Graphemic and Phonemic Networks

(Project 3)

Project Report

Course Instructors-

Prof. Niloy Ganguly

Monojit Choudhury

Animesh Mukherjee

Submitted by-

Ankur Jain (01CS3016)

Markose Thomas (01CS3019)

Table of Contents

Section **Page No.**

1. Introduction

In this paper, we build and analyze the Graphemic and Phonemic Networks for the three languages – English, Hindi and Bengali using the complex networks tools. There are several predictions made, by researchers all over the world, on the evolution of these languages since the time they came into existence. Our aim is to verify the validity of some of these conjectures, based on the results obtained using the analysis tools. For example, a popular theory in the language evolution is that the lexicon is chosen such that it minimizes both the *Articulation Effort* required for various words as well as the *Confusion Rate* among the words of the lexicon. Articulation Effort, for a word *w* is the mechanical effort required by a human being to pronounce or spell *w*. Confusion Rate is the probability of confusing *w* with some other word *w'* while using it in the language. Both are dependent on the word length. However, for lesser Articulation Effort smaller word lengths are preferred, whereas for smaller Confusion Rate larger word lengths are preferred. Since both these functions are required to be minimized, these two are opposing functions and thus there exists a non-trivial optimal lexicon. Our results and inferences does not support the conjecture that lexicon is chosen such that Confusion Rate is minimized. We found out rather contrasting results which we will discuss in details later with other interesting observations.

This report is organized as follows. In section 2, we discuss the motivation behind studying these networks. Section 3 discusses the predictions that are made on these networks, prior to the actual analysis of the networks, based on the popular language evolution and complex networks theories. In Section 4, we have given the description and source of data used in this project. Then we describe how we have modeled these networks as graphs in Section 5. In Section 6, we present the details of the analysis done on the networks and the corresponding results obtained. Finally, we conclude this work and give some pointers to the future work in the Section 7.

2. Motivation

The results of this experiment can have many useful applications. For example, after studying the confusion probability of different words in languages like English, Hindi and Bengali we can tell which words are more frequently confused and thus need to be given special attention in building applications for Speech Recognition and Optical Character Recognition (OCR). Also, we can study the evolutionary aspects of these languages, as discussed in the Introduction section, using these two networks and comment on any similarity or dissimilarity in these aspects among these languages. For example, a popular theory states that in all the languages there is a core which was present since its beginning and around which a peripheral area of words evolves with time. In the field of language evolution and complex networks, there are certain theories and conjectures, like the one suggested in the Introduction section, that we would like the Graphemic and Phonemic Networks to follow. We have enumerated some of those predictions in the next section which are later verified in Section 6.

3. Predictions on these Networks

Since our networks are a type of social networks, we expect to see them exhibit the features of social networks, like Degree Distribution follows Power Law ($y = Cx^{-a}$), Small World Effect etc. Also since the languages have developed through evolution, we expect the graphs to be self-similar or fractal in nature. A graph is said to be *Fractal* if it recursively constructed or is self-similar, that is it appears similar at scales of magnification. Also, if we find out communities or clusters within these networks, they should have similar properties to those of the whole network like degree distribution etc. We also expect the network to support the evolution theory of two opposing functions – Articulation Effort and Confusion Rate acting on the lexicon. In order to reduce the confusion rate of heavily used words i.e. high frequency words, we expect large edit distances between high frequency words, that is high frequency should be mostly adjacent to low frequency words. In other words, the graph should be disassortative with respect to frequency in the high frequency region. This is because confusion rate of a word *w* is dependent on both the frequency of *w* and the frequencies of its neighbors and

also the edit distances between them. According to Vitevitch [1], at a same high degree, words (nodes) with high clustering coefficient are easier to learn than the words with low clustering coefficient. Since the words that are easier to learn are expected to have high usage frequency, thus this implies that nodes with high degree and low clustering coefficient should have low frequencies and nodes with high degrees and high clustering coefficient should have high frequencies. So, we expect our networks to support these predictions too.

4. Data Description

The word lists along with frequencies were already available for the Graphemic Network for Hindi and Bengali languages. The data for Graphemic Network for the English language is taken from the URL: http://www.audiencedialogue.org/susteng.html. It is a word list, along with frequencies, of the 15,000 commonest words taken from the British National Corpus (BNC): a count of 100 million words. However, for each language - English, Bengali and Hindi, we have considered only top 10000 words (frequency wise) so far. The words in Hindi and Bengali were first encoded to ***kgpiscii*** format. For Hindi, the network is similar for both Graphemic and Phonemic Networks. Thus, this data is also for the Phonemic Network for Hindi language.

5. Network Modeling

In these networks, each node represents a word. They are fully-connected graphs i.e. there is an edge between every pair of nodes and edge-weight is the edit-distance between that pair. Each node has a weight equal to its usage frequency. For graphemic network, it is the usage frequency in written English and for phonemic network it is the usage frequency in spoken English. However, in order to reduce the number of edges in the graphs we remove the unnecessary edges by limiting the edit-distance (the edge-weights) to an upper limit called *Edit Distance Threshold.*

6. Analysis using Complex Network Tools

We have analyzed the Graphemic Networks for all three languages and the Phonemic Network for the Hindi language using Complex Network Tools. We plotted the Word

Length Distribution, Degree Distribution, Usage Frequency versus Degree, Degree versus Usage Frequency, Betweenness Centrality versus Degree, Betweenness Centrality versus Usage Frequency, Clustering Coefficient versus Degree, Clustering Coefficient versus Usage Frequency, Average Frequency of the neighbor versus Frequency and Confusion Probability versus Frequency for these networks. We also found out the clusters within these networks and studied the cluster size distribution. For each cluster, we also plotted the Degree Distribution and Confusion Probability versus Frequency of the nodes in that cluster. From these numerous plots we gained some interesting insights about the nature of these graphs. We will now enumerate these results and the inferences made on them, along with the relevant plot.

