

Project Report

Network Analysis Of Languages in the Sub-Continent

Anisa Aisha Ahmed, Farooq Abdul Rehman

Course: Understanding Social Networks



Table of Contents

Abstract	2
Introduction	2
Methodology	3
Analysis	5
Discussion & Conclusion	16
References	18

Abstract

Languages we speak evolve with time. This research project aims to analyze different languages and find out the relationships between each other. Three languages i.e. Hindustani, Sanskrit and Persian have been chosen which have a deep influence on each other. In this project, the words of these three languages are mapped as a network graph. The connections of these words are based on similarities in their meanings, phonetics and etymological background. We then use Gephi, Pajek and Tulip — analysis and graph visualization softwares — to determine interesting quantifiable properties, such as how close a particular word in one language is similar from another.

Introduction

Hindustani, also referred to as Hindi or Urdu, derives much of its vocabulary from Sanskrit and Persian. Interestingly enough, many words of such vocabulary with entirely different meanings are the evolved versions of a single word from a different language. The study of these words is interesting and important, as they determine how a certain word influences other words, and what is the basic structure of words in a language. Using centrality scores, the most influencing words and their distances from some of the commonly used words in current formal and informal communication can be identified. This crucial information helps us understand how close-knitted and interconnected these languages are.

Research Questions and Hypothesis

Our research aims to answer the following questions:

- How are the languages that exist in the sub-continent interconnected?
- Among the words used in this research, which words have diverse meanings?

After studying Shamsur Rahman Faruqi's article, *A Long History of Urdu Literary Culture*, we were convinced that colonial philology played an important role in shaping our

contemporary linguistic identities. Our hypothesis is that the ancient languages, specifically the Indo-Iranian languages, played a crucial role in the formation of these linguistic identities. We prove this hypothesis by analyzing their relationships by forming a network of words of these languages.

Methodology

The methodology for collecting data for this research was manual. As suggested by Dr. Shah Jamal, we consulted Sir Afzal Ahmed Syed to gain resources for our research. As per their suggestion, the data was mainly extracted from Platt's Dictionary [1].

Languages have millions of words, and it is almost impossible for us to map all these words and create a network. Therefore, we picked 3 words from the Hindustani language, **parvarish**, **sayah** and **zabaan**. We limited our analysis to 100 words and established connections of them with words of similar meanings from the languages mentioned above. In our data, 36% words are of Hindustani origin, 23% words are of Sanskrit origin, 24% words are of Persian origin, and 17% words are of Urdu and Persian origin.

Following pie chart illustrates the frequency of words of each language.

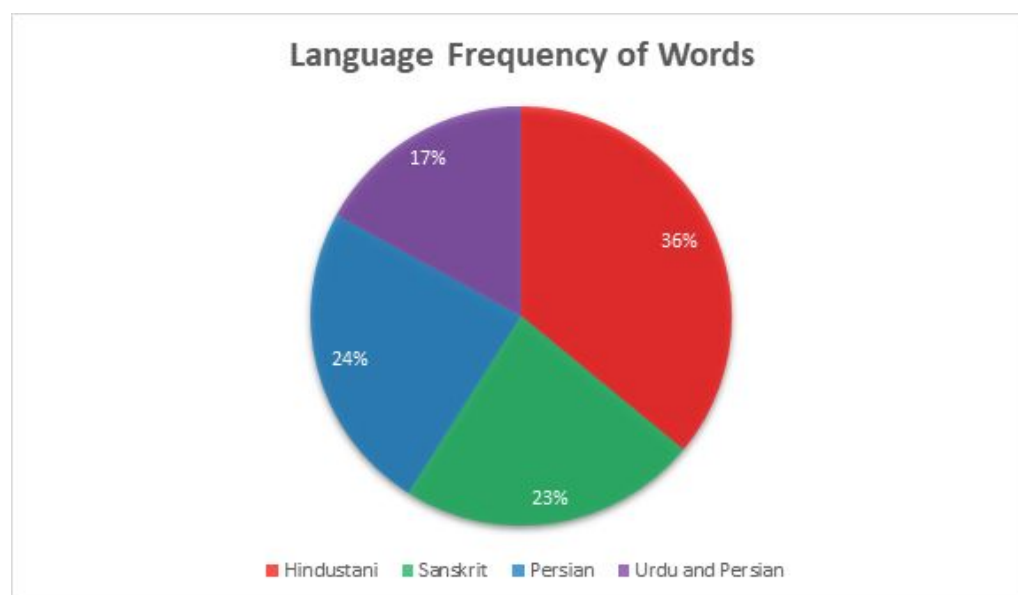



Figure 1:
Language frequency of words in the network



Data for this network was manually entered and formatted in a **.net** file. The **.net** file consists of node labels as words, and were colored differently to represent the three different languages under study. An edge list was created which represented connections of words with other words. A **GML** file was also created to analyze data. This data analysis was performed using Gephi, Pajek and Tulip softwares. We then applied *Kamada Kawai* algorithm to layout our network in an aesthetically-pleasing way, but more importantly, help us distinguish between the two sub-networks and focus on the central nodes.

The analysis showed extraordinary results as the interconnectivity of languages was more than what we expected.

