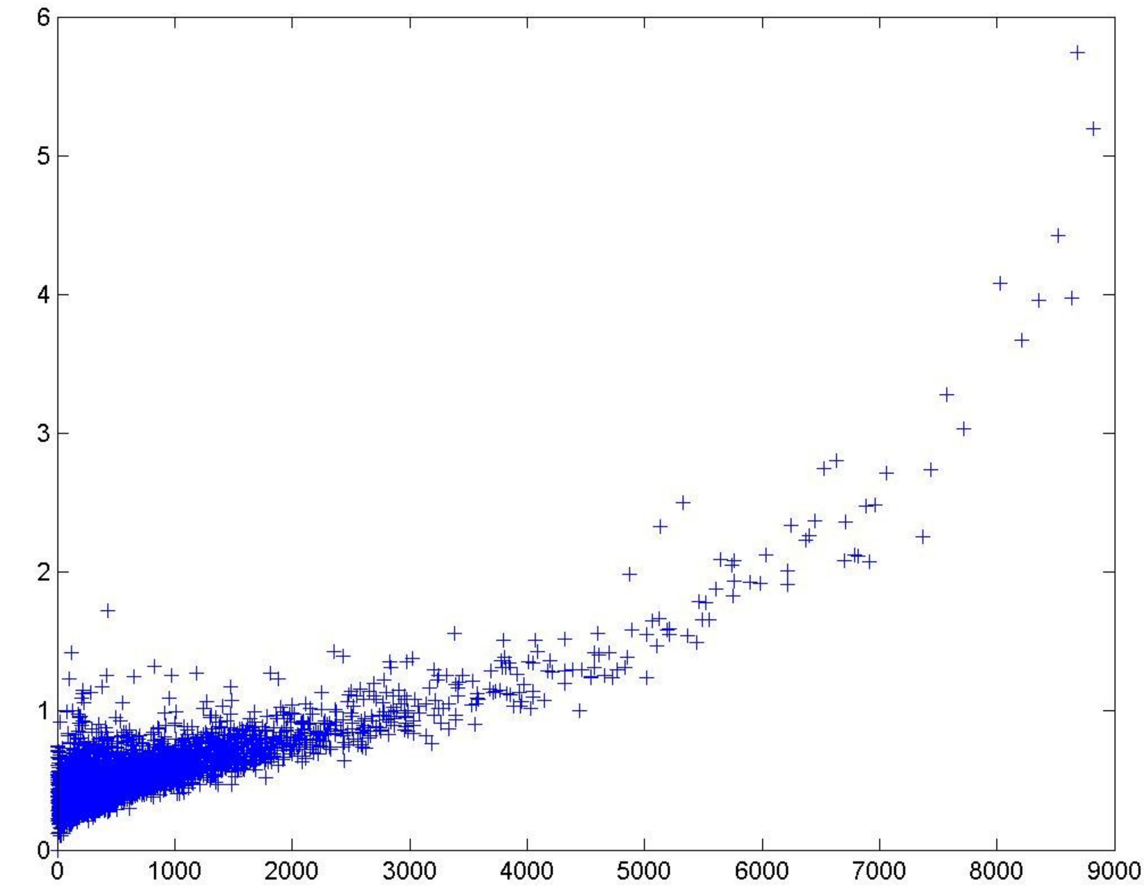

5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

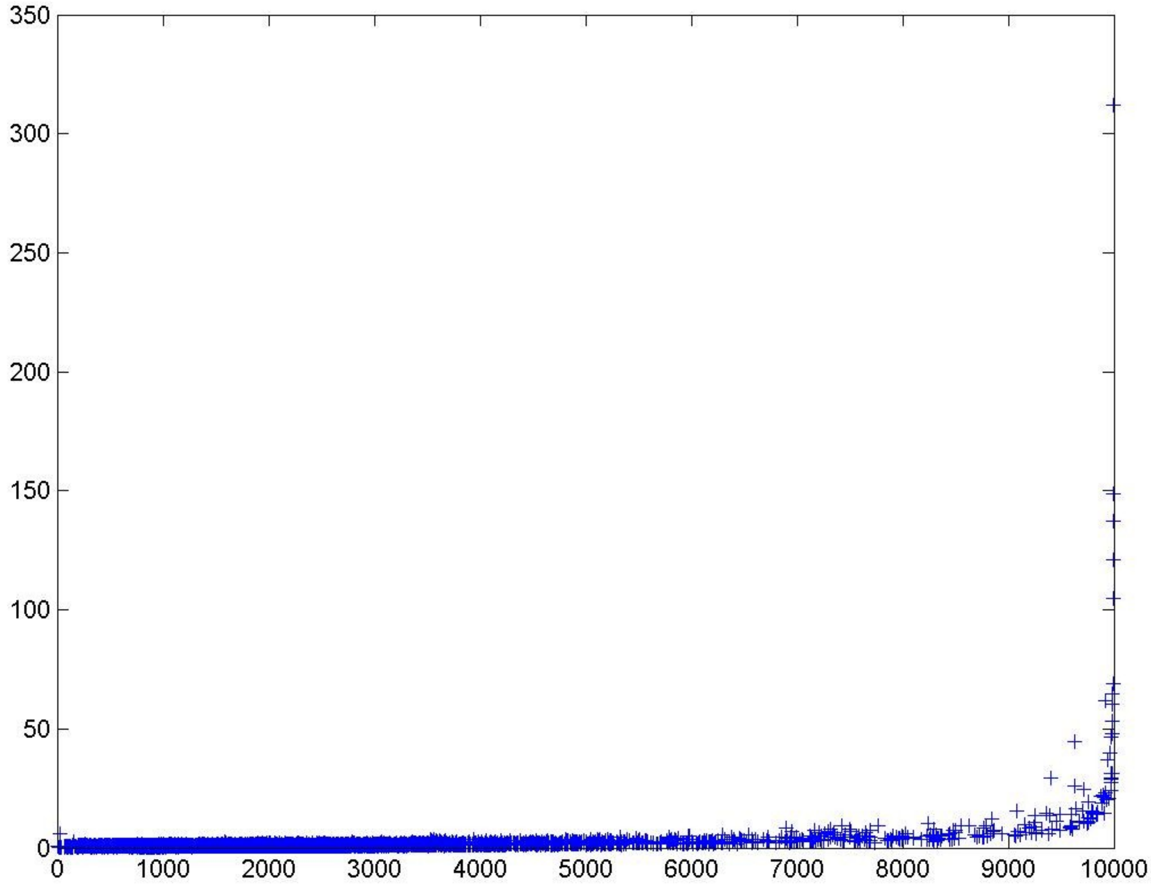


The above three figures show the graph of degree(X axis) vs Mean degree of neighboring vertices (Y axis) for Hindi, Bengali and English respectively.

(b) Degree vs mean of weight of connecting edges







The above three figures show the graph of degree(X axis) vs mean of the edge weights from that vertex (Y axis) for Hindi, Bengali and English respectively.

