The Architecture of Wikipedia

Project Status

Justin Lindsey
Dave Long
Alex Mozdzanowska

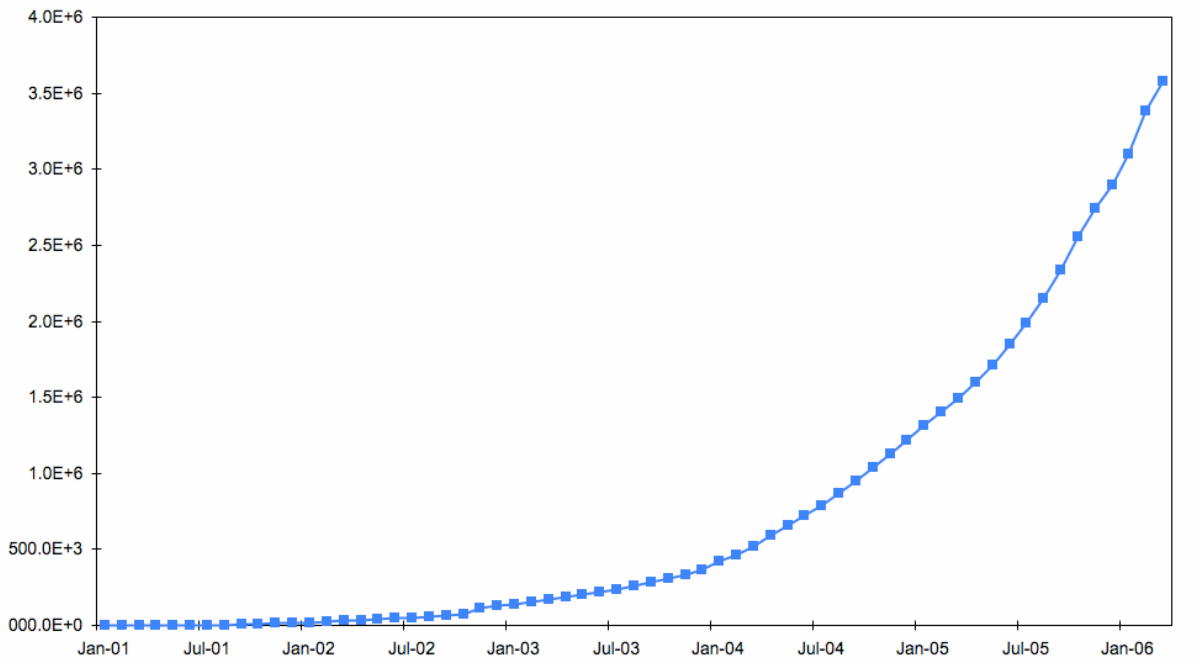
March 23, 2006

Overview

- Project Description
- The Data Set
- What We've Done – So Far…
- What's Next

What is Wikipedia

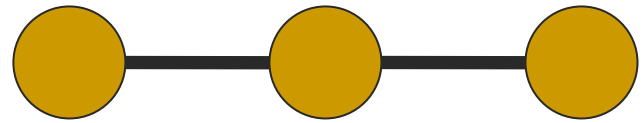
- Online encyclopedia
  - Can be edited by anyone who wishes to do so
  - Available in 214 languages
    - Language sites range in size from 1.03M articles to 1
- Wikipedia started in 2001 and has grown rapidly



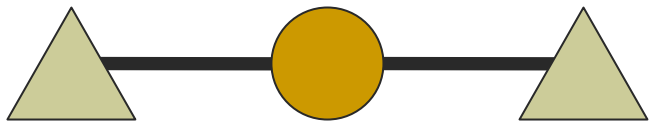
Project Description

- The Architecture of Wikipedia
  - Purpose: Advance knowledge of Interoperable Information-sharing environments and rich social network data by understanding Wikipedia architecture
  - How: Study relationship between the structure of information on Wikipedia and the authors who generate the structure
- Description
  - Analyze links between pages, between authors, and the emergent network
    - Pages linking to each other are linked
    - Authors that edited same page are linked
    - Pages edited by same author are linked