Analysis

We have analyzed our network on three different levels, on the whole network level, on group level and lastly, on individual level.

Our network was formed by making connections with respect to the meaning of the three words that we chose initially. The nodes in this network represent the words of various languages, and edges represent the similarity in meaning between words. The different color of nodes represent the different languages which are shown in the network diagram.

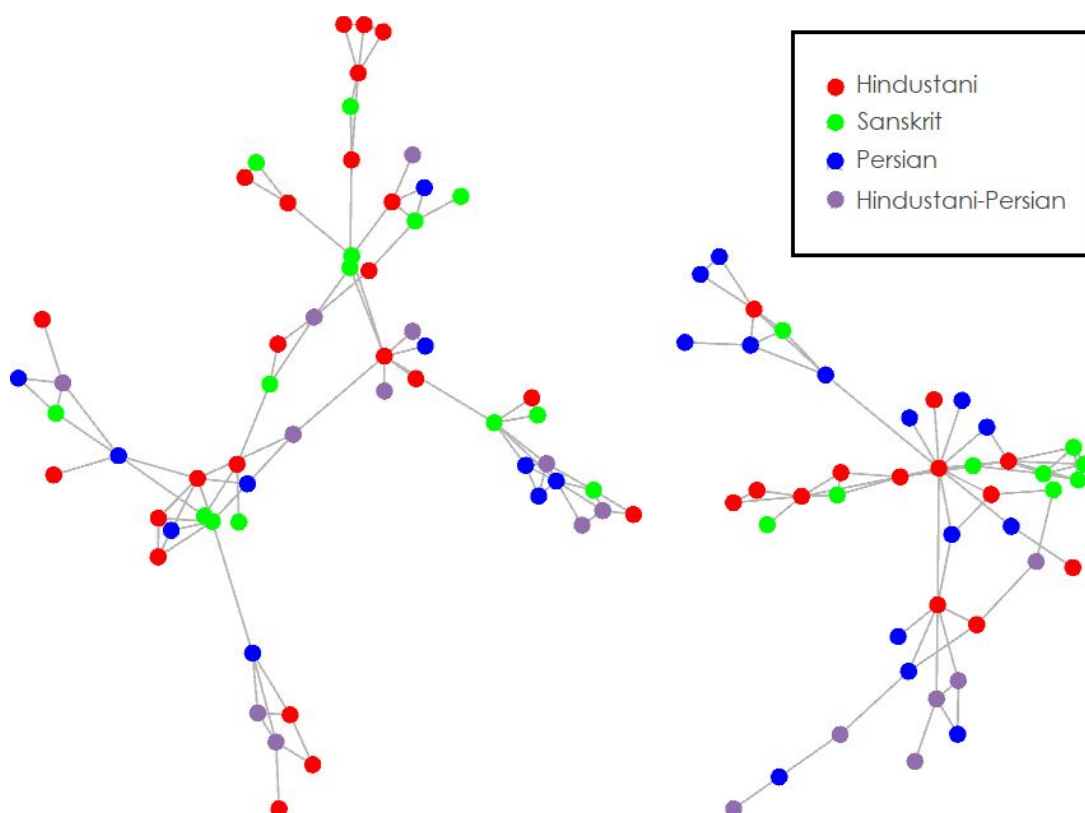


Figure 2: Network diagram using Kamada-Kawai algorithm

Density and Degree Distribution

The density of the network was significantly low, and was calculated as 0.031. Low measure of density indicated this network is sparse.

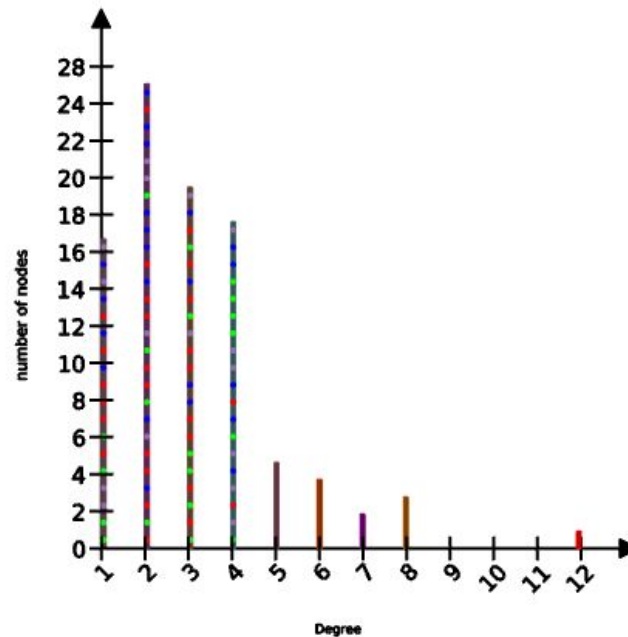


Figure 3: Degree Distribution of network

The degree distribution of the network was moderately varying since most of it lied between 1 and 4. The average degree calculated was 3.1. By looking at the histogram we can observe the degree distribution of the network follows the Poisson distribution to some extent. However, it is important to note that our network is based on limited data, hence a distribution of this form is observed. It is our assumption that for a large data set of words, the network will follow a different distribution, possibly the power law.

