Wikipedia: working with large networks

Justin Lindsey
Dave Long
Alex Mozdzanowska

May 9, 2006

Project Goals

- Analyze and understand the Wikipedia networks
  - Understand the growth patterns of Wikipedia by analyzing different languages
  - Chose and analyze subsets of data by selecting specific repeating patterns
  - Analyze the patterns of an entire language network
- Work with a large set of data

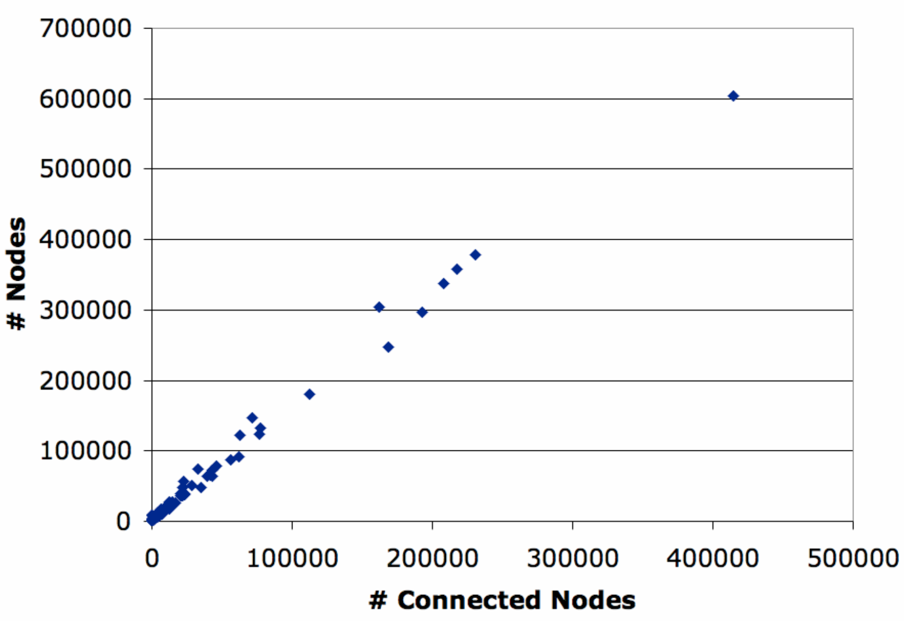
Wikipedia Properties

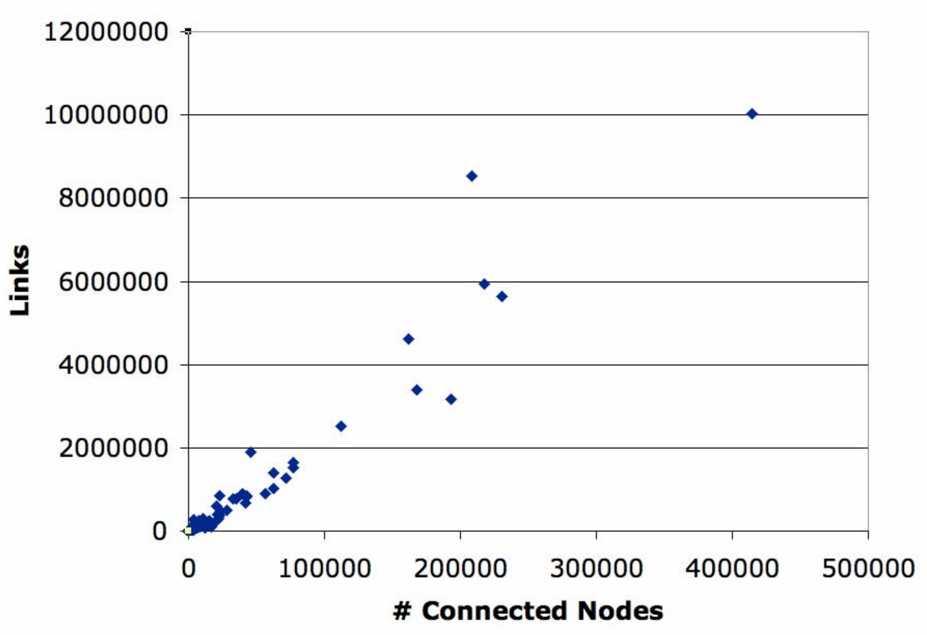
Data: Wikipedia Online Encyclopedia

- Nodes: Wiki pages
- Links: html links between pages
- 206 of the 229 Languages included in the study
    - Excluded are:
        - 10 smallest
        - 2 largest (German and English)
        - 11 Languages with no data

