

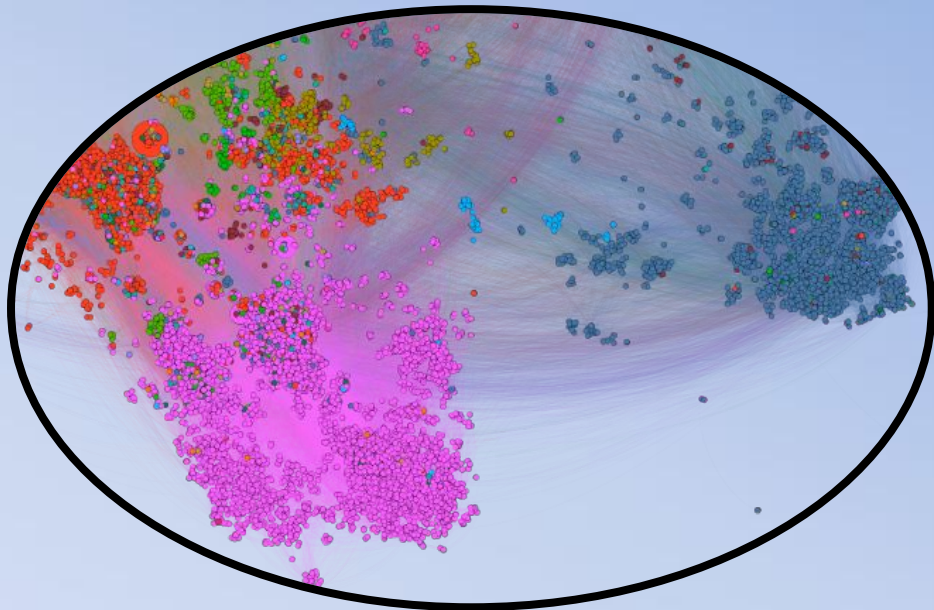
Exploring WikiLinks

Sean O'Neal

- Graph of Wikilinks
- Cluster Analysis
- Degrees of Separation
- Link Recommender System



WIKIPEDIA
The Free Encyclopedia



The Dataset

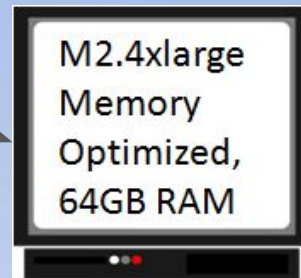
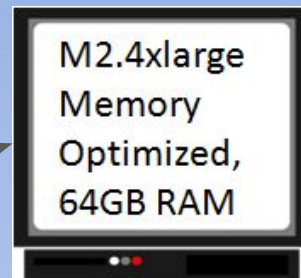
. **53GB**