6.1. Important Observations-

1. *The Degree Distribution of the graphs follows the power law*. Consider the plot of the Degree Distribution for the phonemic network of Hindi at threshold equal to 1 (Figure 1) and 3 (Figure 2). The straight lines in these log-log plots denote that this is a power law distribution.

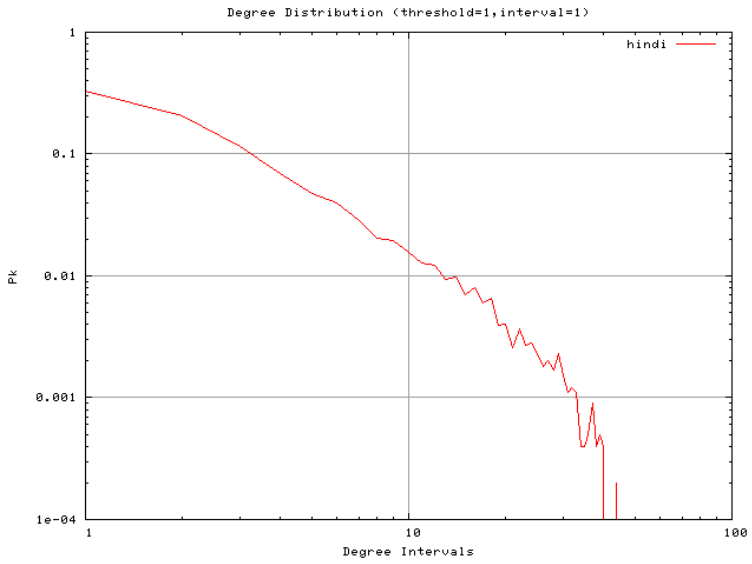

Figure 1: Degree Distribution for the phonemic network of Hindi at threshold = 1

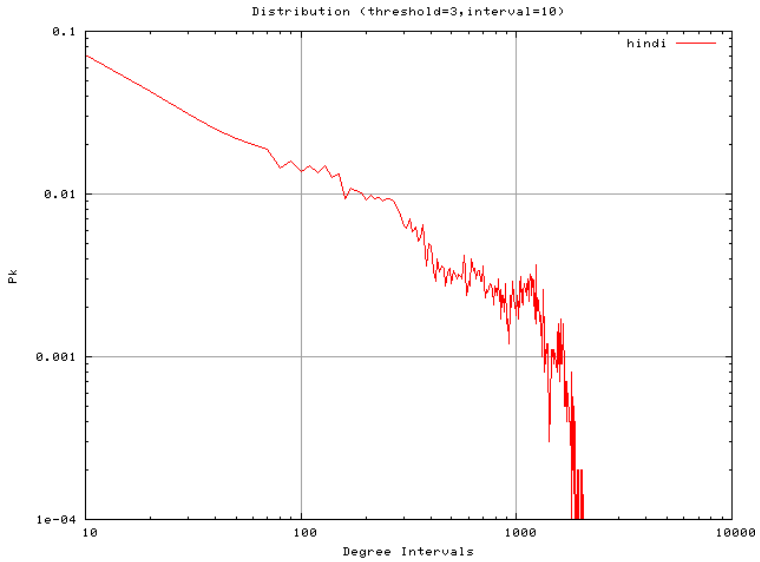

Figure 2: Degree Distribution for the phonemic network of Hindi at threshold = 3

2. From the Clustering Coefficient Vs Degree plots for the Phonemic Network for Hindi language at threshold = 1 (Figure 3) and 3 (Figure 4), we get that the clustering coefficient is more or less proportional to the degree. Thus, high degree nodes have high clustering coefficient and low degree nodes have generally low clustering coefficient. This implies that high degree words are adjacent to one another (in few very large clusters) and low degree words are adjacent to one another (in many small clusters). That is, the **graph is assortative with respect to degree**. This prediction is supported by the assortativeness coefficient of the networks. It is 0.70 for threshold = 1 and 0.36 for threshold = 3.