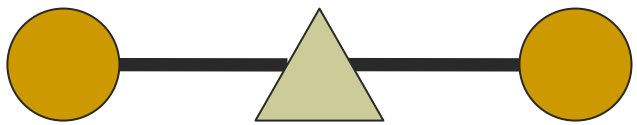
Taxonomy for Structures



Pages connected to each other
Actual link



People connected by a page
Inferred link



Pages connected by a person
Inferred link

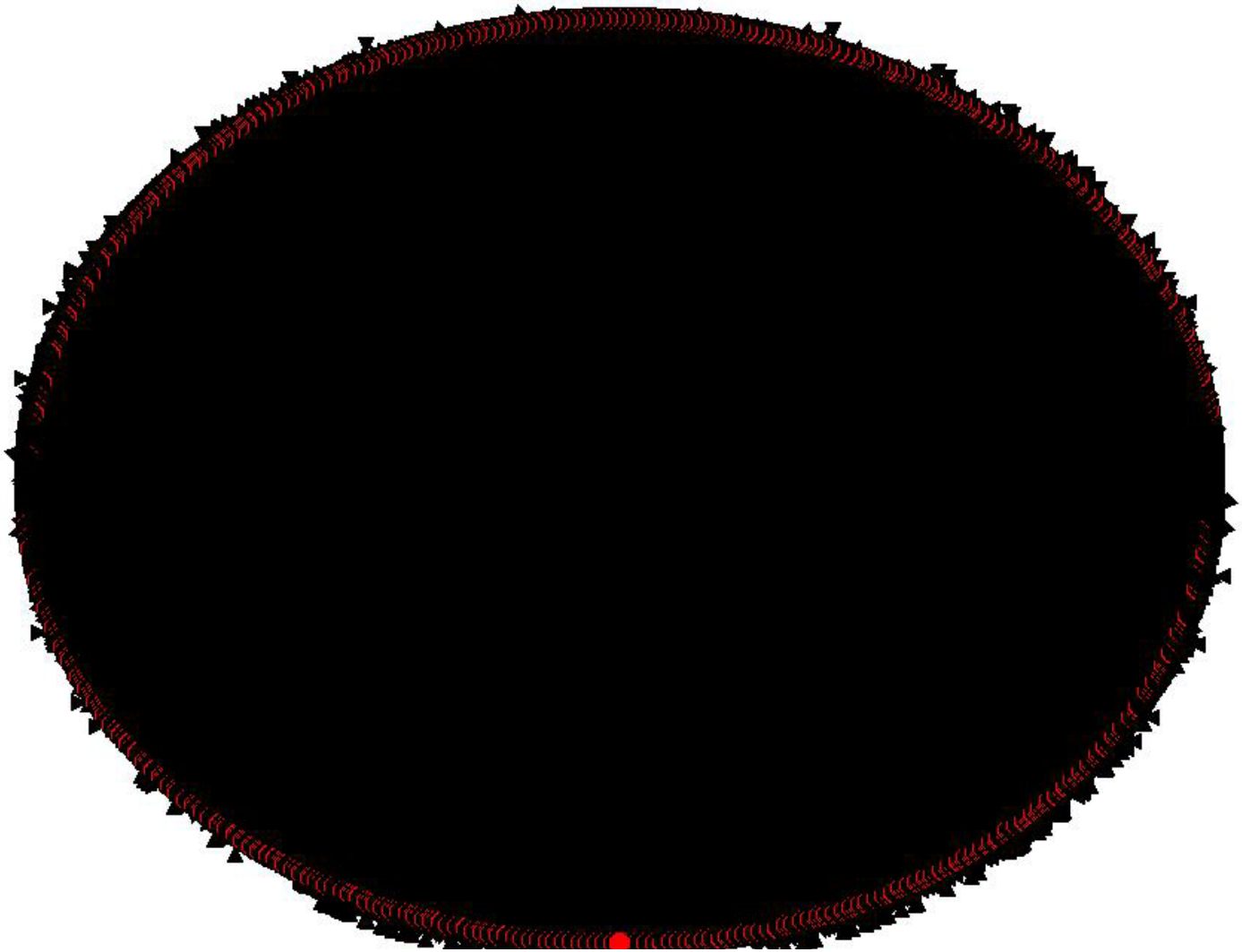
What we've done so far

- Identified a 3 layer network of pages and authors that will be studied
- Downloaded wikipedia data
  - Polish
  - Latin
  - West Frisian
- Transformed page link data into formats usable by Matlab and UCINET
- Began using page link data