- ~5 million articles
- ~150 million links

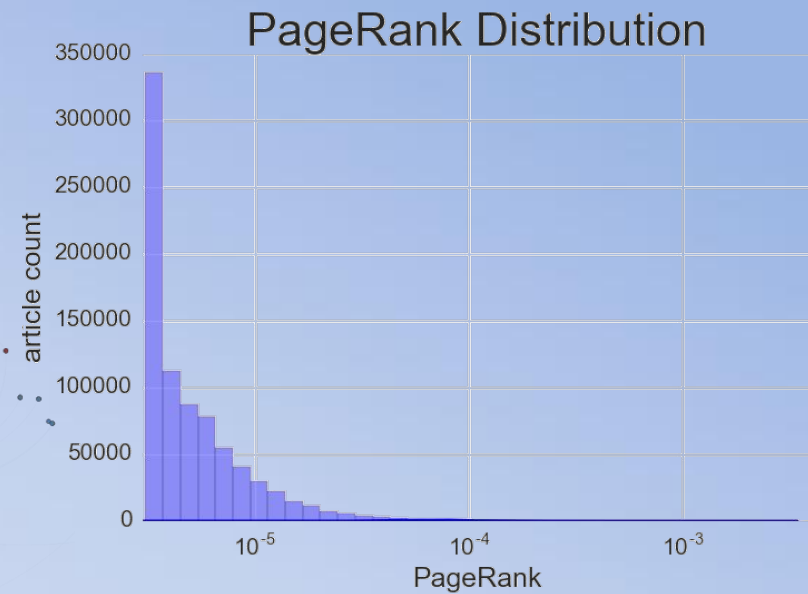
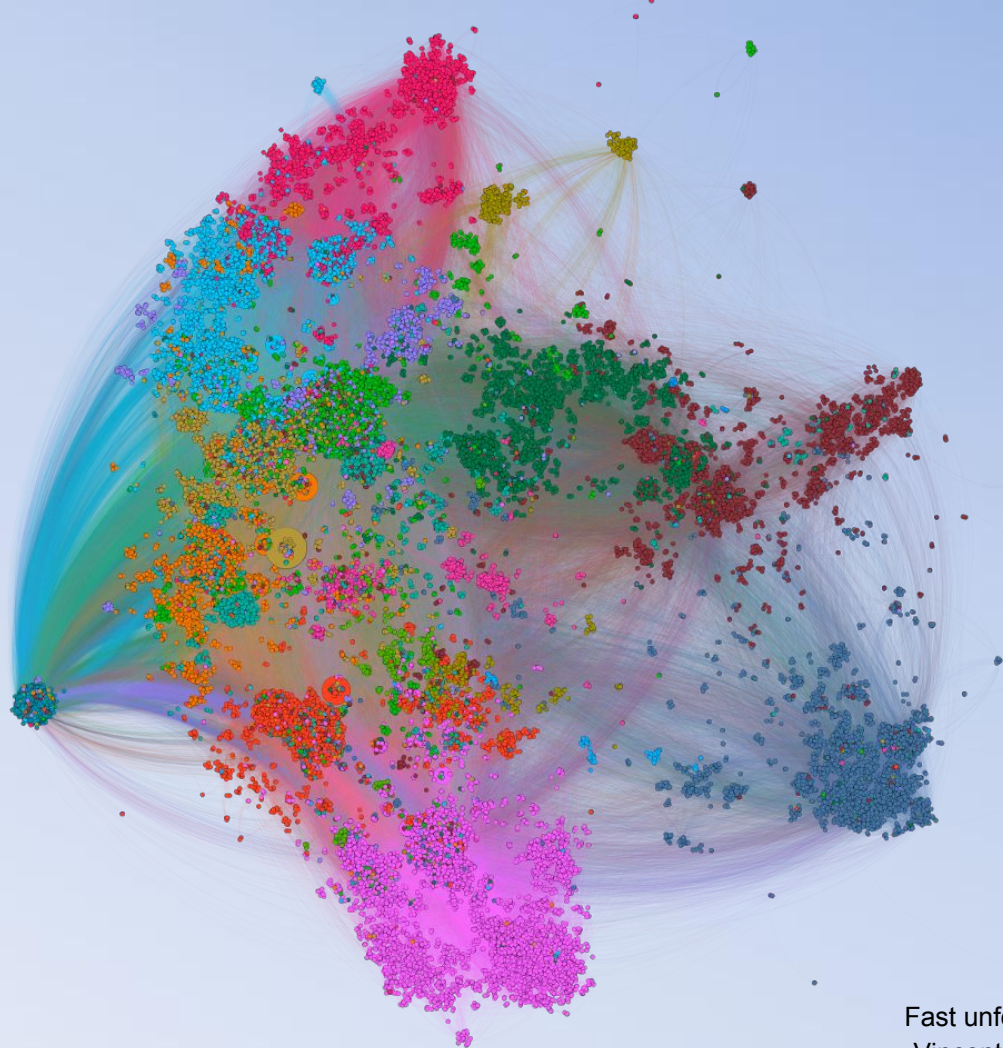
The Tools