Language	# Nodes	# Links
Smallest	1,544	3
Largest	604,611	10,017,000

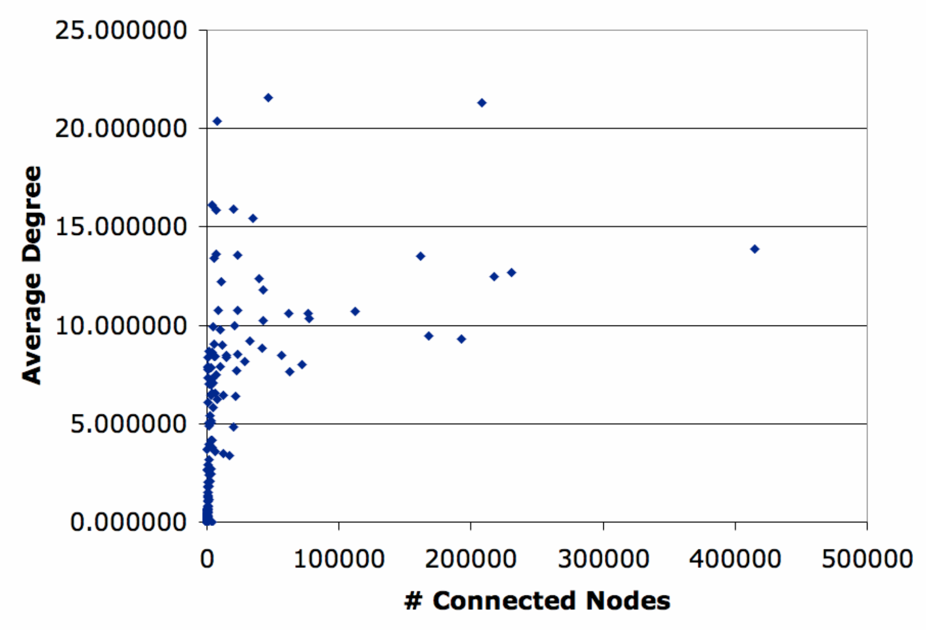
Nodes and Links

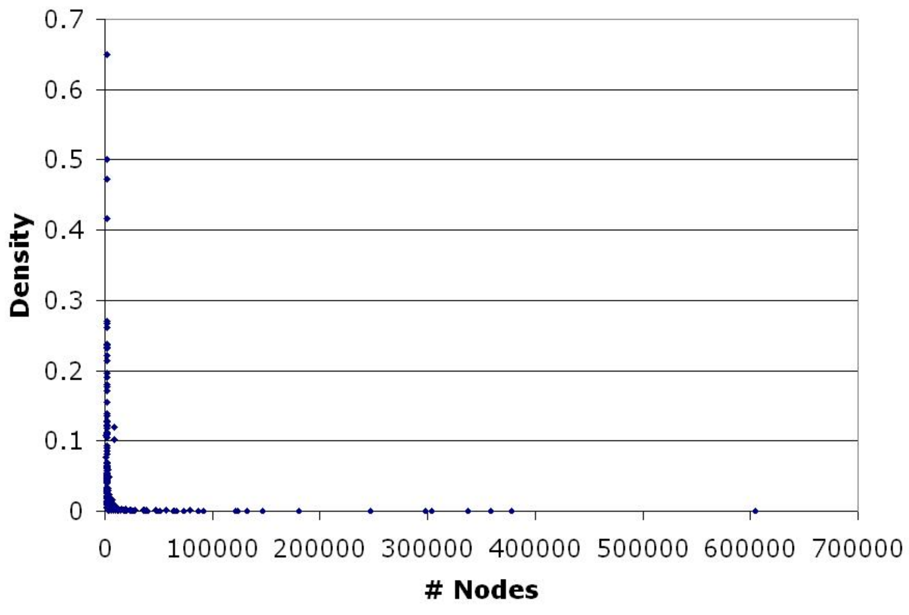




- Number of Connected Nodes Used to Represent Total Nodes
- Different languages give an indication of how Wiki networks grow over time