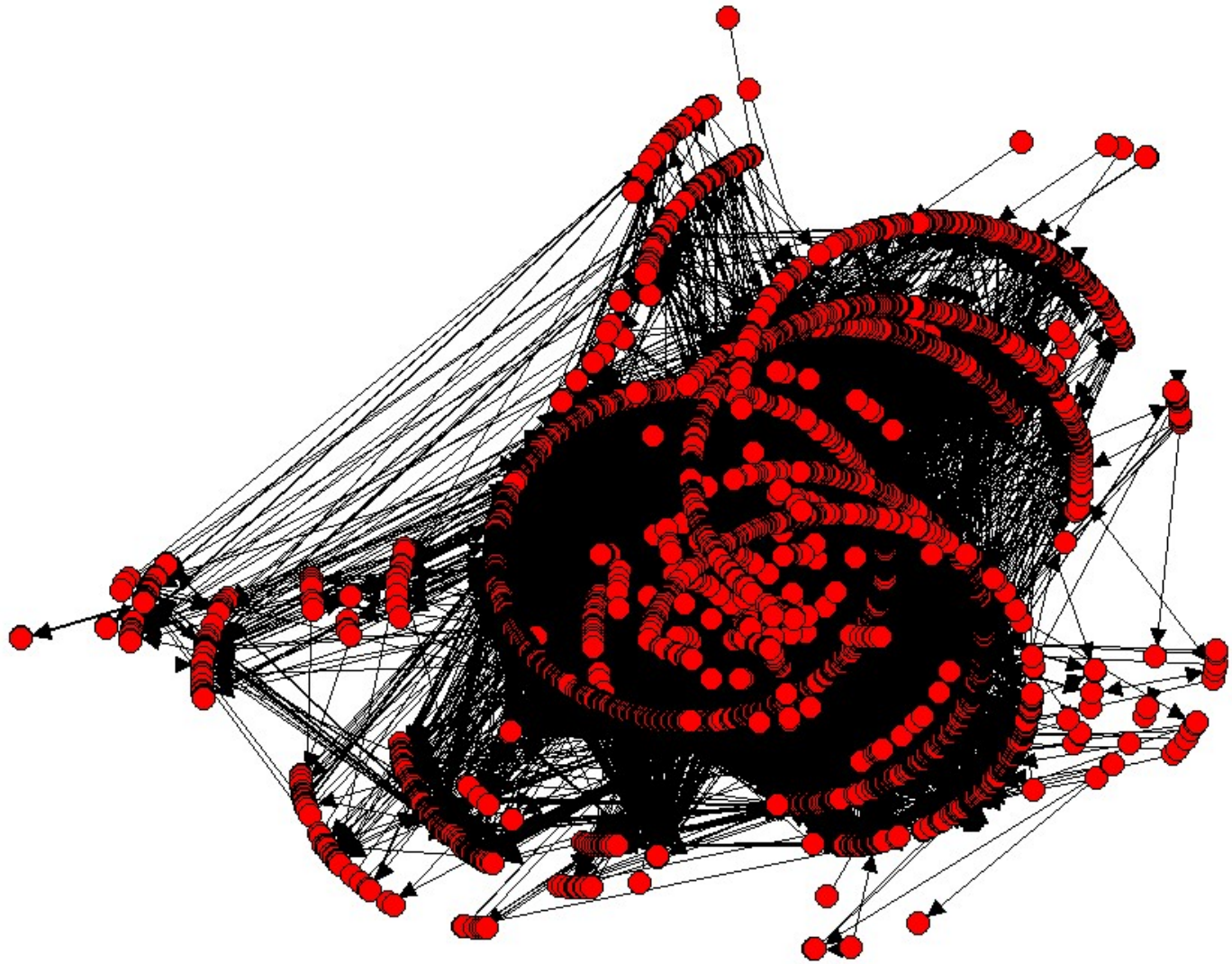
Data Challenges

- Wikipedia has huge amounts of data
  - The data can be downloaded
  - But computationally we don’t have the tools to work with it
- Attempted solution: work with a subset of pages chosen by language
  - English 1.03 million pages
  - Polish 200K pages
  - Latin 6K pages - 70K node pairs
  - West Frisian 2K pages
- Alternate solution
  - Chose a subset of data by topic