The standard deviation of degree of the network was calculated as 1.9 which indicates low variance from the mean value (3.1), in resemblance with the high degree distribution on short interval i.e. [1,4]

Bi Components and Cut-vertices

An important characteristic of this network is that it is divided in two components. The largest component of the network has diameter 10, which indicates that two words who appear to have completely different meanings are still linked together by such a short distance. Another important thing to note here is that this network doesn't have any isolates, meaning there are no such words who are not linked with some other word. This contributes towards proving our hypothesis that words are very closely interlinked with one another. Our network has 10 cut-vertices which can be seen in the below figure (colored light grey).

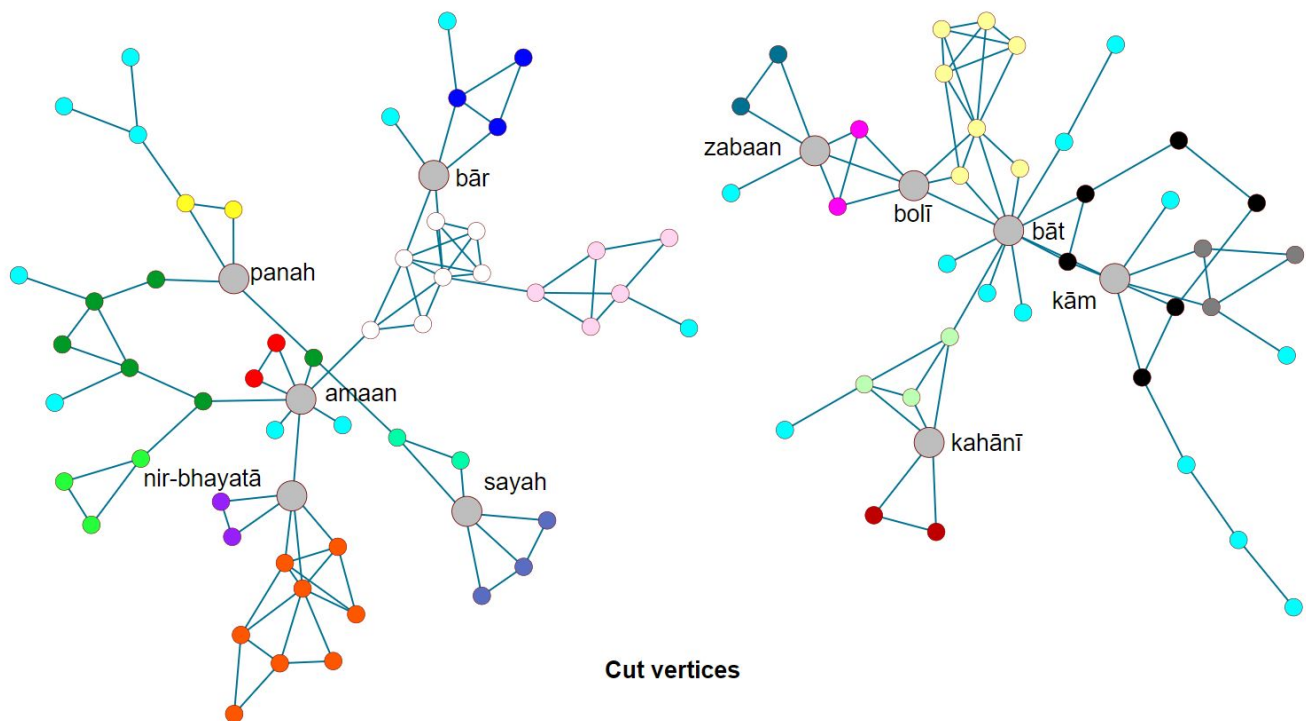


Figure 4: Cut-vertices in the network

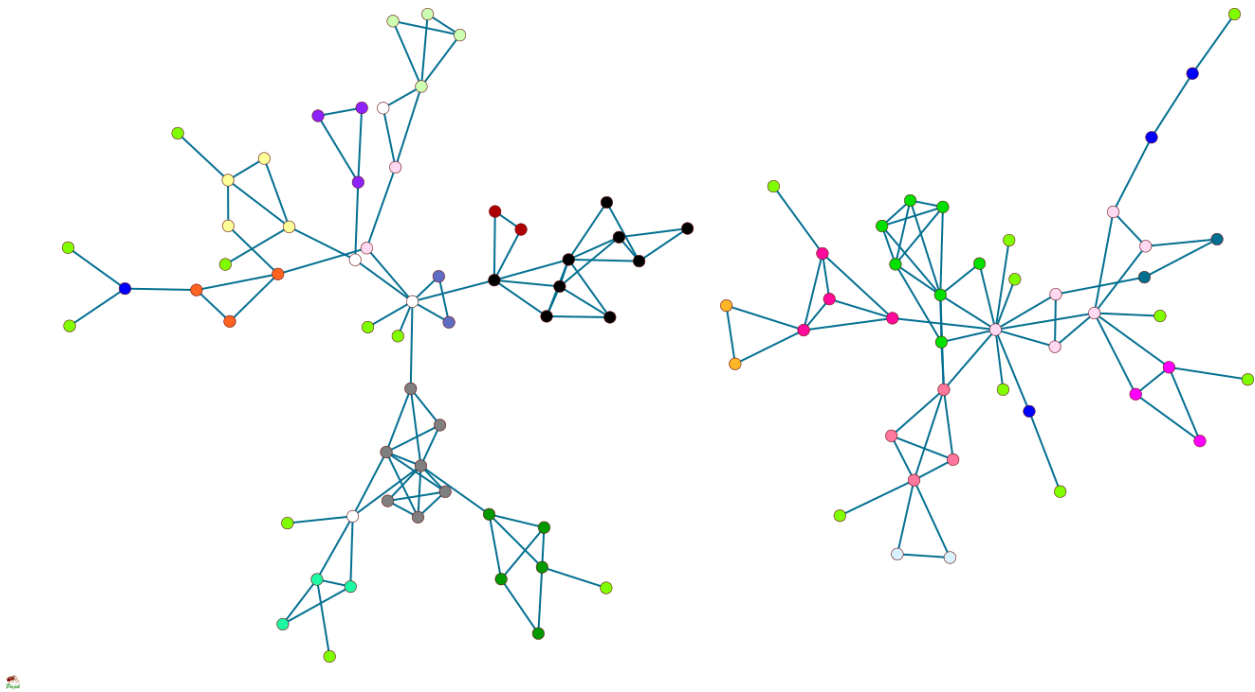


Figure 6: n -cliques (Different colors indicate different cliques)

Cliques were also identified in the network to analyze cohesive groups. Here we observe that there are quite a large number of cliques in our network. There are only a few words which are not a part of any clique, and on analysis we see that there is a