Wikipedia Network Analysis: Working with Large Data Sets

ESD.342 Project Report
May 16, 2006

Justin Lindsey
Dave Long
Alex Mozdzanowska

1. Introduction

The original goal of this project was to understand the structure of data on Wikipedia. The Wikipedia is an online encyclopedia that can be edited by anyone who wishes to do so. As a result, the content of the encyclopedia and how that content is structured emerges over time and is the result of the input of many individuals. This rich data set provides ample opportunity to analyze how network structure emerges as the network grows. One of the initial questions framing our project was whether there is an inherent structure of information that emerges regardless of what language is being studied. However, it rapidly became evident that the amount of data available on Wikipedia made the computation time to calculate network parameters either very costly or impossible. Faced with this difficulty, three options were available. The first was to choose a small subset of data and limit analysis only to that set. The second choice was to limit the analysis, but find a way to analyze the entire set of data efficiently. The third option, and the one chosen by the group, was to combine approaches one and two.

Three types of analysis were conducted as part of this project. The first analysis looked at a large number of language networks and conducted an analysis of network properties for those networks. Since the language pages are of different ages and sizes, and size was shown to increase with age, studying different languages provided a way to analyze how a Wikipedia network grows over time. The second and third analyses focused around pattern matching. In one instance specific branching patterns from a node were identified and all nodes matching that pattern were selected. The network properties of these subsets of the Wikipedia data were then analyzed. For the final analysis, the pattern of all nodes in a network was identified and plotted resulting in a pattern signature of the entire network.

Studying the growth and properties of the Wikipedia data provided some understanding of how data was structured and how the Wikipedia networks grow. That understanding proved highly useful in conducting the pattern analysis of both single nodes and the entire network.

Studying the branching pattern of nodes and networks was conducted in order to determine if the information contained within the data (1) has a structure that can be identified and (2) if the unique structure of a network can be used to identify and classify that network. In addition to the specific motivation mentioned above, studying patterns in networks has a broader interest. Understanding patterns can help quickly identify and locate subcomponents of networks of interest. Finding and identifying patterns could help track and identify terrorist or drug cells. For example the patterns of phone calls could be identified for a cell. After identifying the patterns, a particular cell could be found even if the cell changes geographic location or telephone numbers because the pattern of interaction between the members would be the same. Pattern identification methods could also be used to study organizational response to a crisis, determine if a military configuration indicates that a unit is posed for battle, and other applications.

2. Data

This study utilized Wikipedia data segmented by language. The Wikipedia is an online encyclopedia which can be edited by any user. For the purpose of this study, the network nodes were defined as Wikipedia pages and the network links were defined as html links between pages.

The growth analysis of the Wikipedia languages used 206 of the 229 existing Wikipedia Languages. Ten of the smallest languages were excluded because they did not contain any links, the two largest sets of data (English and German) were excluded due to their prohibitive size, and eleven other languages were excluded because the data was not available. The smallest included language contained about 1,500 nodes with 3 links and the largest included language contained about 600,000 nodes with 10,000,000 links.

The pattern analysis of a subset of networks was conducted on the West Frisian Wikipedia. West Frisian is a language spoken in the northern part of the Netherlands and has an estimated 300,000 to 700,000 speakers. This language was selected because of its relatively small size. The data set has about 6,000 nodes and 72,000 links. Out of this data, nodes matching the simple pattern of 31 out degrees were identified. Of the twenty nodes that matched this criterion, five were randomly selected for analysis.

The analysis of patterns across an entire network was conducted both on the West Frisian language as well as Latin. Latin is a slightly larger data set containing about 10,000 nodes and 100,000 links.

3. Approach: Large Network Analysis

The largest obstacle to conducting an analysis of the Wikipedia networks was the large sizes of the data sets. In order to accommodate the complexity and time and space requirements of working with this data, more efficient algorithms needed to be developed. Adjacency matrices had to be replaced with sparse matrix representations. Working with lists of node pairs or lists of node children also proved to be a more efficient representation than the adjacency matrix. Figure 1 shows a representation of the methodology. The following two subsections describe in more detail aspects of the methodology that pertain to all three types of analyses conducted for this project. Specific analysis methodologies will be discussed later in the document.

