Node and Network Signatures in Wikipedia
Project Status

Justin Lindsey
Dave Long
Alex Mozdzanowska

April 25, 2006

Project Goal

- Understand the structure of data on Wikipedia
- Work with a large set of data
- Conduct node and network signature analysis (to a specified depth)
  - Node signature analysis
    - Look for node structures that characterize the branching pattern from a node
    - Look for other nodes that have the same branching pattern
  - Network signature analysis
    - Analyze the node signatures of all nodes
    - Plot the distribution of node signatures for the network

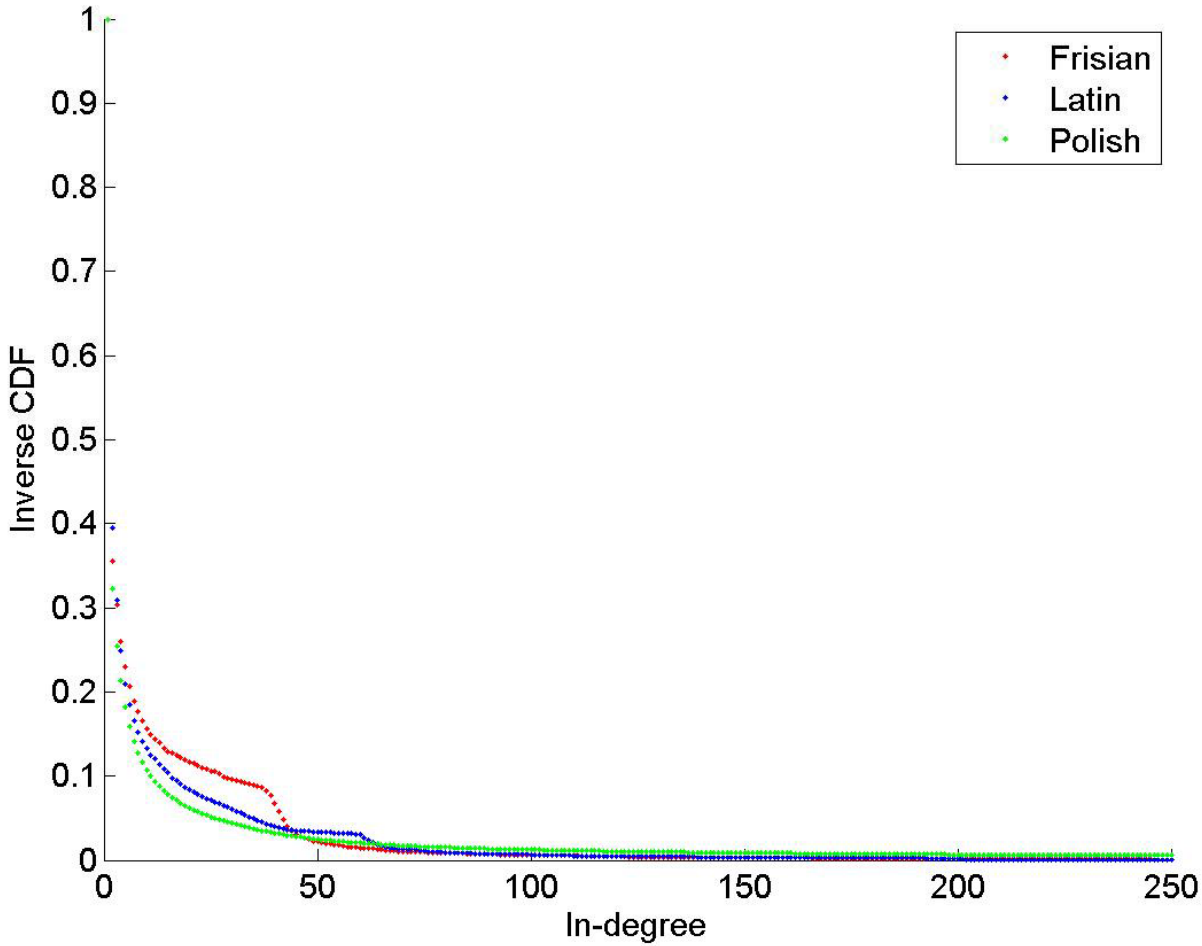
Data: Wikipedia Online Encyclopedia