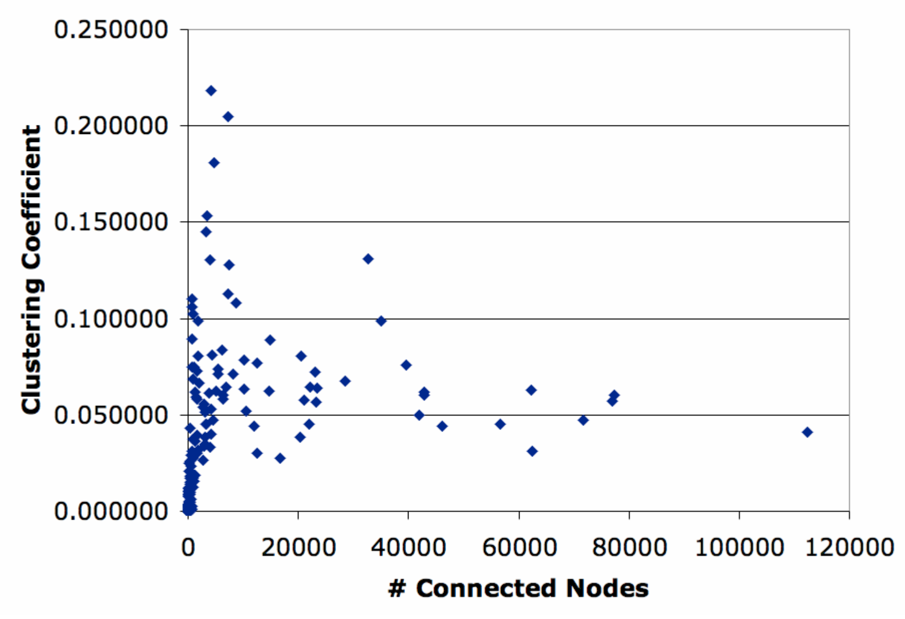
Average Node Degree and Network Density

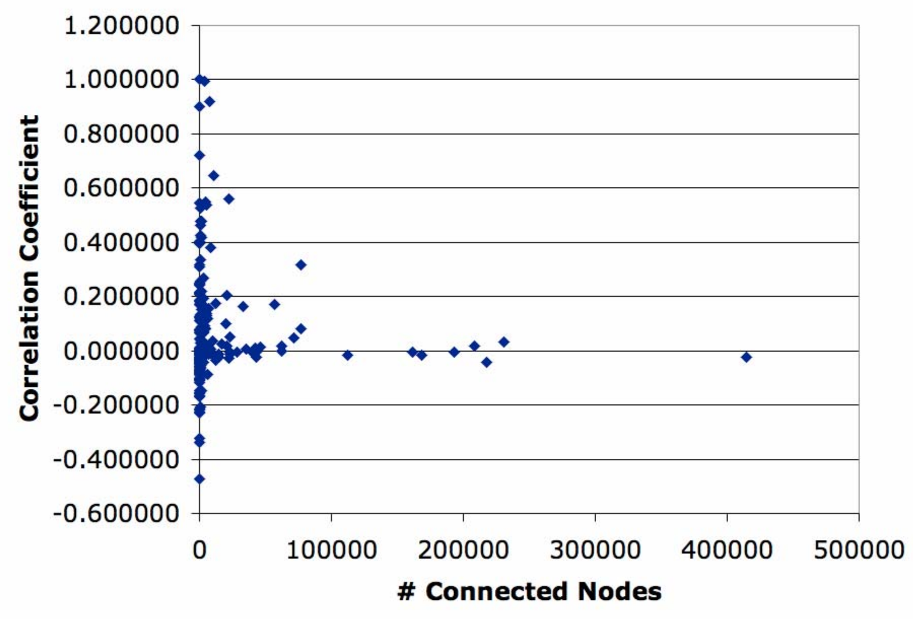




- Node degree increases during initial growth, but growth slows
- Density is high initially, but drops off to close to zero