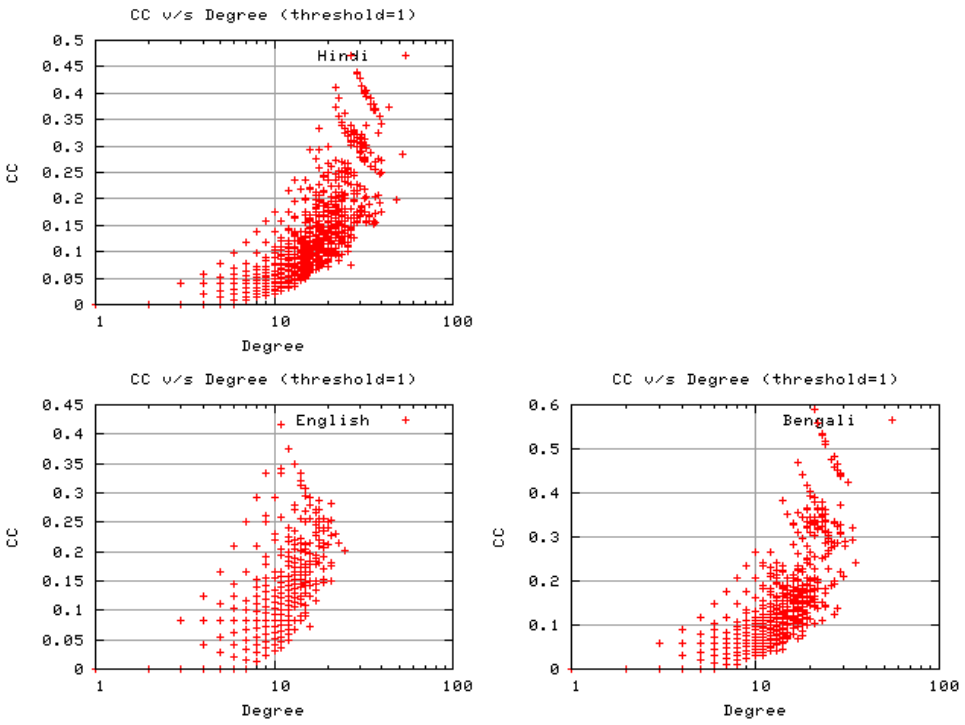

Figure 3: Clustering Coefficient Vs Degree for threshold = 1

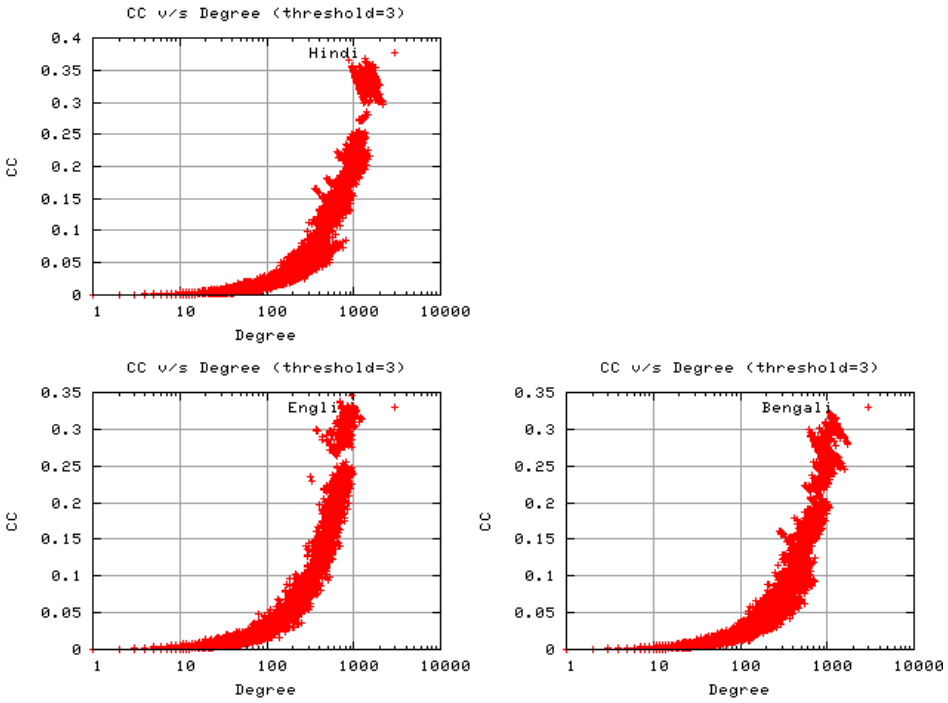

Figure 4: Clustering Coefficient Vs Degree for threshold = 3

3. In the above figure (Figure 3), we notice that if we split the graphs areas into 4 quarters, we will see that there are no or very few points in the quarters for 1. Low degree and high Clustering Coefficient and 2. High degree and low Clustering Coefficient. Thus, it is difficult to comment on the Vitevitch's prediction [1] that high degree and low clustering coefficient implies low frequency for a node.

4. Our basic hypothesis was that language evolves in such a manner such that confusion between words is minimized. Let the probability of confusion between two words W_i and W_j be Pr(i,j). So we have, Pr(i,j) is proportional to F_i, F_j, and 1/E(i,j), where F_i is the frequency of W_i, and E(i,j) is the edit distance between W_i and W_j. For a word W_i, given its frequency F_i, the probability of it getting confused with rest of the lexicon is given by-

$$\Pr(i \mid F_i) = \sum_j G(F_j, 1/E(i,j)) ,$$

Here, G is some function of F_j and E(i,j). Therefore we cannot directly claim that high frequency words should be connected to low frequency ones for low confusion rate. This is since the degree of the high frequency word (the value of *j*) is not considered. Also, it does not support high frequency words cannot be adjacent to other high frequency words. That is possible if the edit distance is high, thereby keeping the confusion rate still low. Thus, **our hypothesis should be for any high frequency word W_i, Pr(i | F_i) should be relatively small so as to keep the confusion rate low.**