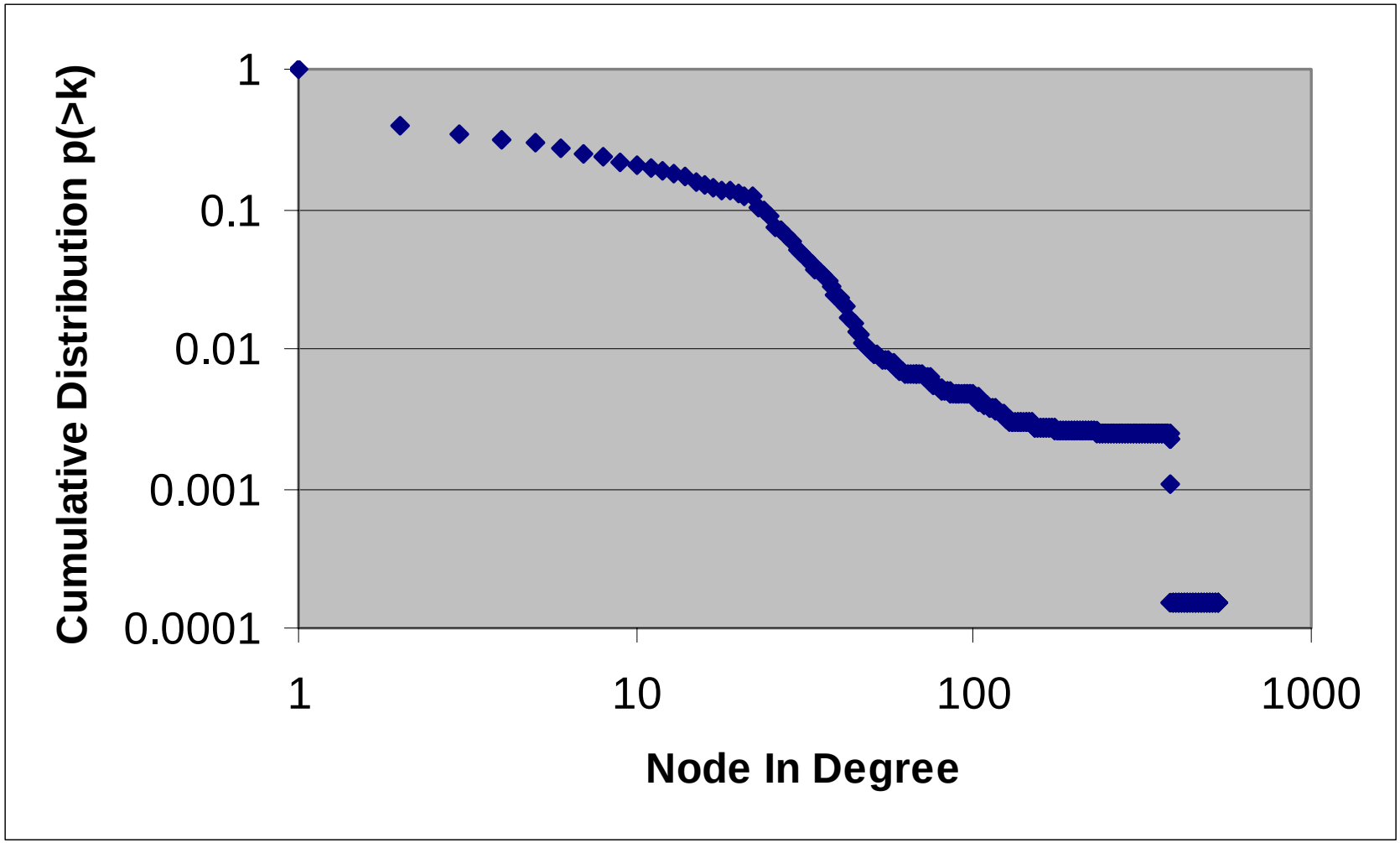
Network of Page Links (circle)