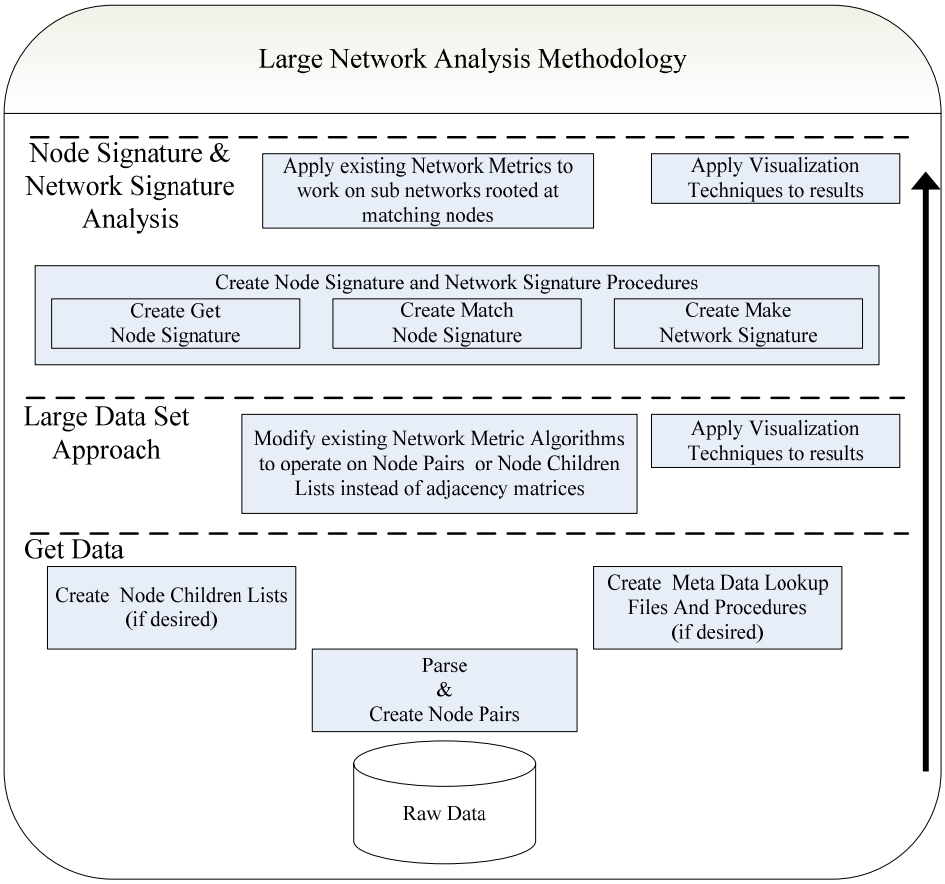

Figure 1: Methodology Representation

3.1. Phase I: Get Data

The first phase of the methodology was to retrieve the Wikipedia data. While this may seem simple, the size and complexity of the data available made this step challenging. The selected files of data had to be downloaded and parsed into a meaningful and usable representation. Code was written to extract data and convert it to a node pair list or a node children list. The node list contains two columns with a FromNodeID in the first column and a ToNodeID in the second. An example is shown below:

128 119
128 138
128 157
157 122
157 147

For some aspects of the node and network pattern analysis a node children list was used to further compact the available data. An example of how the above sequence changes when represented as a node children list is shown below:

128 119 138 157
157 122 147

In addition to producing code that generates node lists and node children lists, methods to retrieve or find nodes were generated. More specifically, the ability to retrieve a node ID given a node name (or the reverse) was useful to transition between a numerical representation of a node and the actual topic page in Wikipedia. The ability to move between a node ID and the actual page in Wikipedia was crucial to understand the results of the node pattern matching analysis. Without this ability a match could be found, but it would not have been possible to determine which subjects resulted in the match.

3.2. Phase II: Large Data Set Approach

In order to analyze the obtained sets of Wikipedia data, existing routines required modifications to use node pair lists. Matlab was discovered to be capable of loading large lists of node pairs and converting those lists to sparse matrix representations which were much more computationally efficient than using adjacency matrices. A sparse representation allowed analysis of significantly larger data sets, which in an adjacency representation would have the memory of Matlab and our computers.

In addition to using a sparse matrix representation of data many functions needed to be written or adapted in order to be more efficient. Working with large data sets requires a glass box approach where the specific workings of all algorithms need to be understood and structured for maximum efficiency.

4. Growth of Wikipedia Networks

4.1. Results

In order to understand the growth patterns of Wikipedia networks, the size of a language was used as a proxy for network age. Figure 2 shows figures of networks' sizes, as indicated by (1) the size of the circles, (2) a function of both the age of the network, and (3) the number of language speakers for each language network. The two figures show the same information, but the right hand side figure plots the x-axis plotted on a logarithmic scale to spread the data points to more clearly visualize the trend. It can be seen from the figures that the network size increases both with age and with the number of language speakers. An outlier can be seen on the left hand side of the left pictures. This outlier is China. The language was started later than those for other countries with a similar number of speakers, and the language is also small given the high number of speakers. These factors can be explained by the fact that China has high restrictions to free speech on the internet and access to computer and internet technology is not easily available to everyone.