We have verified this hypothesis by plotting Pr(i | F_i) Vs F_i (Figure 5). This is more accurate and informative than the Average Frequency v/s Frequency plot. It was plotted for various thresholds. **The observations, however, do not support the hypothesis. We get that high frequency nodes have high values of Pr(i | F_i).** The function G(x,y) was taken to be x'/y^2 , where x' is the normalized x. If we consider these plots at the cluster level within the network, we will make similar observations. First let us discuss the details of the clustering done. The network has many disconnected components (at threshold = 1). So, we ran a clustering algorithm on the largest component, of size 4273. We obtained a total of 106 clusters, with the mean cluster size of 40.3, minimum size of 2, and maximum size of 879 nodes. The clustering algorithm used is the hierarchical agglomeration algorithm (presented in [2]). As mentioned earlier, for the clusters at threshold 1, we found the plots of Pr(i | F_i) v/s F_i, to have almost the

same nature as of the entire network (see Figure 6 and 7). **Thus, the network is not disassortative with respect to frequency in the high frequency region.**

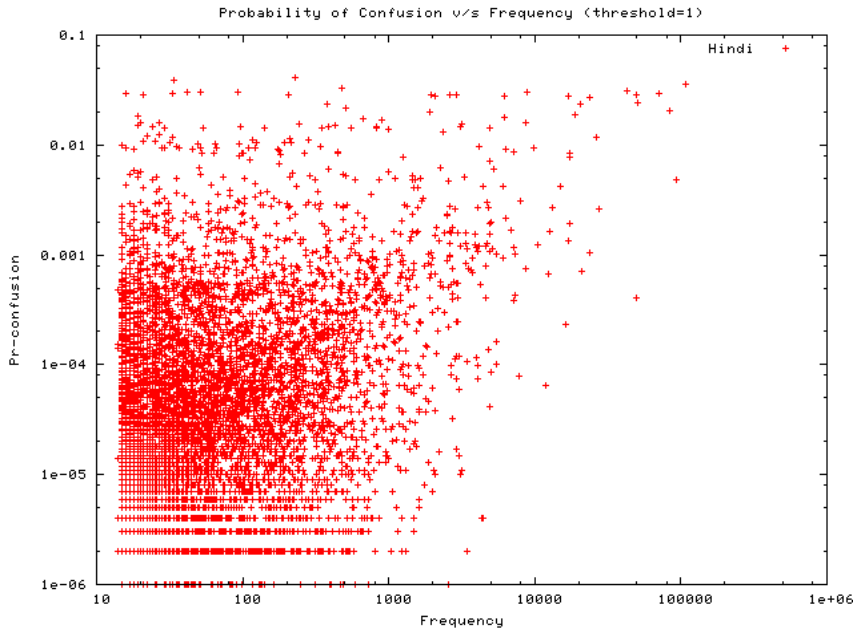

Figure 5: The confusion probability vs frequency for the whole network (Threshold = 1)

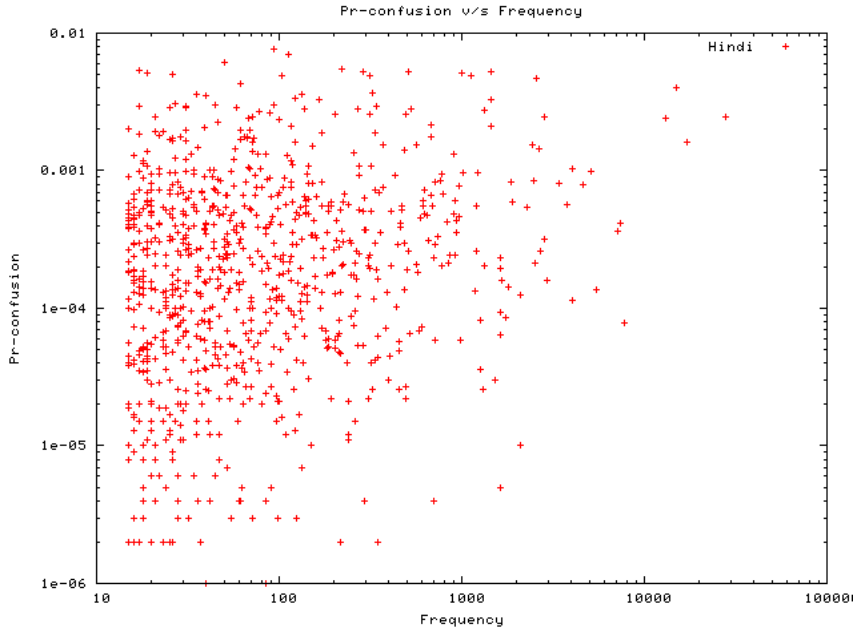

Figure 6: The confusion probability vs frequency for the largest cluster of 879 nodes (Threshold = 1).