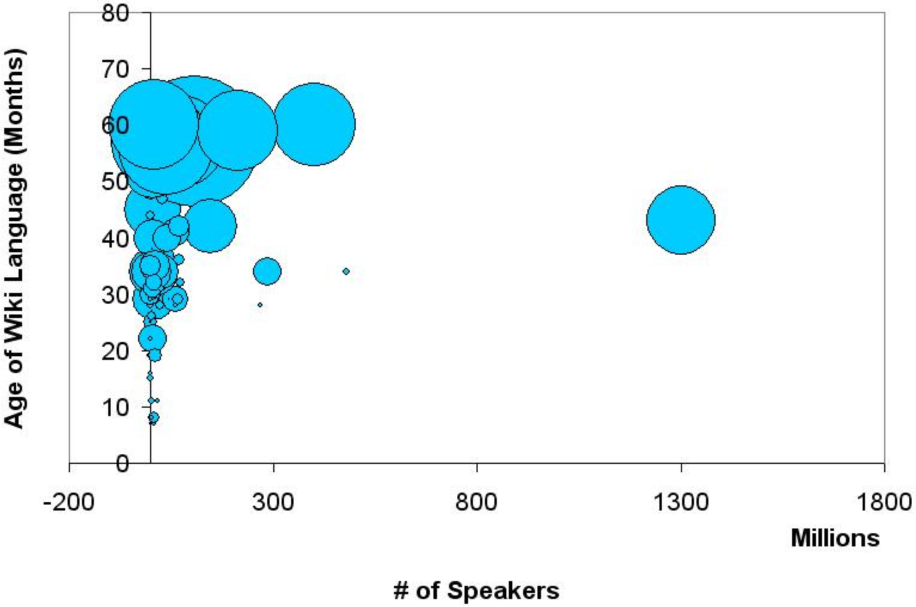
Clustering and Correlation Coefficients

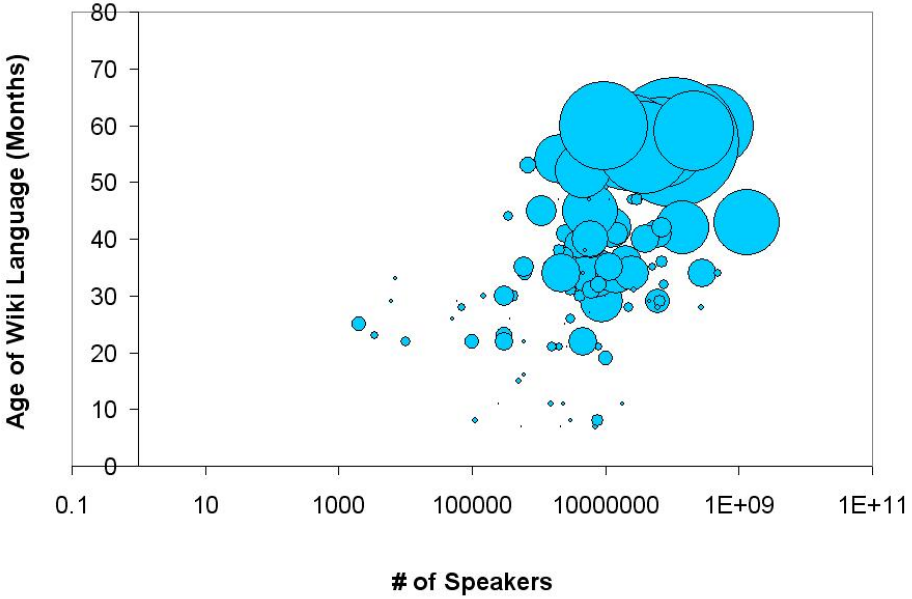




- Clustering and correlation settle around a constant number as the networks grow
- Low clustering and correlation may indicate that networks grow in a random fashion

Size of Network as Determined by Age and Speaking Population





Pattern Analysis

Data: Wikipedia Online Encyclopedia

- Nodes: Wiki pages
- Links: html links between pages

Language	# Nodes	# Links
Latin (LA)	~10 thousand	~100 thousand
West Frisian (FY)	~6 thousand	~72 thousand

Network Analysis Approach - I

Match Analysis - This Approach is focused on finding a region of interest and zooming in for more analysis

The Approach:

- Identify useful signature
- Scan network for nodes that match signature
- Create sub networks with matching nodes as roots
- Perform detailed analysis on sub networks

Simple Signature – The 31

Example of approach I:

While looking at FY.wiki, we became curious about nodes that had the simple pattern of 31 out edges

- There were 20 such nodes in the network.
- We selected 5 of them
- We created 5 sub networks with one of each of the 5 matches respectively as root nodes.
- We then performed analysis on the sub networks.

Here are the results:

Network Analysis Approach - I

West Frisian Dataset	Nodes	Links	Clustering coefficient	Ave Path Length	Degree Dist (deg)	Deg Dist (vertex)	C_ave
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

Network Analysis Approach - II

Aggregate Analysis - This approach is focused on summarizing the state of the entire network. Much as you could use Fourier series to identify the frequency content of a signal.

The Approach:

- Identify appropriate “frequency content” equivalent
- Calculate Network Signature based on specific “frequency content” component
- Compare Network Signature to other Network Signatures of interest:
  - The same network over time (normal vs attack mode)
  - Comparable networks for which detailed behavior is known (classify a network as a specific type or in a specific mode)