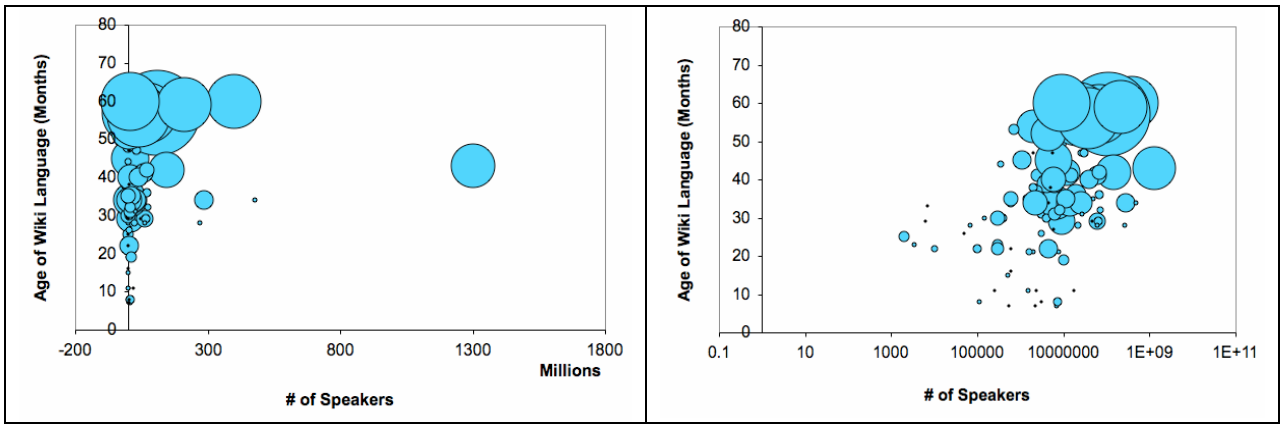

Figure 2: Network size as a Function of Age and Language Speakers

For a large number of the network analyses conducted, the number of connected nodes, rather than the total number of nodes, was used. A connected node has one or more links to another node(s). Figure 3 shows the relationship between the total and the connected number of nodes. This relationship is linear and the number of connected nodes increases with the number of total nodes. For large language data sets, as many as one-third of the total nodes were not connected and the number of not-connected nodes was higher for smaller data sets.

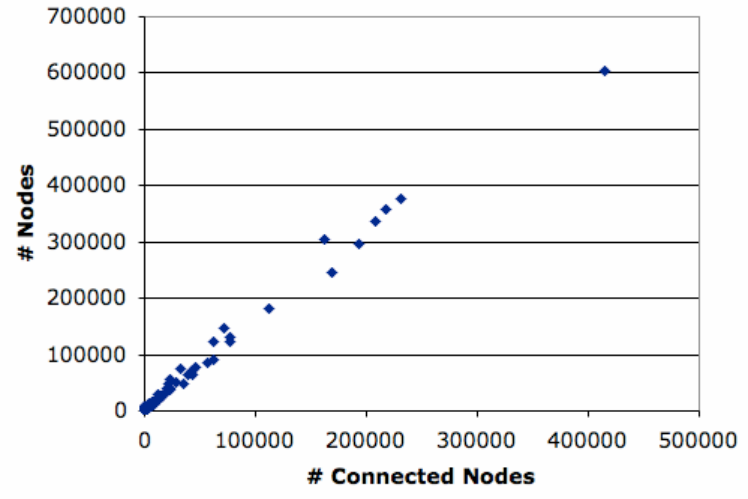

Figure 3: Relationship between Total Nodes and Connected Number of Nodes

Figure 4 shows the growth in the number of links as the number of connected nodes increases. Similar to the relationship between total nodes and connected number of node, the relationship between the total number of links and connected nodes is linear.