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

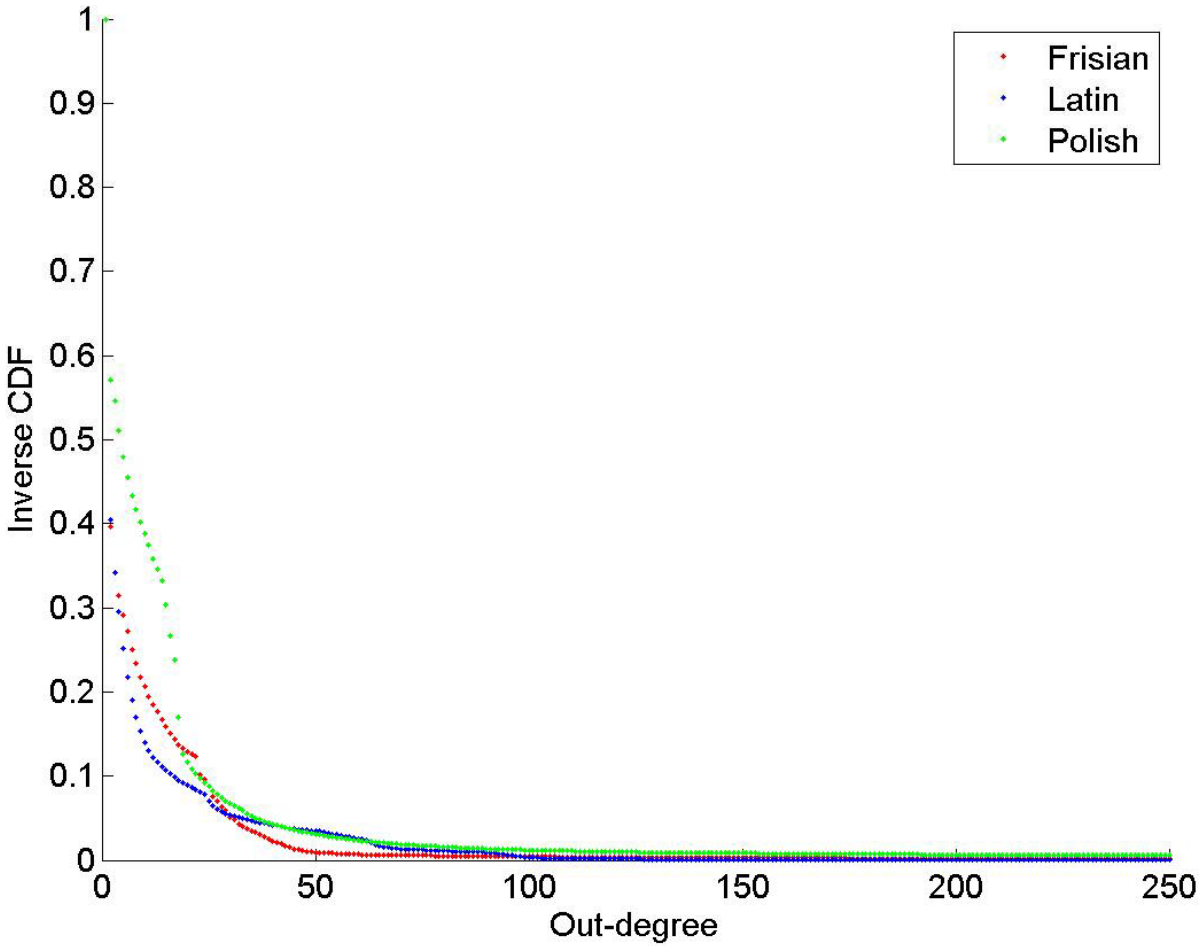
Language	# Nodes	# Links
English	~3.8 million	~46 million
Polish	~350 thousand	~5 million
Latin	~10 thousand	~100 thousand
West Frisian	~6 thousand	~72 thousand

- Visualization:
  - UCINET can handles up to 10,000 nodes
- Analysis:
  - UCINET can handle ~5,000 nodes
  - Matlab can import large lists of pairs, but can only perform certain computations

In Degree


Frisian
Latin
Polish
Inverse CDF
In-degree
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

Out Degree


Frisian
Latin
Polish
Inverse CDF
Out-degree
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