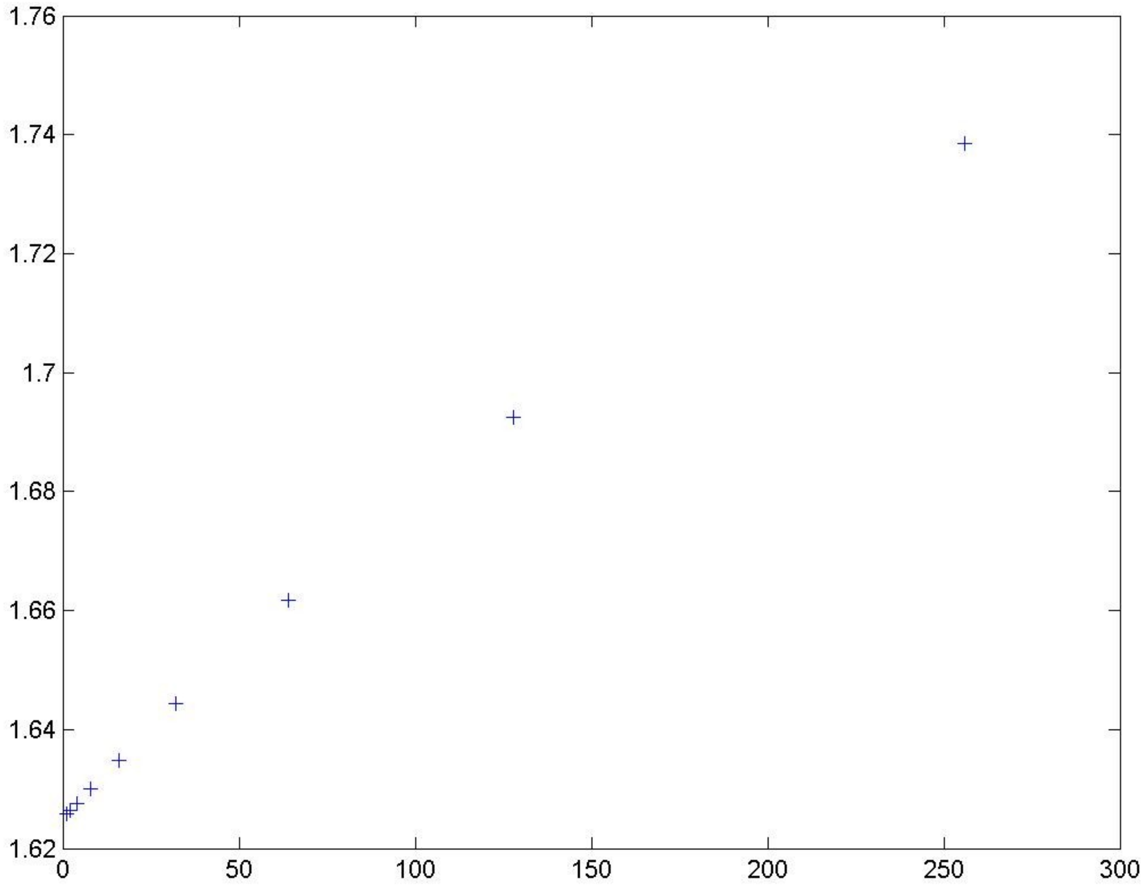
The plot (a) was similar for all the three languages English, Hindi and Bengali. The plot suggested lack of assortativeness.