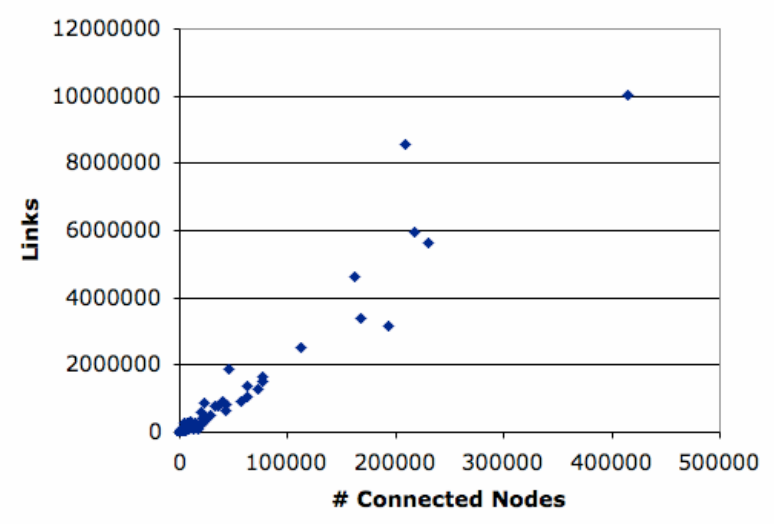

Figure 4: Relationship between the Total Number of Links and Connected Nodes

Figure 5 shows the average degree for each language network. For small networks, the average degree varies significantly anywhere from 0 to more than 20 degrees. Such a variation is expected for small networks because a steady state has not been reached. A hypothesis is that when the network is small, pages with more important substance are added and result in more connections which increase the averages. As the networks grow, the average begins to level between 9 and 14 degrees. It would be interesting to continue tracking this data as the networks increase in size to see if the average decreases. A decrease would indicate that new pages have fewer links than older pages. Such a result could be consistent with the hypothesis that earlier pages contain more substantive information.

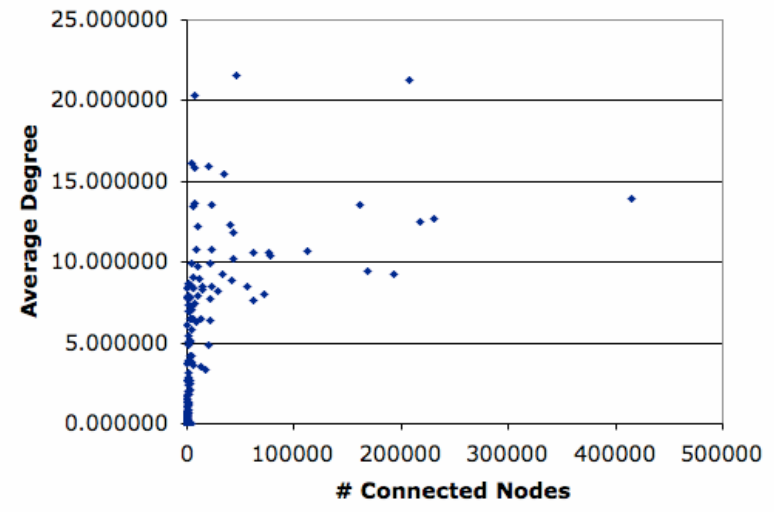

Figure 5: Relationship between Average Degree and the Network Size

The density of the networks as a function of network size was also analyzed and the results are shown in Figure 6. It can be seen that once again for small networks there is a large density variation, but density rapidly settles close to zero as the size of the networks increases. The low density indicates that the network utilizes only a very small percentage of all possible connects and is, in fact, very sparse.

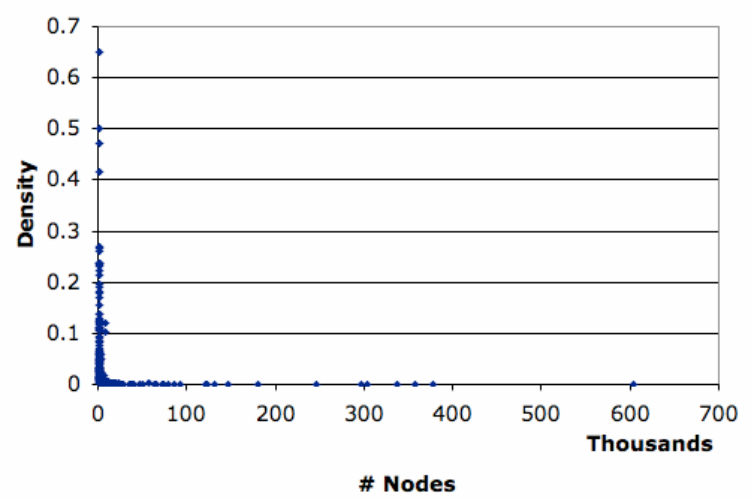

Figure 6: Network Density as a Function of Network Size

The clustering and correlation coefficients (Pearson degree correlation) were also calculated and are shown in Figure 7 and Figure 8. Once again the pattern of high variability moving to a steady state with increasing network size can be observed. In the case of the clustering coefficient, the values settle around 0.05 which is very low. Even the highest values of the clustering coefficient are below 0.25. This result indicates that the networks are not clustered and was highly surprising; the team expected to find large clusters of highly connected pages much like a hub and spoke network. The correlation coefficient shows that the steady state value hovers around zero. This value does not provide much meaning, but seems to indicate that the nodes are consistently connected to neither well-connected nor not-well-connected nodes.