fairly large number of cliques containing more than 3 words. These cliques signify that the words are closely connected with each other, regardless of the language they belong to. The words which are not a part of large cliques can be seen as intermediate words which connect words from one clique to another. This is important because rather than majority words being connected to a single word, words are connected within groups and one word from that group maintains a connection with words from other groups (in real world networks, we could call this node as the broker). Hence, we observe that even a network of words exhibit some properties from real world networks.

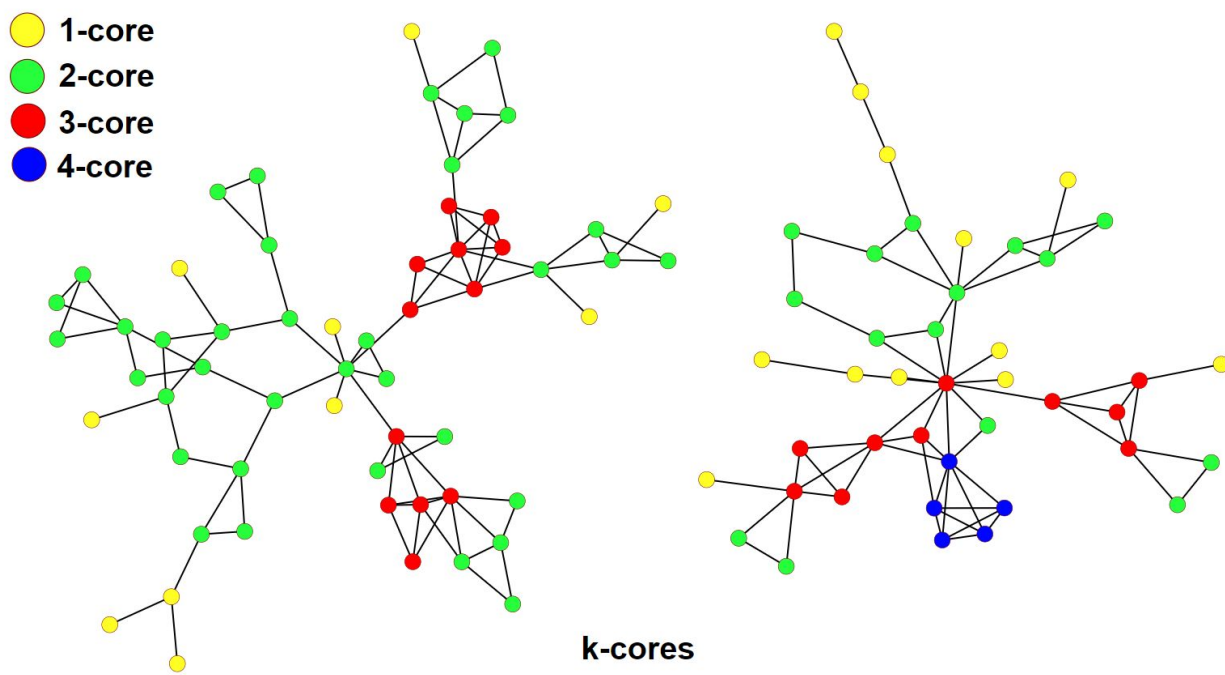


Figure 7: *k*-cores of the network

K-cores are an effective way to find cohesive groups in the network. In our network particularly, they serve an important meaning: similar words tend to form groups. One interesting observation we made is how the similarities between words increase as we move to a higher valued core. In other words, the higher the value of k is, the higher the extent to which words in that particular core are similar. For example, 4-cores are the most tightly connected subgroups which means 4-core words are more similar to each other than words in less valued cores.