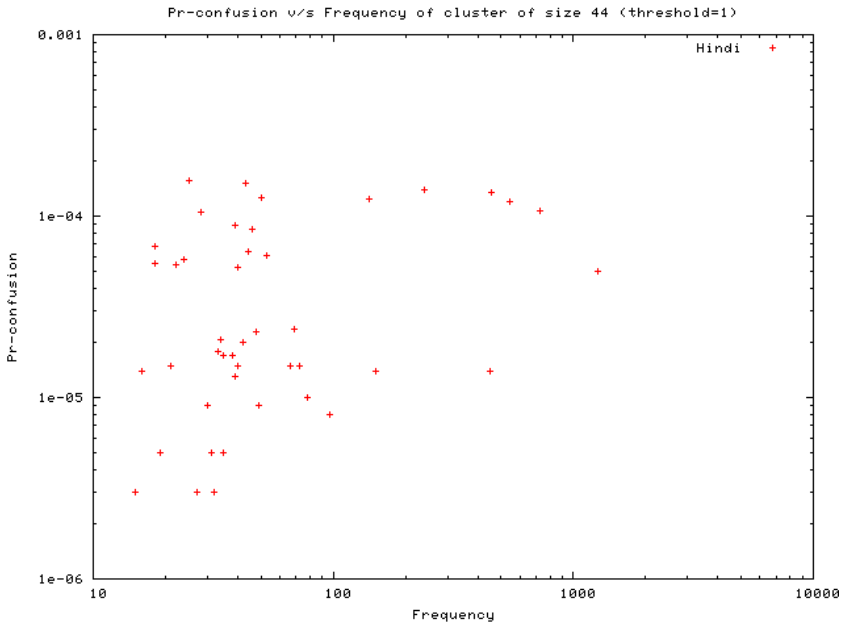

Figure 7: The confusion probability vs frequency for the cluster with size of 44 nodes, near to the mean size (Threshold = 1).

5. From the Degree Distribution Vs Frequency plot (Figure 8), we observe that low frequency nodes may have any degree but high frequency words tend to have high degrees only.

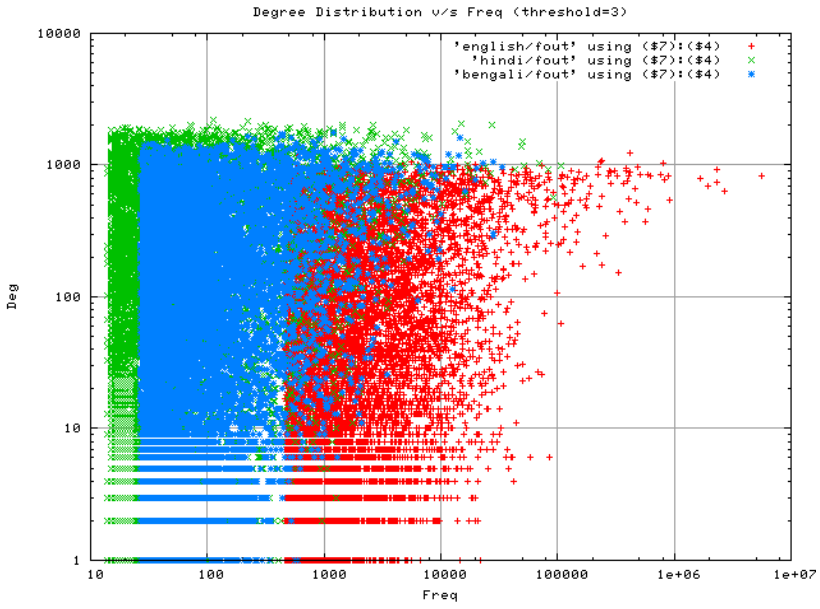

Figure 8: Degree Distribution Vs Frequency at Threshold = 3

6. From the Degree Distribution v/s Frequency plot and Clustering Coefficient v/s Frequency plot, we infer that very high frequency nodes have high clustering coefficient and high degree. That is, **high frequency implies high clustering coefficient and high degree.** However, Vitevitch suggested the other direction that is, high degree and high clustering coefficient should imply high frequency. This is *not* supported by our observations. Words with high frequency and high clustering coefficient spans over the whole frequency range i.e. they can have any frequency. This is can be verified as follows. We earlier concluded that clustering coefficient is almost proportional to the degree. Thus, high degree nodes have invariably high clustering coefficient only. Since, high degree does not imply high frequency (refer Figure 8), thus **high degree and high clustering coefficient does not imply high frequency.** Consider Figure 8 for Degree Distribution Vs Frequency and Figure 9 for Clustering Coefficient Vs Frequency.