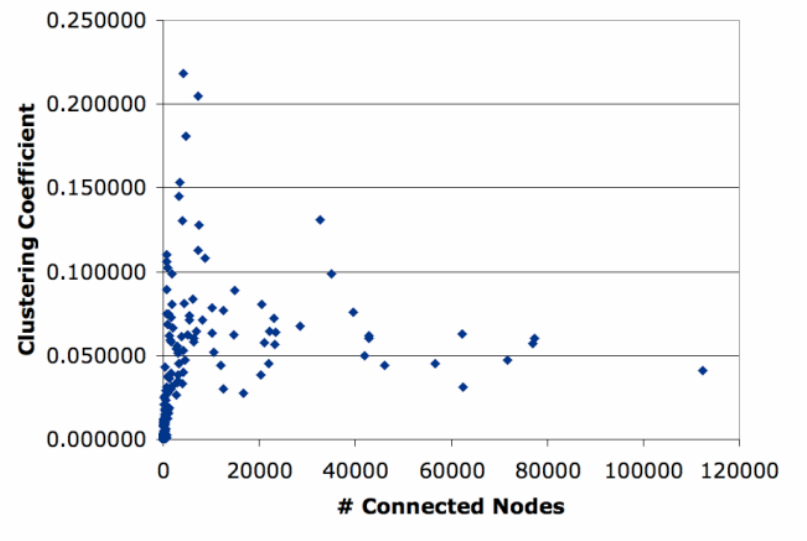

Figure 7: Clustering Coefficient as a Function of Network Size

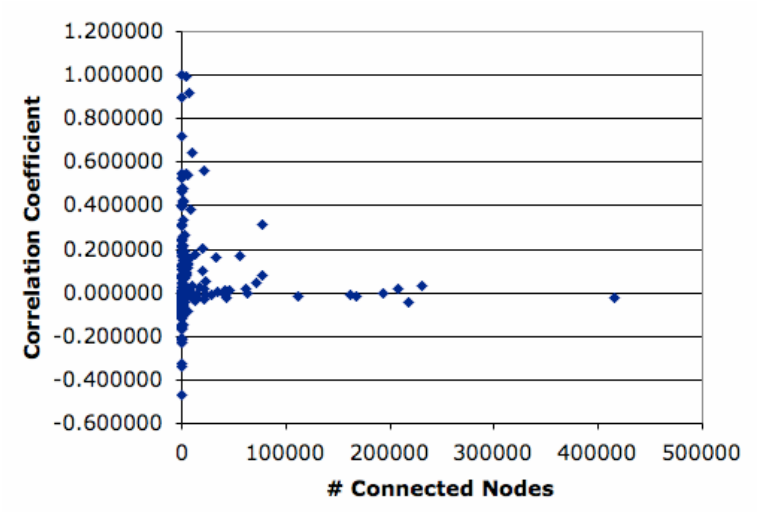

Figure 8: Correlation Coefficient as a Function of Network Size

Based on these results the Wikipedia networks appear to grow randomly. The networks were found to be sparsely connected, not clustered, and not have any significant correlations between nodes. This result was also highly unexpected. The team expected to find large, well-connected clusters that acted as hubs between other smaller nodes. It was also expected that the networks would grow based on preferential attachment where new nodes would be likely to link to highly connected nodes.

4.2. Methodology

The Wikipedia size and growth analysis was conducted using a Perl script to download data and convert it to a list of node pairs. Python scripts were used to obtain and format additional data (such as initiation dates and the number of speakers) from Wikipedia.

Matlab scripts calculated network statistics. Excel was used to collect and plot the resulting data. Some Matlab scripts were written specifically for this project, but others were obtained from the course website. The Matlab scripts were written or modified to read lists of node pairs which are much smaller than adjacency matrices. The scripts also operated on a sparse--rather than a full--matrix representation making it feasible to work with significantly larger sets of data.

5. Nodes Pattern Analysis

The Wikipedia network analysis followed three courses of investigation. After the large network analysis at the language level, pattern analysis was performed at the node level. Working with the node patterns allowed traditional network metric analysis, such as the calculation of path length and betweenness, to be performed on the Wikipedia data sets. Selecting a subset of data requires finding a region of interest, based on the branching pattern from a node, from the entire data set. In order to select unique sub-networks for study, a node pattern was chosen and all nodes matching that pattern were found. The topology was represented as a directed graph originating at a particular node and proceeding to its children and to its children and to their children to some specified depth. While this model quickly creates a signature for a node, the node is not guaranteed to be unique. The probability that the signature becomes unique increases with greater depth or increasing the generations of children considered.