Clustering

Our network is divided into two large components, both being highly clustered, which will get more apparent as we calculate the local clustering coefficients for the nodes and distribute them over a plot.

Our network consists of a significant amount of triangles, consequent with the high local clustering coefficient we obtained which tells us that many words have one or multiple synonyms. The entire network is not a connected graph and is divided into two large

components. There are 66 triangles in our network and the average clustering coefficient is 0.574 which is quite large but common in real networks.

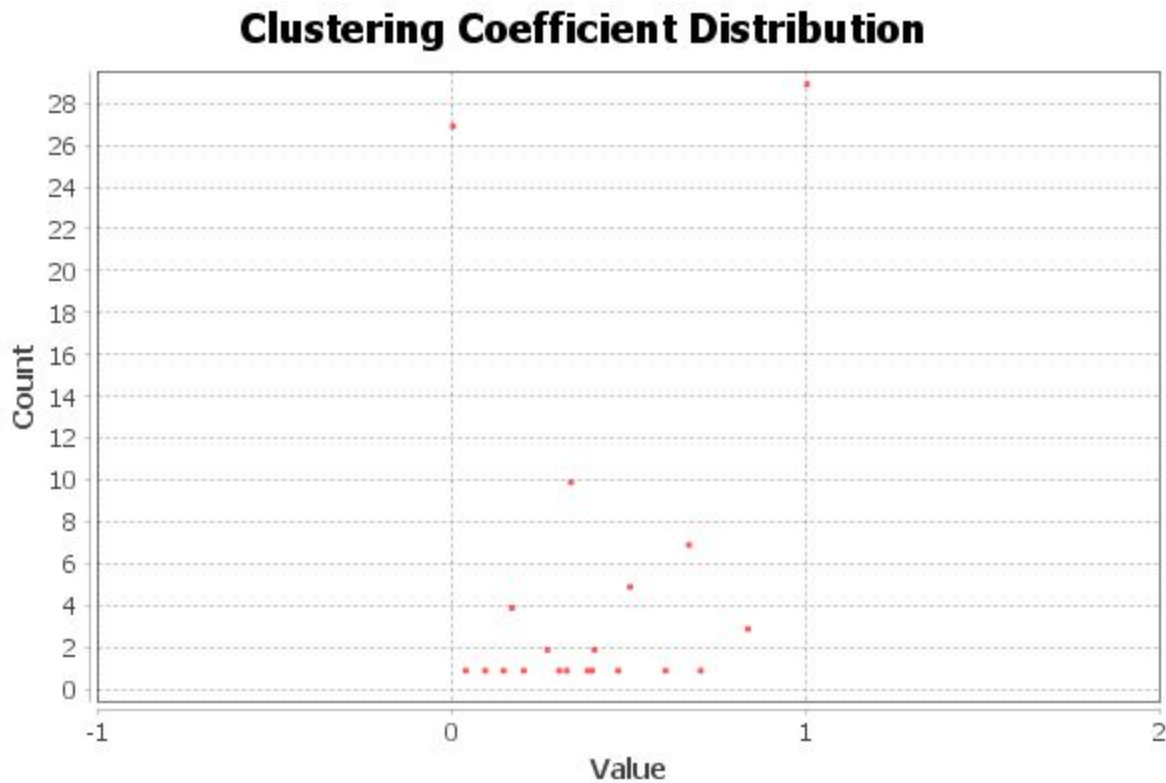


Figure 8: Clustering coefficient distribution of the network

From the clustering coefficient distribution plot which we generated using Gephi, we can see that out of 100 nodes, 29 nodes have a local clustering coefficient of 1 which makes this network a highly clustered one.

Eccentricity

Most nodes in the graph were found to have high eccentricity, which is indicated in the network diagram as well as the eccentricity distribution histogram. High eccentricity of nodes reveal that the nodes are far apart. Thus, the graph is sparse, and nodes have fewer connections with other nodes, but for every node there exists at least one link, which specifies that the every word is connected to some other word in some way.