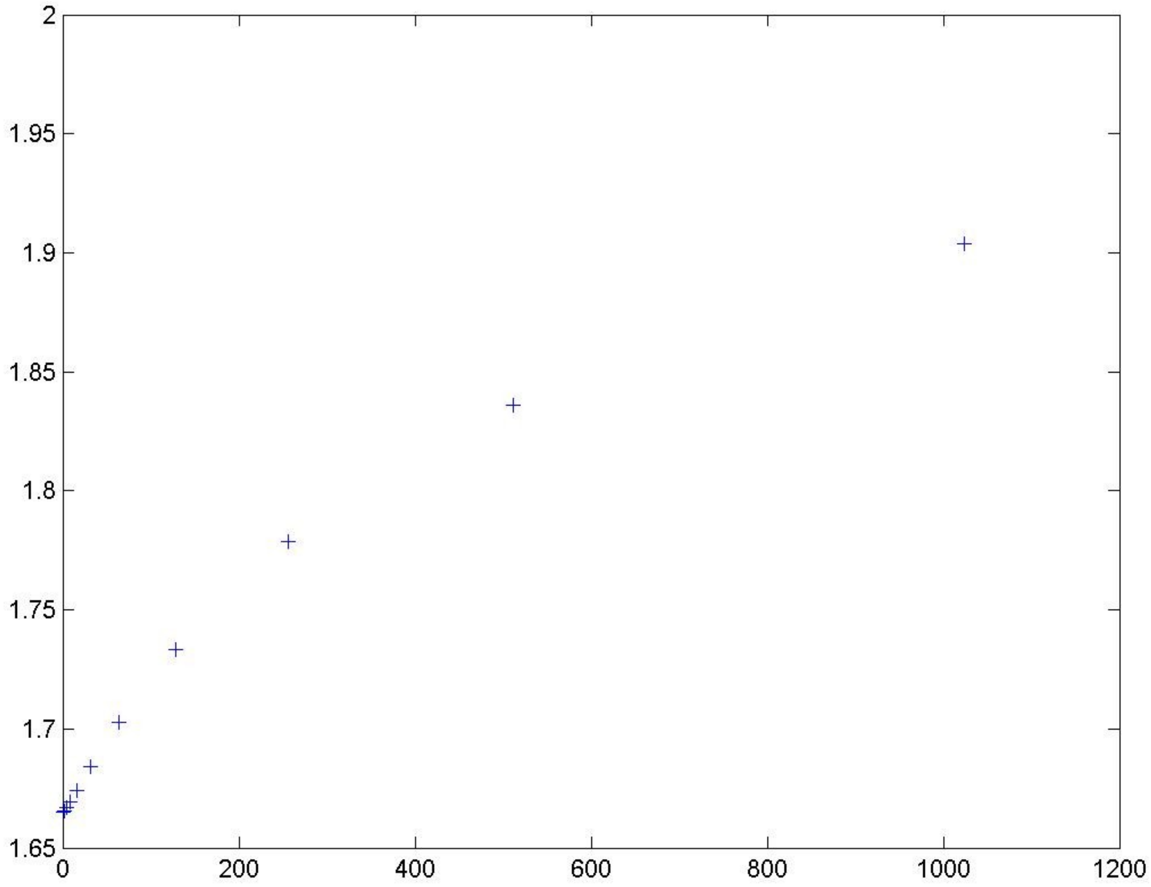
While plot (b) suggests English is less assortative than Hindi and in turn Hindi is less assortative than Bengali.