5.1. Results

Five sub-networks were selected for study from the West Frisian Wikipedia language set. The data sets were randomly picked from a list of 20 sub-networks that all had an out degree of 31. A network analysis of the 5 data sets was performed yielding the results shown in Figure 9.

West Frisian Dataset	Nodes	Links	Correlation Coefficient	Ave Path Length	Degree Dist (deg)	Deg Dist (vertex)	Clustering Coefficient
A	607	2186	0.022	135	4.7M	3.3M	0.000290
B	953	5579	-0.059	72.8	7.5M	6.8M	0.00181
C	1086	31,750	-0.033	24.2	29M	21M	0.0231
D	1199	33,943	-0.031	5.63	34M	25M	0.0244
E	1144	35,541	-0.032	23.3	35M	26M	0.0236

Out edges = 31
Connected = Yes
Simple = No (loops exist)
Directed = Yes
Symmetric = No

Figure 9 Network Metrics for Sub-network Nodes

As can be seen from the figure, the sub-networks are connected, not simple, directed and not symmetric. Inspection of the data shows similarities between data sets C, D and E in the areas of node and link counts, degree distances, and average clustering. After the similarities were noted, the networks were compared to their associated subjects. Networks A and B were "Magnetic North" and "Winter Sports". The networks designated C, D and E, however, all dealt with the same topic: Month of March. This correlation between (1) clustering, (2) degree distances, and (3) node and link counts, suggests that data mining in the form of network metrics analysis can be used to identify similar events.

5.2. Methodology

The nodes with out edges of 31 were selected without any bias of knowing the subject of the node. There were 20 of these nodes with 31 out edges. From these 20 nodes, 5 nodes were selected for additional network analysis. The traditional network properties were calculated to characterize the sample set.

After the characteristics were tabulated, the data were inspected for trends, similarities and differences. After the blind inspection (not knowing the subject matter of the five sampled datasets), conclusions were made based on data trends. This approach allowed the identification of nodes that matched a given signature as a method to choose a region for more in depth analysis.

In summary, the Node Signature Matching approach (1) identified a useful signature, (2) scanned the network for nodes that matched the signature, (3) created sub-networks with the matching nodes as roots, and (4) performed detailed analysis on the sub-networks.

To perform the node pattern analysis, a network topology had to be selected. Any network topology is permissible as the node will always map to a tree structure based on the number of children in a particular node. This mapping schema is applied regardless of where the node exists in the network. In this mapping they are the children of the particular node and hence a tree.

For this model to become useful in a calculation that can be performed by a computer a coding method for the topology had to be developed. Figure 10 shows the scheme used for encoding the patterns.