Simple “frequency content” – Depth 2 – Out Degree

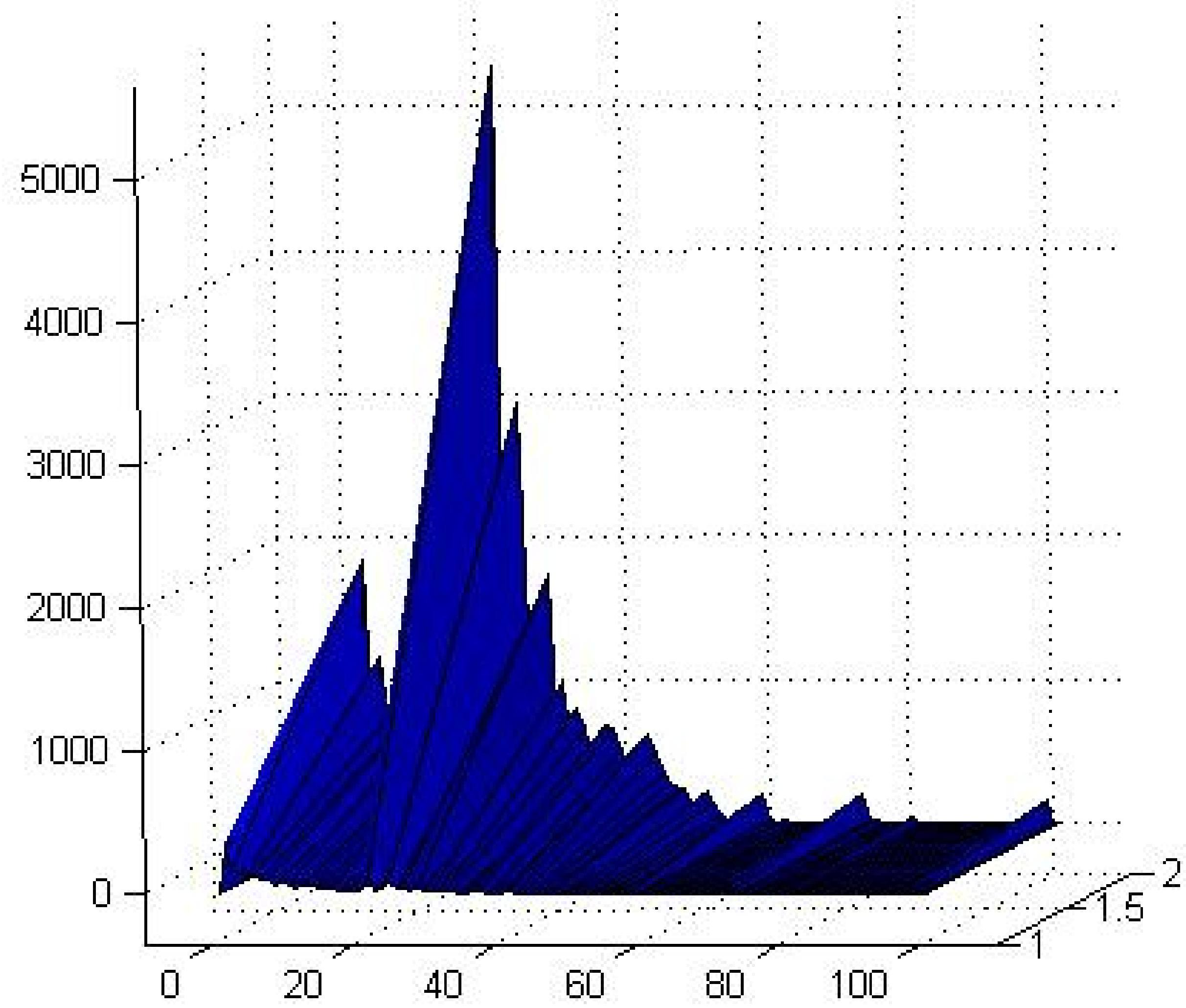
Example of the approach II:

In many command and control structures the number of reports (or people with whom they converse) that my reports (or people with whom I converse) have is a useful indicator