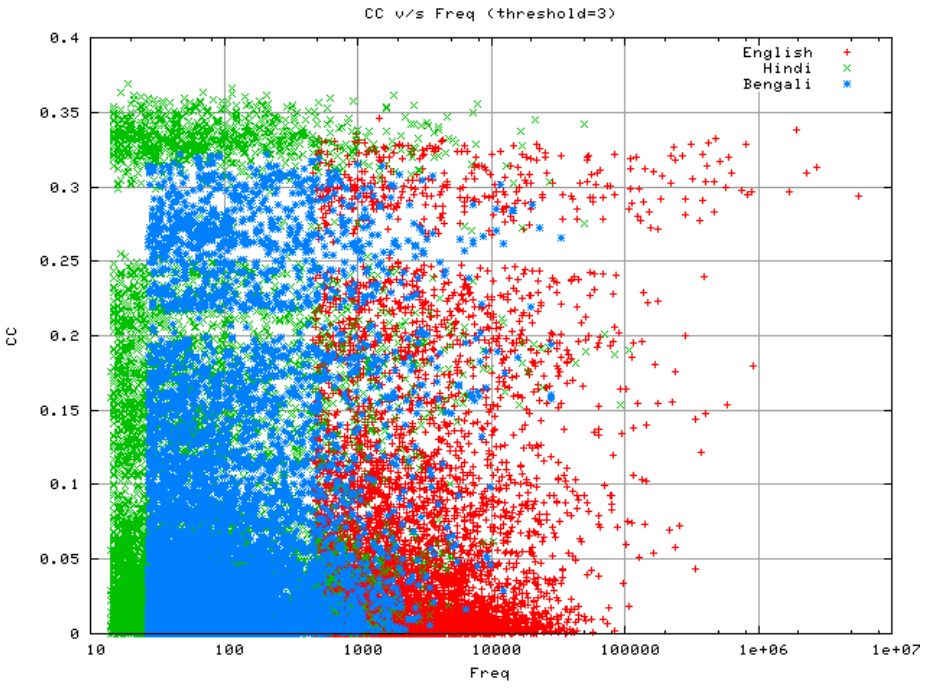

Figure 9: Clustering Coefficient Vs Frequency at Threshold = 3

7. We discussed the clustering details earlier for threshold 1. At threshold 3, we obtained the maximum size of a component to be 9550 nodes, and there are 61 clusters in this

component with maximum size of 6310, minimum size of 2 and mean size of 156.56. Now, consider the distribution of the cluster size for the Phonemic Network for Hindi language at threshold 1 and 3 (Figure 10 and 11). We observe that the distribution roughly follows the Power Law, thereby suggesting some kind of self similarity in the network.