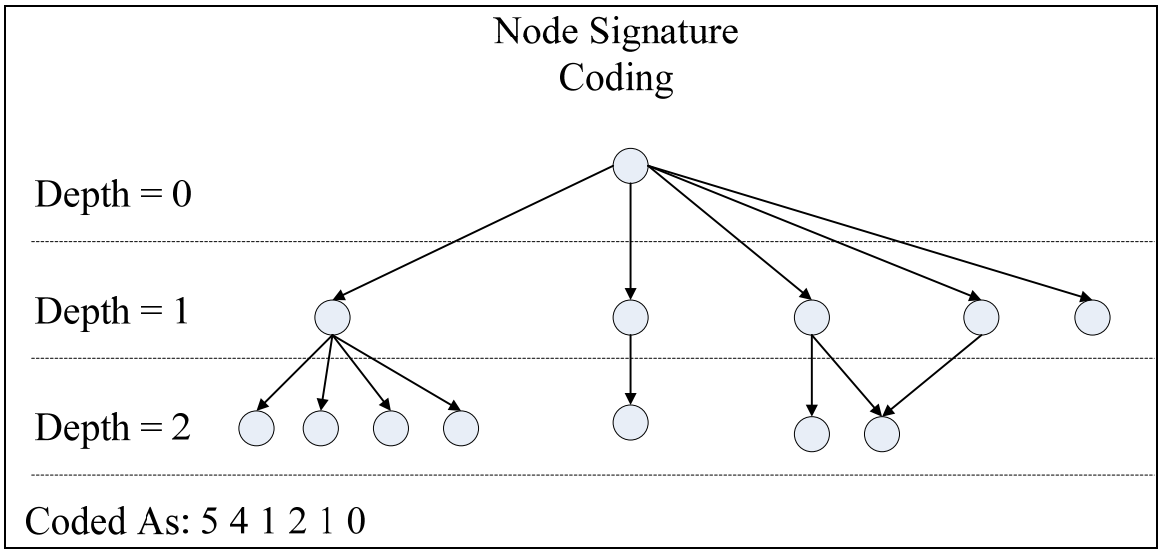

Figure 10: Node Signature Coding

6. Network Pattern Analysis

For the final part of the Wikipedia analysis, the pattern information of an entire language set was summarized into a single network signature. This signature analyzes the node signatures of an entire network and compiles them into a network metric. This is analogous to representing a specific frequency in the overall "frequency content" of the network. A Fourier series does the same type of analysis for signal frequency content.

6.1. Results

The signatures of the West Frisian and Latin language networks were compared. In order to generate a signature, the signatures of all nodes to a depth of two were found using the following method:

1. For each node with x children, increment the count of those nodes. This means that if 2 nodes had 31 children then at the x position of 31, y would have a value of 2.
2. Repeat the process for the second depth of the network.
3. Plot the resulting surface map.

The results of this analysis can be seen in Figure 11 and Figure 12. The x-axis in both figures is the number of children. The z-axis is the depth or layer of the network, and the y-axis is the number of nodes that have the specified number of children and the specified depth. The key takeaway from these figures is that the two sets of data produce distinctly different signatures.

While this analysis is still highly simplified, the results indicated that more detailed work could produce classes of signatures allowing the comparison of a network under examination to a known archetype. This comparison would in turn facilitate analysis.

An example of such an approach can be shown through a historical event. In his book Code Breakers, David Kahn relates a story from World War II. Long before the Allies understood the messages being communicated by the Japanese navy, they were able to tell the difference between "normal" and "attack" mode because they realized that *who* was sending messages to *whom* changed as the mode changed. With aggregate analysis, the Allies determined the mode of the Japanese fleet. This determination enabled the Allies to react appropriately. Similar value can be derived from such an approach in the current day conflicts. The network signature approach may be able to help identify the transitions from one mode to another.