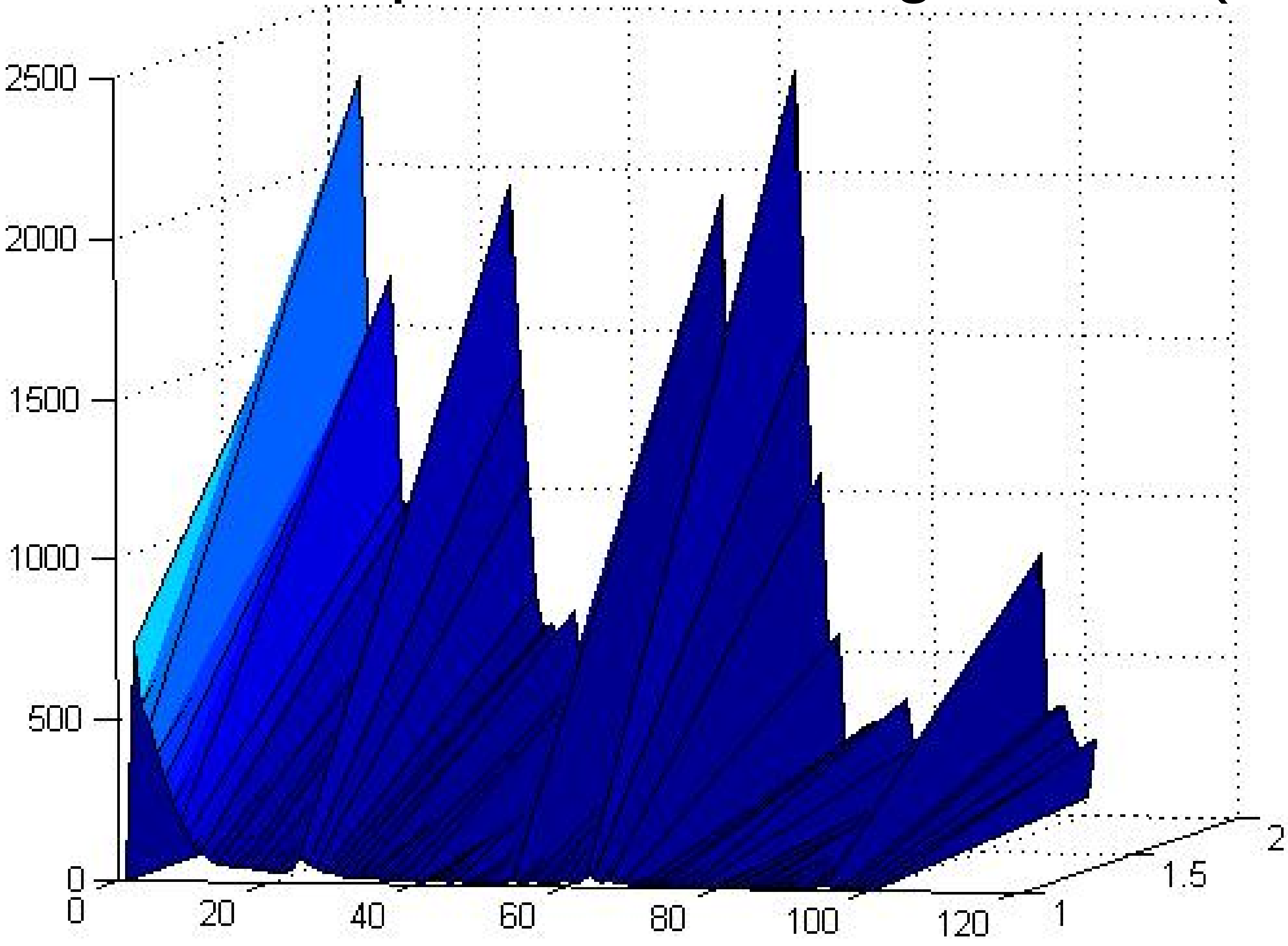
So we

- Selected 2 level Out Degree as a useful “frequency content” sifter
- We calculated the number of nodes that had a given out degree and then the out degree of each of their children
- Here are the results:

Surface Map of Network Signatures (FY)



Surface Map of Network Signatures (LA)



What we learned

- Dealing with very Large networks requires different methods that respect computational time and space requirements.
  - Adjacency Matrices don’t cut it – Sparse Matrices, Node Pairs, or Node Children lists are required.
- Match Analysis allowed us to treat a large network as a collection of smaller networks of interest which in turn allowed use of traditional “small network” analysis tools.
- Aggregate Analysis allowed us to finger print the state of large networks and compare in time and to other networks of interest.

Challenges

- In Match Analysis care must be taken to identify an appropriate signature that has sufficient relaxation as to catch more than the single node from which the signature was derived and still be a meaningful indicator
- In Aggregate Analysis, whatever “frequency content” sifter is selected will be applied over the entire network and can quickly explode computationally if not selected with care
- Domain knowledge is required to determine most interesting signature or “frequency content” to use for meaningful analysis

Possible Future Projects

Possible Continuations of Project

- Add an author and user dimension to the analysis
  - Look at how the number of authors grows with the size of the Wiki
  - Look at how the number of users grows and how often they visit the website
- Analyze the evolution of one or more Wiki languages networks
  - Limited data is available, contacting the managers of the page may be necessary
  - Develop theory for quickly establishing communities