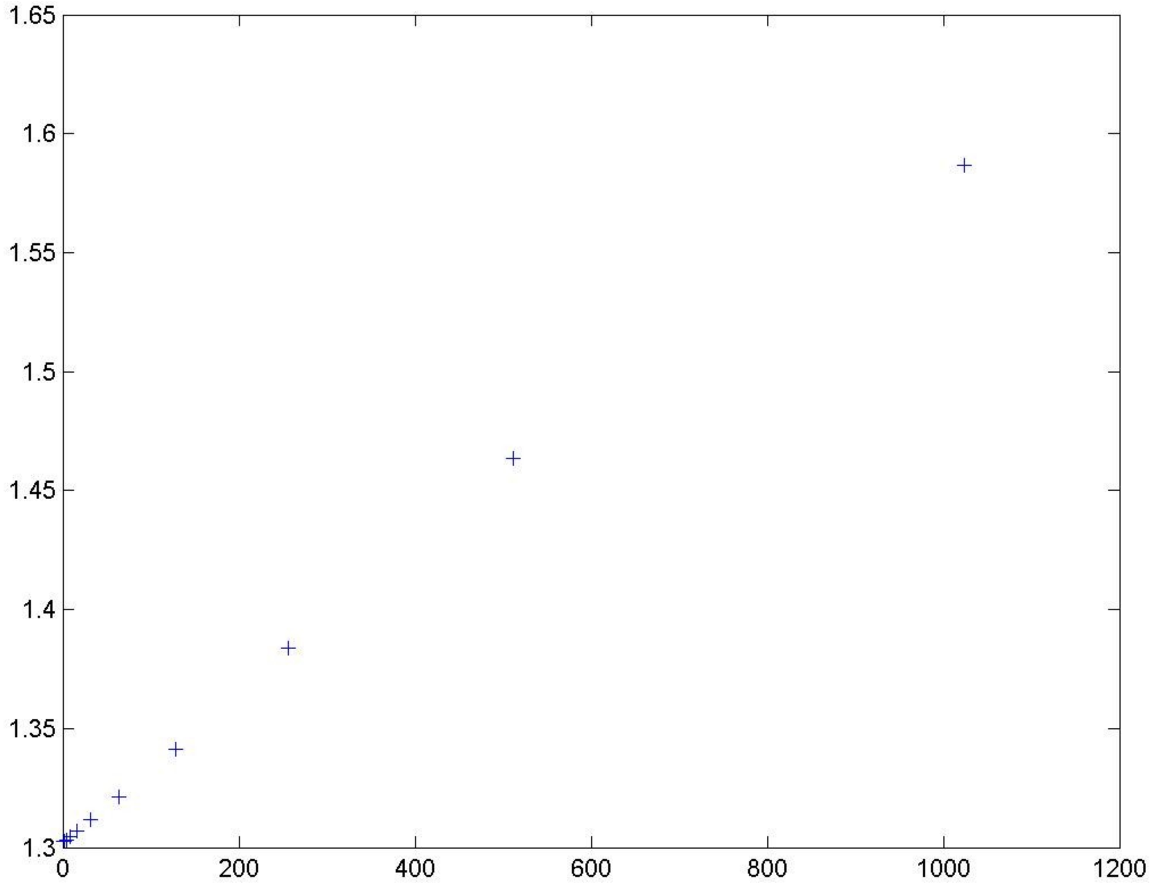
D. Diameter

The Diameter of Hindi and Bengali networks is 3 while the diameter of the English network is 2.If we start the most frequent occurring words then many frequent words are removed to increase the diameter. In fact, the diameter is increased by only 1 before the network becomes disconnected. The reason of small Diameter is the connection between every words occurring in a sentence. Hence most of the non-linking words are connected through linking words making shortest path 2.Similarily two non-adjacent linking word is linked through a non-linking word. The shortest path of length 3 may be between a linking word and a non-linking word which are not adjacent. Even removing the high frequent words does not alter the diameter much increasing it by at most 1 before making the network disconnected.

Average Shortest Path







The above three figures show the graph of number of vertices removed(X axis) vs Average shortest Paths(Y axis) for Hindi, Bengali and English respectively.

Average shortest path for English is 1.3 and for Hindi it is 1.6 and for Bengali it is 1.66 for 2100. On removing vertices shortest path increases linearly before the graph becomes disconnected.

4.DISCUSSION

This article argued that there are statistical universals in language networks, which are similar to the features found in other complex networks. It points to new types of universal features of language, which do not focus on properties of the elements in language inventories as the traditional study of universals but rather on statistical properties. Second, the pervasive nature of this network features suggests that the three languages must have been subjected to the same sort of self-organizational dynamics. The study of co-occurrence networks of the words and the identification of their universal statistical properties provides a tentative integrative picture. The explanation of universal statistical network features is even more in its infancy. There are several reasons for this. First of all the explanation of these features generally requires that we understand the forces that are active in the building of the networks. Second we must develop more complex models.

Different nontrivial traits have been identified which are as follows:

(1) Non-assortative mixing in our word co-occurrence network tells us that labor is divided in human language.

(2) Small world ness is a necessary property by construction of the network

(3) Regardless of the heterogeneity in the languages, common patterns have appeared, in all the three languages. Apart from small differences in the plots, they are almost same. Hence, the plots presented here are candidates for linguistic universals. The minor differences are due to different types of word joining (like vibhakti in Bengali with content words) and the range of frequencies covered by the high frequency words in the languages.

(4) The number of triangles (and hence the clusters) are few in comparison to the number of nodes.