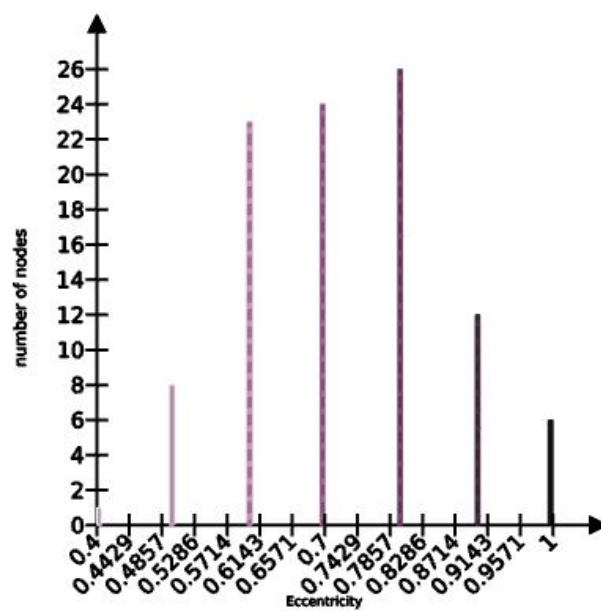
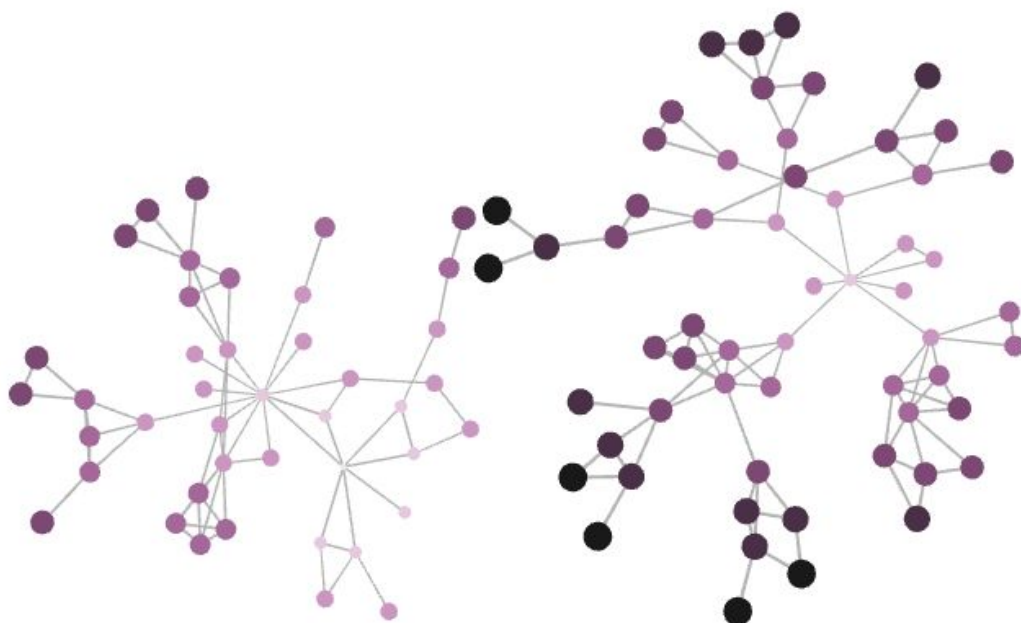


Figure 9: Eccentricity of nodes (highest to lowest from black to pink transition in color)

Centrality - Degree, Betweenness and Closeness

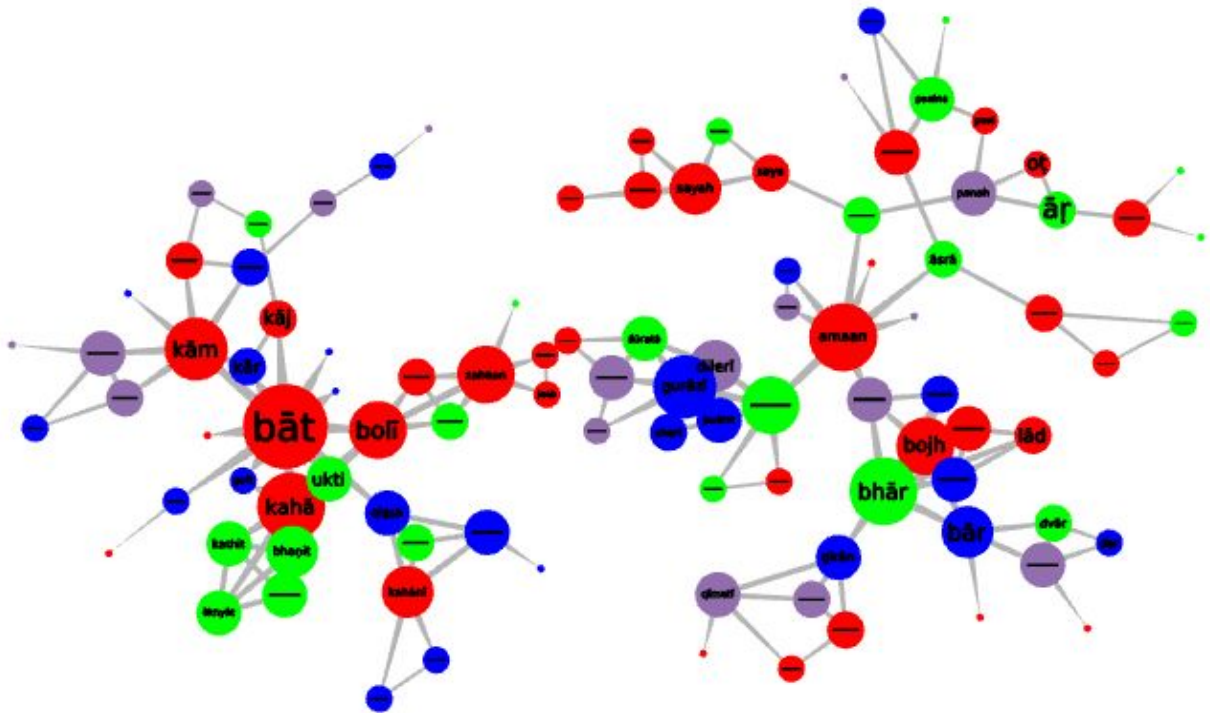


Figure 10: Size mapping of degree centrality of nodes

The most influential words in the network can be determined using centrality scores. Degree centrality of the network was found to be 0.09173. This is significantly low, suggesting that the difference between the largest and smallest value of degree centrality is not high. As represented in the above figure, word **bāt** has the highest degree centrality. This indicates that it has a very diverse meaning, and is connected to several other words in the network. Words **kahā**, **bhār**, **amaan**, **kām** and **g_urāzī** also exhibit high degree centrality, having variety of meanings.