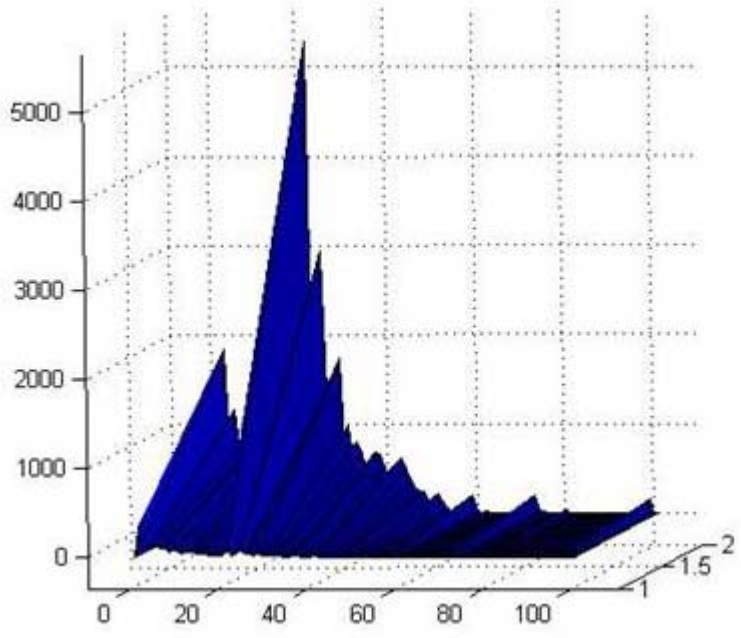

Figure 11: Surface Map of the West Frisian Network Signature

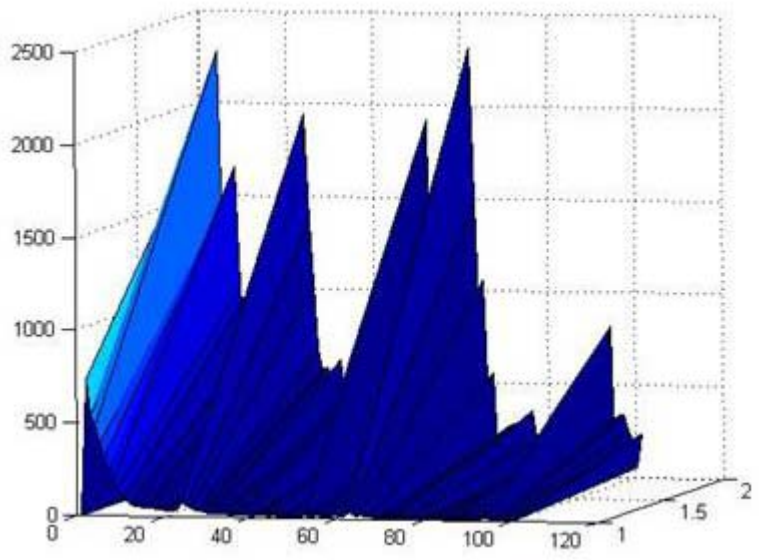

Figure 12: Surface Map of the Latin Network Signature

6.2. Methodology

In this phase of the analysis, the tradeoffs between the extent of the analysis and the memory and disk usage--as well as the time to run algorithms--were an important concern. While the code written for this assignment allowed simple network analysis, more work would need to be done for deeper studies of the data. In some instances during this section of the project (when the data set being analyzed was small), less efficient algorithms were used because of their greater ease of implementation. As a result, the methods generated for the network signature analysis should be seen as proof of concept rather than a final approach. While generating more sophisticated algorithms, particular attention should be paid to growth in data storage space and processing time.

7. Conclusions

7.1 Hypotheses

(1) When the network is small, pages with more important substance are added and result in more connections which increase the average number of degrees.
(2) Each node has a unique signature--much like each person has a unique fingerprint.
(3) A network has a unique pattern signature.

7.2 Summary

Analysis of the growth and structure of Wikipedia data showed that new pages and links appear to be added randomly to form a sparse and unclustered network. Analysis also showed that using different languages is a good approximation for the study of network growth. However, the number of speakers of a language--as well as other social, economic and political factors--can influence how the Wikipedia network growth occurs.

Node and network signature analysis allowed treating a large network as a collection of smaller networks of interest. This, in turn, allowed use of traditional "small network" analysis tools. The results of the node signature analysis showed that even simple patterns can yield powerful results by identifying similar or the same subjects by looking only at the node structure. This method located matching pages in a Wikipedia language that no team member spoke.

The network signature analysis created a signature of a network by identifying simple node patterns. Such a method has potential implications for rapidly sifting through networks and identifying characteristics or developments over time that are significant.

It was determined that domain knowledge is highly useful and potentially required in order to conduct a successful node and network signature analysis. Understanding the network facilitates selecting meaningful patterns that can be studied.

While working with the Wikipedia data, it was determined that dealing with very large networks requires different algorithms and methods that respect computational complexity, time and space requirements. Adjacency matrices are not adequate for manipulating large matrix data sets; rather a sparse matrix representation as well as lists of node pairs and node children are required. The sparse matrix representation in particular proved to be a powerful tool for working with large sets of data. The Wikipedia data sets, and most likely many other large networks, are very sparse. For example, the English Wikipedia has over 5 million nodes most of which have less than 100 edges. That means most rows in a NxN matrix would have 99% of its entries be 0; the sparse matrix representation takes advantage of this fact.

While methods for dealing with large data sets were identified and utilized as part of this project, the team determined that visualizations of such massive amounts of data would not be possible. Many visualization routines expect on the order of 1000 to 5000 nodes and cannot handle more. Some open source programs (such as Walrus) will visualize large networks, but that visualization is not highly useful as it cannot be changed and manipulated to gain insight about the data. Until significant additional work is done on algorithms to visualize and study large networks, scientists and engineers will be limited to specific subsets of analysis or to working with subsets of data.

8. Future Work

Several opportunities for future analysis of Wikipedia data exist. These include:

1. Modify the pattern matching algorithms to include a method for selecting approximate and not only exact matches.
2. Author dimensions could be added to the analysis. The community of Wikipedia authors could be proportional to the size of the Wiki. Or,
3. User dimensions could be added to the analysis. Determine how the number of users grows as a Wiki language or subject matures.
4. Editor/author behavior. Learn about the characteristics of Wiki users who primarily create new articles versus those who primarily edit existing pages and subjects. Do create to edit ratios have any meaning?
5. Analyze the evolution of one or more Wiki language networks. Currently, limited data is available on the temporal aspects of the language growth. However, by contacting the page managers, additional data may be accessible.
6. Establishing communities. Because Wikipedia subjects and languages can grow very quickly, the Wiki data might provide a platform to develop theory about how communities can be established quickly.
7. Death of a Wiki. What user and author behaviors indicate the fall, demise, or death of a Wiki subject, language, or entire Wikipedia community. How can self-forming/managed teams learn from the death of a Wiki?