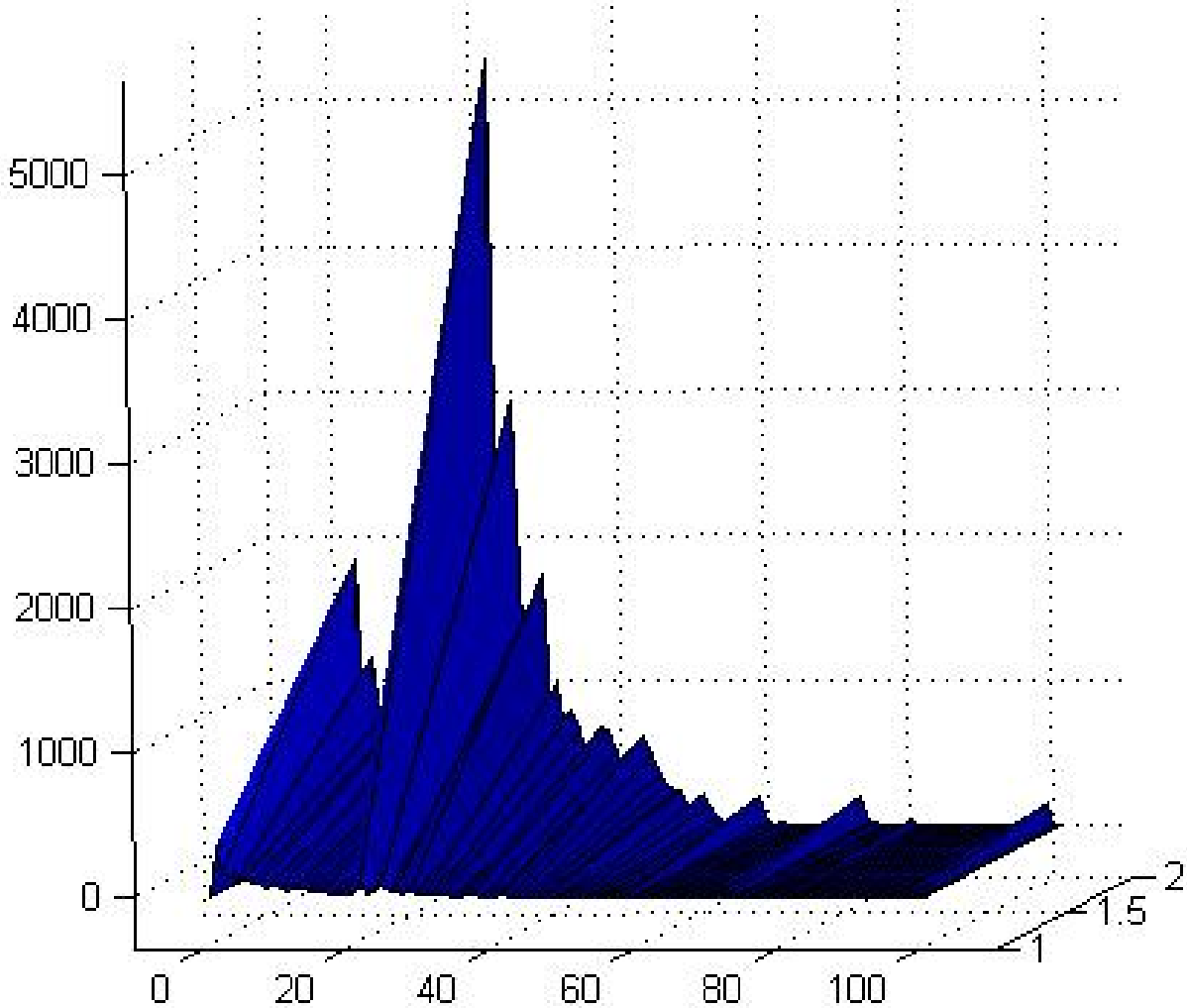
Average In and Out Degrees and Clustering Coefficient

	Average In and Out Degree	Clustering Coefficient
English	-	-
Polish	12.5	-
Latin	6.4	0.059
West Frisian	7.1	0.044