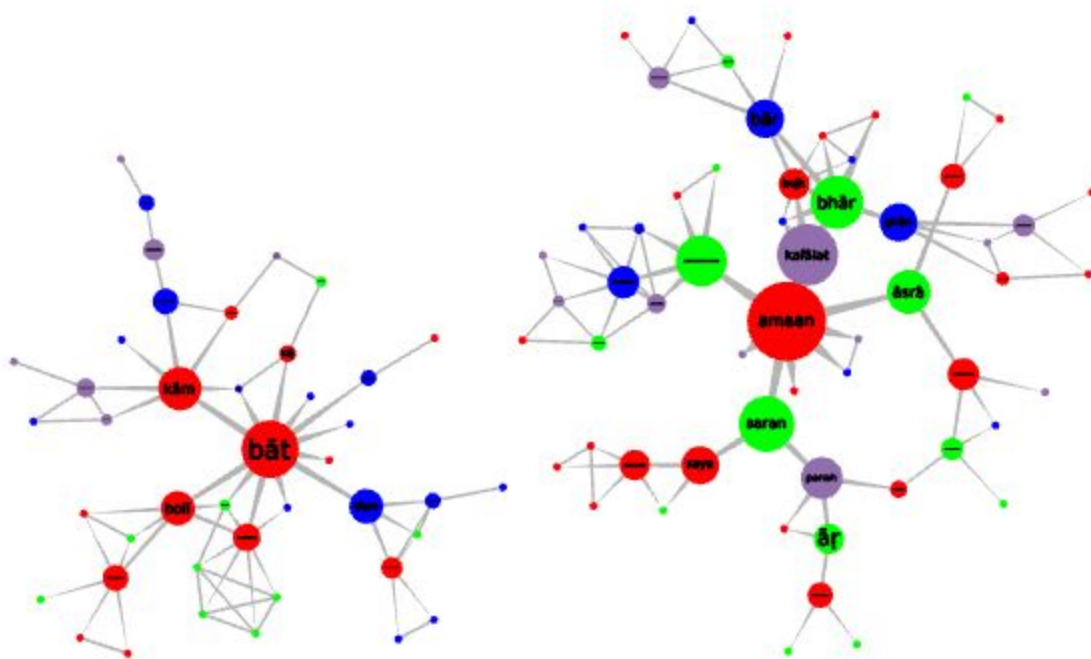


Figure 11: Size mapping of betweenness centrality of nodes

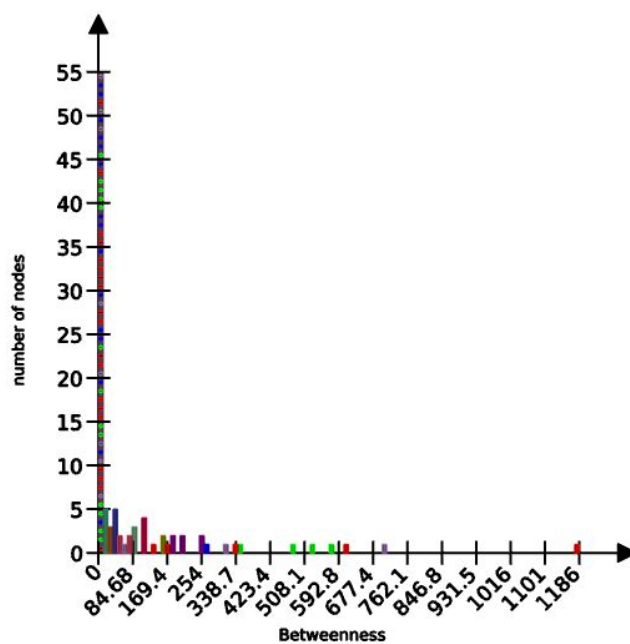


Figure 12: Distribution of betweenness centrality of nodes

Betweenness centrality of the network was calculated as 0.2287. The above figure represents the nodes with high betweenness centrality, word **amaan** having the highest measure. High betweenness indicates that the word is prestigious, as it acts as an intermediate synonym between two words. As we can see that the word **amaan** connects **kafālat** and **āsrā**, which differ relatively in meaning. Similarly, it connects several other words having contrary meanings. It is also interesting to note here that the node with highest degree centrality (**bāt**), has lower betweenness centrality than **amaan** and **kafālat**. Having diverse meaning thus does not necessarily indicate that it will act as an intermediate synonym between words.