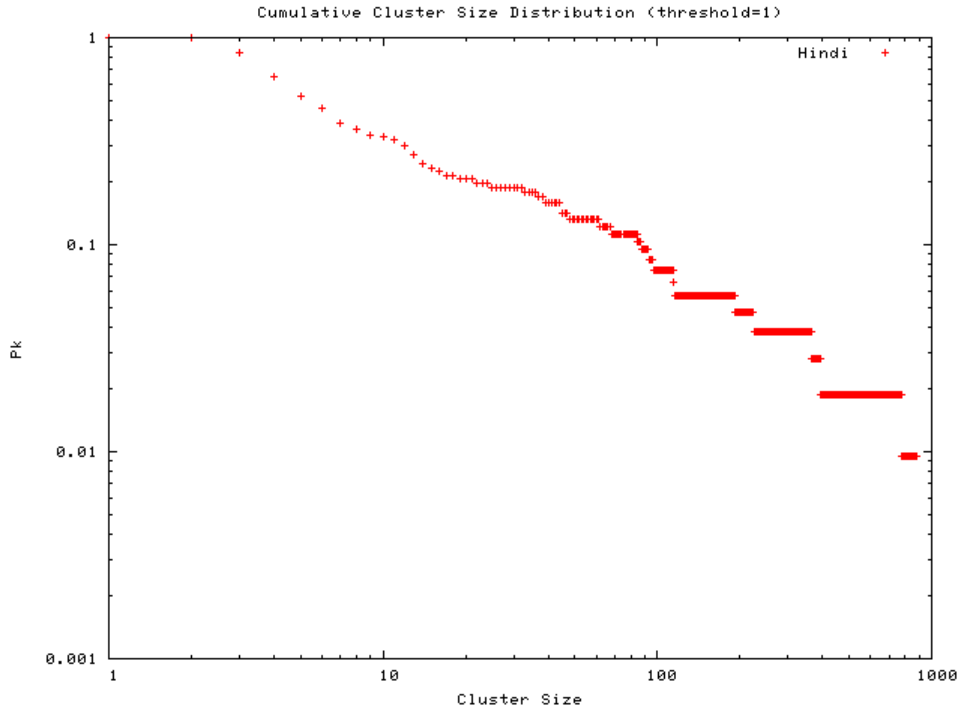

Figure 10: Cumulative Cluster Size Distribution at Threshold = 1

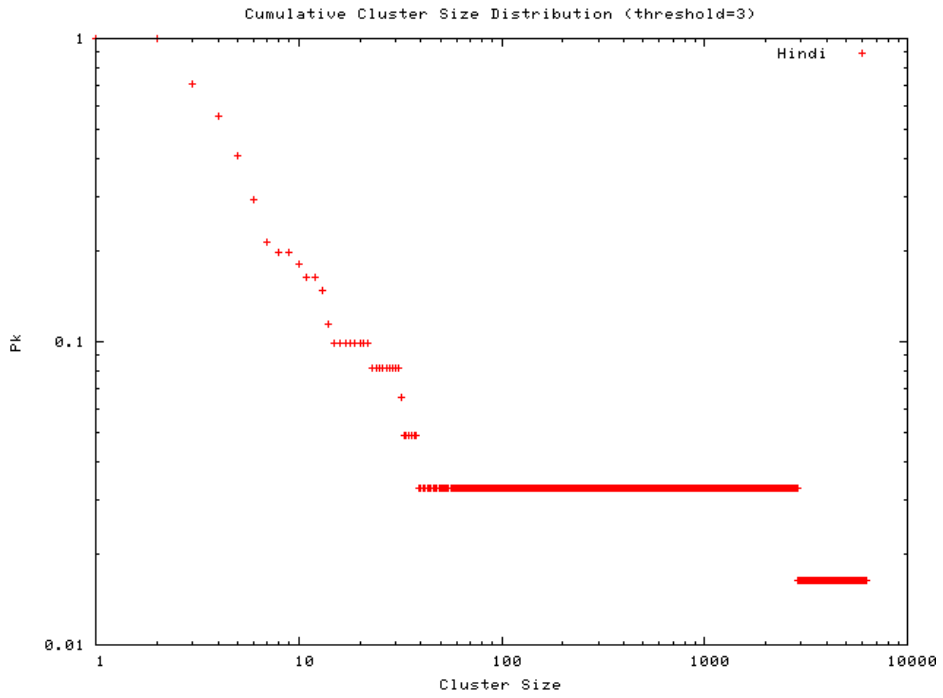

Figure 11: Cumulative Cluster Size Distribution at Threshold = 3

8. The Degree Distribution within each cluster also follows the Power Law. **This shows that the network is self-similar in structure.** Consider the Degree Distribution of the largest cluster and the mean sized (approximately) cluster for the Phonemic Network of Hindi language at threshold 1 (Figure 12 and 13).

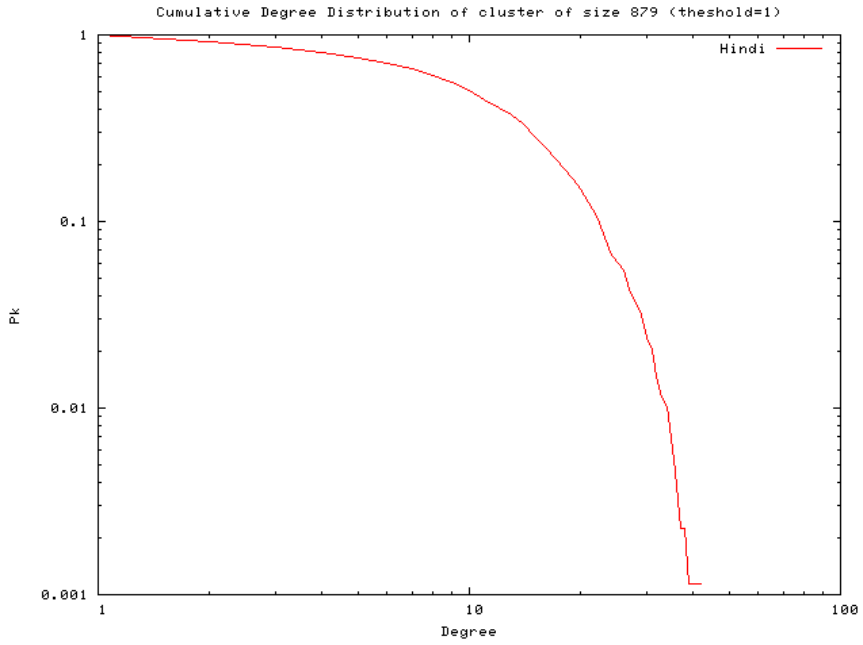

Figure 12: Cumulative Degree Distribution of cluster of size 879.

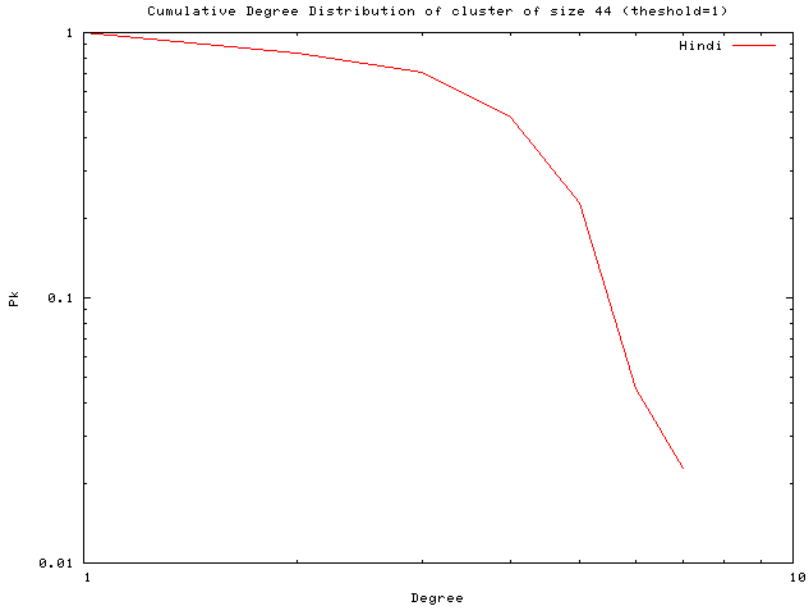

Figure 13: Cumulative Degree Distribution of cluster of approximately mean size (44 nodes).

6.2. Other Interesting Observations

1. If we increase the Edit Distance Threshold to 8, the graph almost becomes fully connected (for all three languages) (Figure 14). This means that almost all the words in these languages are within edit-distance 8 from each other in the Graphemic Networks.