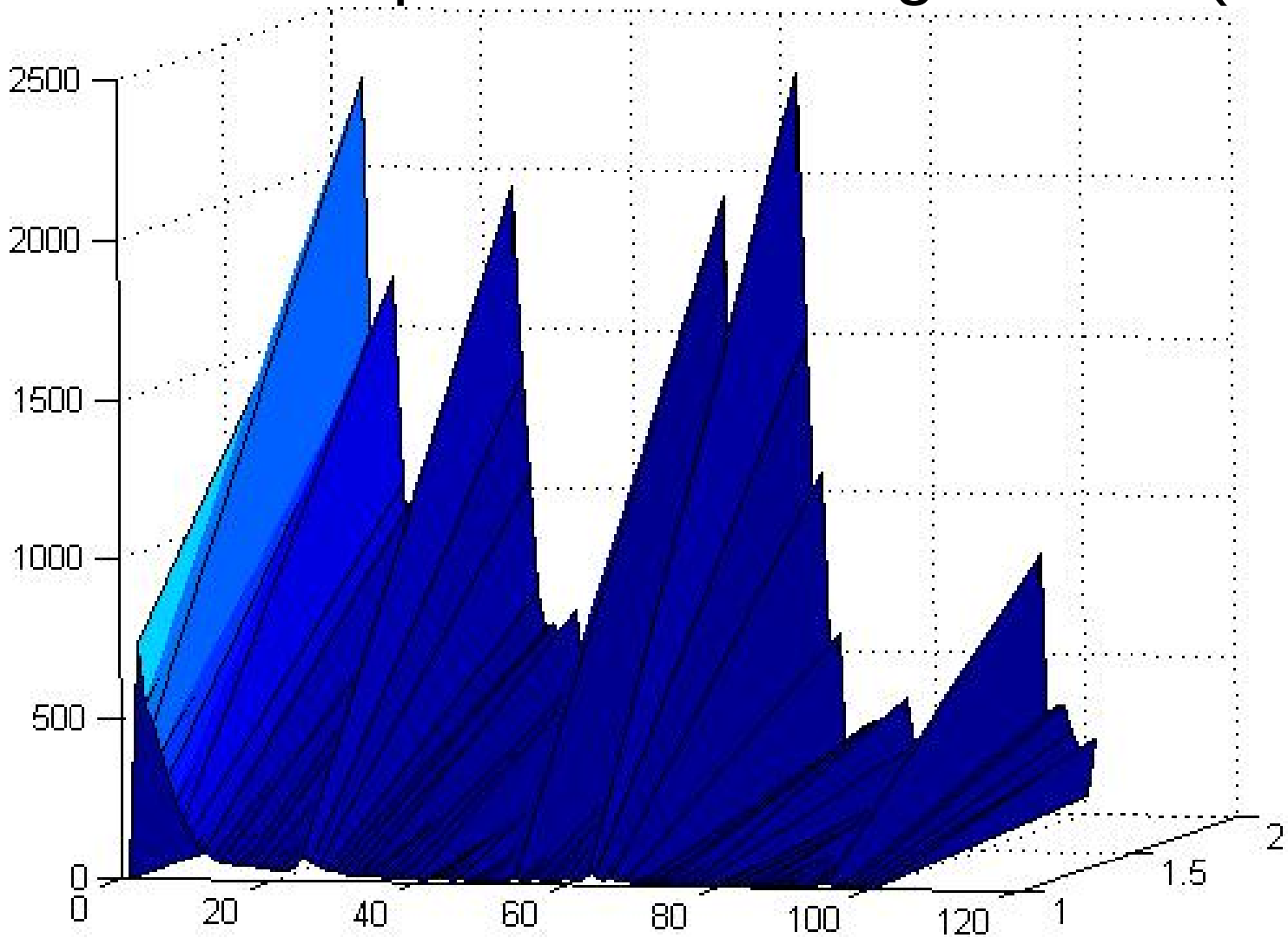
Code

	Order
Extract Wiki data (pairs and links)	O(N)
Pairs and links to node child list	O(NlgN)
Determine network signature	$O(N^{\wedge d}) \rightarrow dO(N^{\wedge 2})$
Find node signatures	O(N)

Surface Map of Network Signatures (FY)



Surface Map of Network Signatures (LA)



Applicability

- Developing node and network signature analyses which operate on large data sets
- Tools can be used to:
  - Study organization of information on wikipedia
  - Identify terrorist or drug cells by communication patterns
  - Identify patterns of activity within organizations during crisis management
  - Determine if military forces are in attack mode