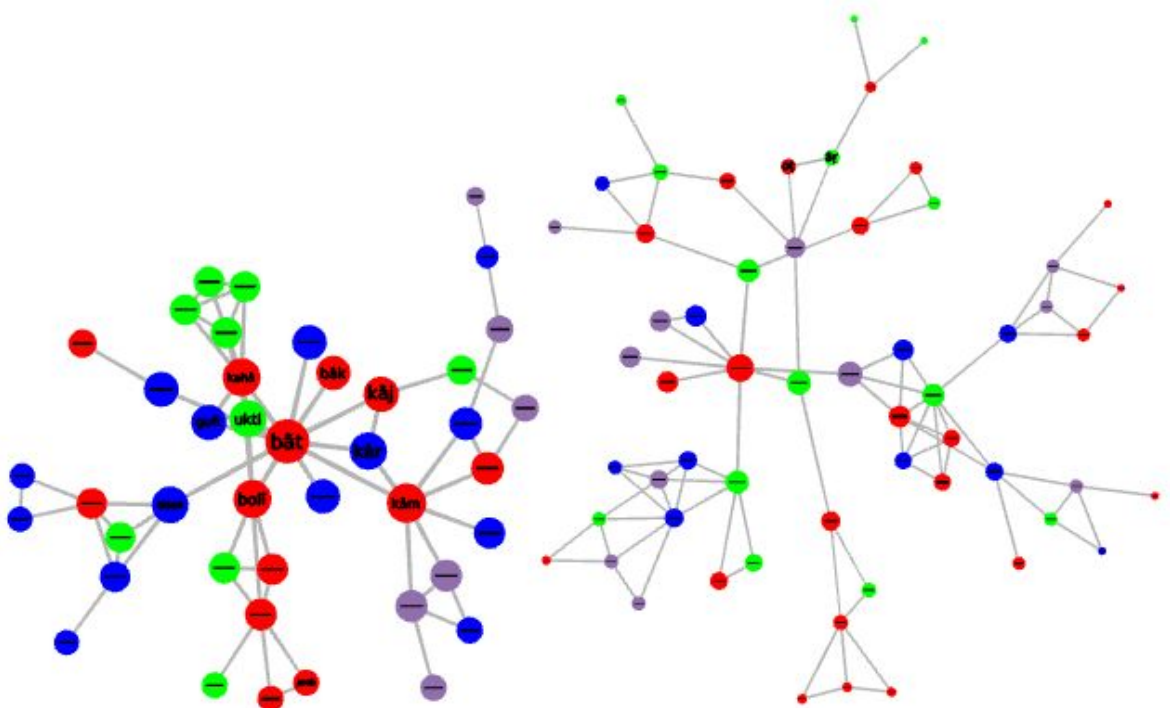



Figure 13: Size mapping of closeness centrality of nodes

Closeness centrality measure for majority nodes was found to be notably low. Word **bāt** has the highest closeness centrality of 0.0117647, so it connects with words of its own cluster only, and not with the words from the entire network. This behaviour is observed in almost all the nodes in this network.




Analyzing the above calculated centrality measures, it is interesting to note that words from Hindustani origin occupy a prestigious position in the network. A specific example would be of word **bāt**, which has the highest degree and closeness centrality. It is also noted that most of the words in the network do not have high degree and closeness centralities, but have fairly high betweenness centrality. This determines that most words act as intermediate synonyms between other words. Thus, a word has multiple meanings, and it is connected to its respective meanings.

Discussion and Conclusion

The analysis gives us a better idea about the connectivity and relationship between aforementioned languages from three different aspects. On a network level, the density and average degree suggests a moderately dense graph, even with just 100 words. It is quite possible that extending such network would result in a more dense network. Then the diameter of the network (10) gives a surprising feature of languages, that if we start our path from a certain word, we can reach to a word with completely different meanings. If we look at the network on a subgroup level, we come to discover some really interesting characteristics of languages. Identifying bi-components, cores and cliques proves our hypothesis and signifies the fact that words with similar meanings tend to be close-knitted, irrespective of the language that they belong. Cut vertices we found could potentially be root words for connected components. On individual level, we look at the importance of each word through centrality scores. We came to the conclusion that every word is no less important than its neighbors and contributes to the evolution of languages as a whole.

The nodes which appear to have more prestige were considered as the most influential words of the network. Betweenness centrality was the most useful measure in determining the prestige, as it determined the key words which were necessary for the network to stay connected, which was the prime question of this research. Henceforth, the most influential words in this sample network were found to be **bāt** and **amaan**. A general observation would be that words from Hindustani origin were more influential than words from the other two languages.



Finally, this study can be expanded to a larger set of words which will help us discover more interesting properties as well as improve already established features of these languages. We believe that the Hindustani language and its roots can be explored in greater details with an extensive and curated analysis of its relationships with other languages, particularly Persian, Sanskrit and Arabic. We can also expand our resources to the literature to study the evolution of words. All in all, this research carries significant potential and could motivate linguists and social scientists to explore the area in better ways using modern network analysis tools.



References

[1] Platts, John T. (John Thompson). A dictionary of Urdu, classical Hindi, and English. London: W. H. Allen & Co., 1884.

<http://dsal.uchicago.edu/dictionaries/platts/>