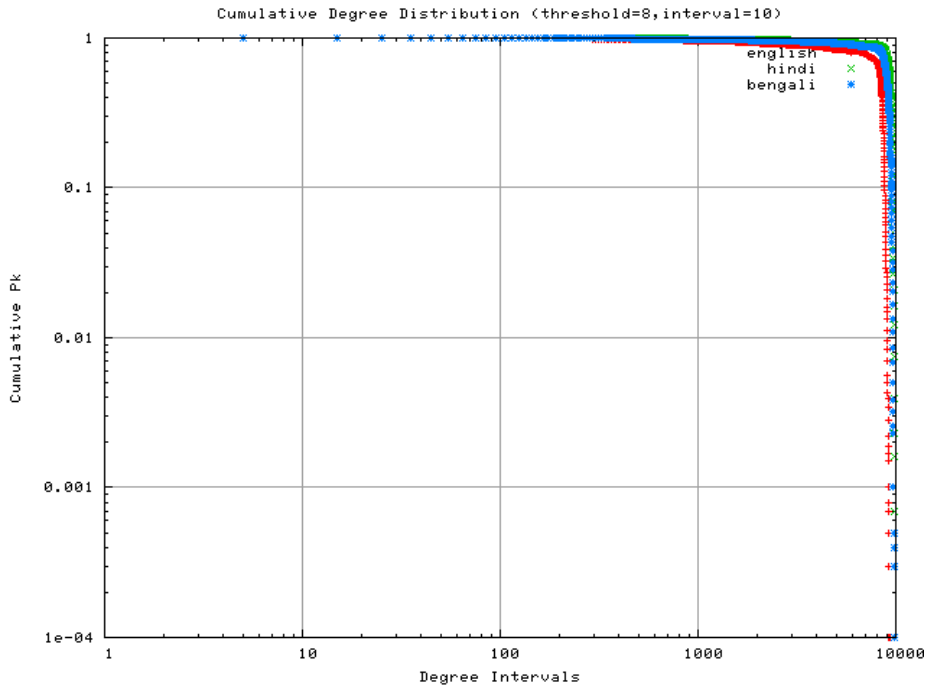

Figure 14: Cumulative Degree Distribution at Threshold = 8

2. Consider the word length distribution for the phonemic network for Hindi at threshold =1 (Figure 15). We note that for length = 6 the frequency is highest.

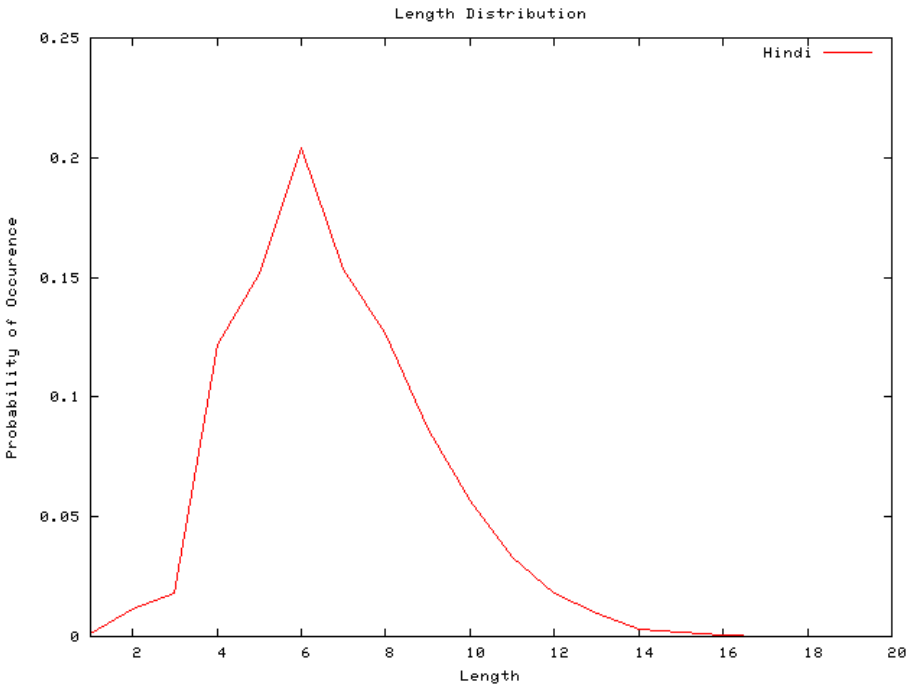

Figure 15: Probability of Occurrence Vs Length

7. Conclusions and Future Work

In this work, we obtained some really interesting results, some were intuitive and some were counterintuitive. We showed that the networks exhibit the properties of social networks like power law degree distribution, self-similarity even at the cluster level etc. Our earlier hypothesis that the networks should be dissassortative with respect to frequency in the high frequency region proved to be wrong. The high frequency words tend to have high confusion rate. *This may be an indication that confusion rate is primarily dependant on the usage context rather than on edit-distances from it to other words.* We also found out that clustering coefficient is proportional to the degree in these graphs and thus, the graph is assortative with respect to degree. We tried to verify the predictions made by Vitevitch and we found out that our observations do not support his predictions. However, we established the reverse direction that is, high frequency words tend to have high clustering coefficient and high degree, empirically. Since we have obtained sufficient empirical results, our future work would be to extend a theoretical model based on the two opposing functions *Articulation Effort* and *Confusion Rate.*

8. References

[1] “Phonological Neighbors in a Small World: What Can Graph Theory Tell us About Word Learning?”, Michael S. Vitevitch (Presentation)

[2] “Finding community structure in very large networks”, Aaron Clauset, M. E. J. Newman, and Cristopher Moore, *Phys. Rev. E* **70**, 066111 (2004).