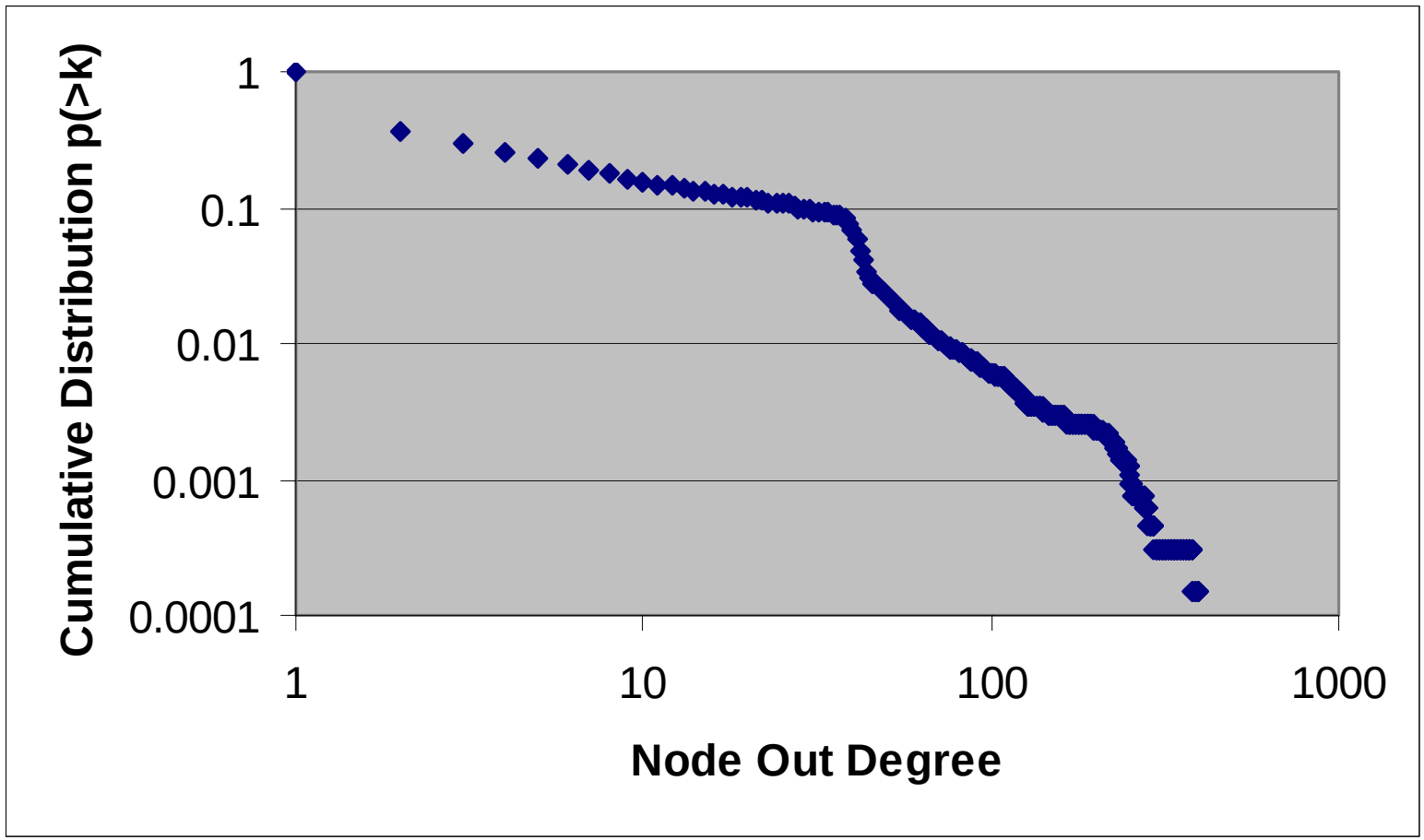
Network of Page Links (spring)



Network Statistics



Network Statistics



What's Next

- Transform author data into a usable format
- Analyze the three layers of the network separately to identify their properties and understand the structure of each layer
- Analyze and attempt to understand the connecting between the structure of the three layers