Cluster Configuration

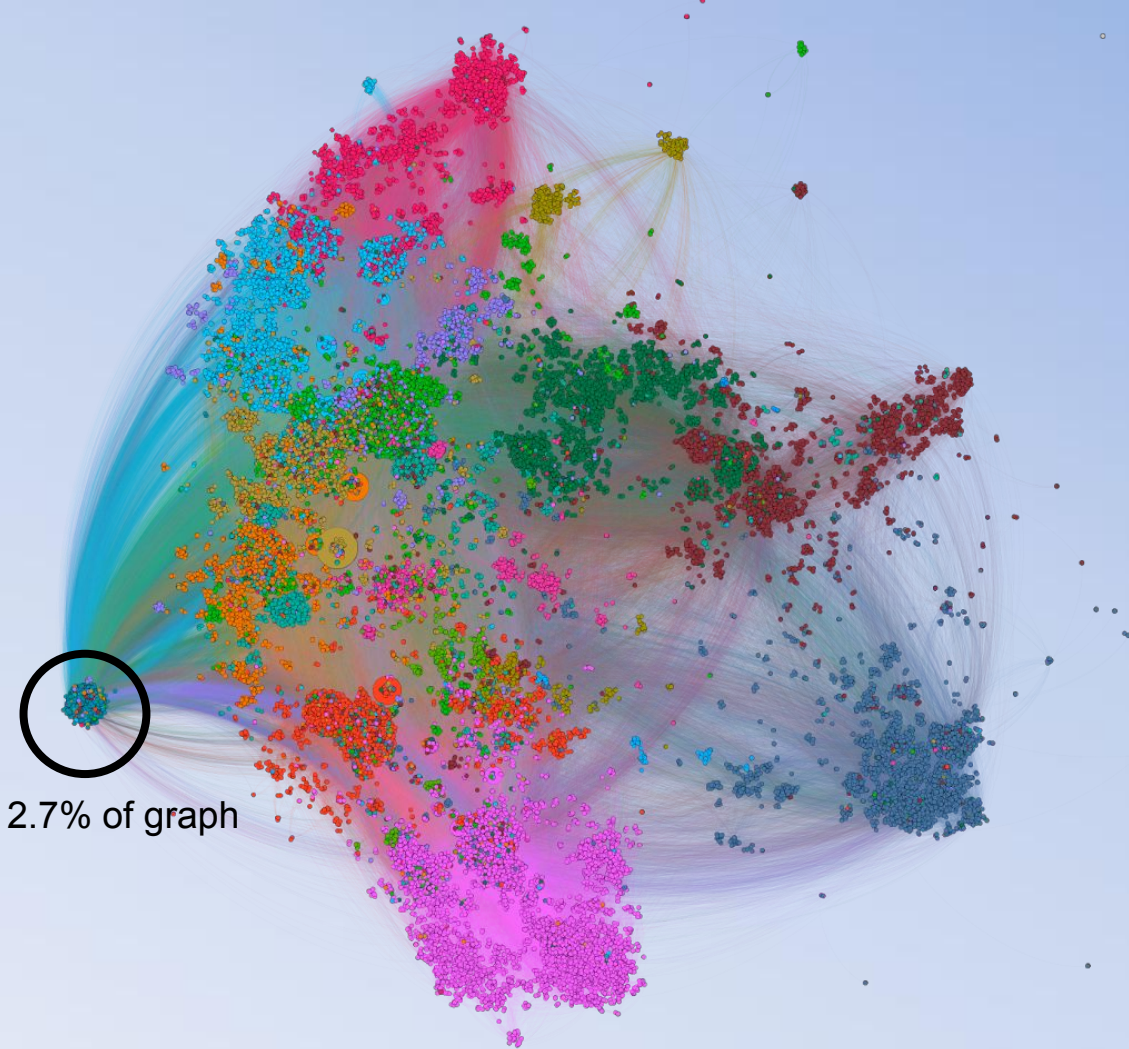


- 36 cores
- Memory optimized
- Minimize network communication
- Hold all of wikipedia in memory at once!



Fast unfolding of communities in large networks

-Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre



2.7% of graph

Top Nodes:

1967

1933

2014

November 20

March 15

1942

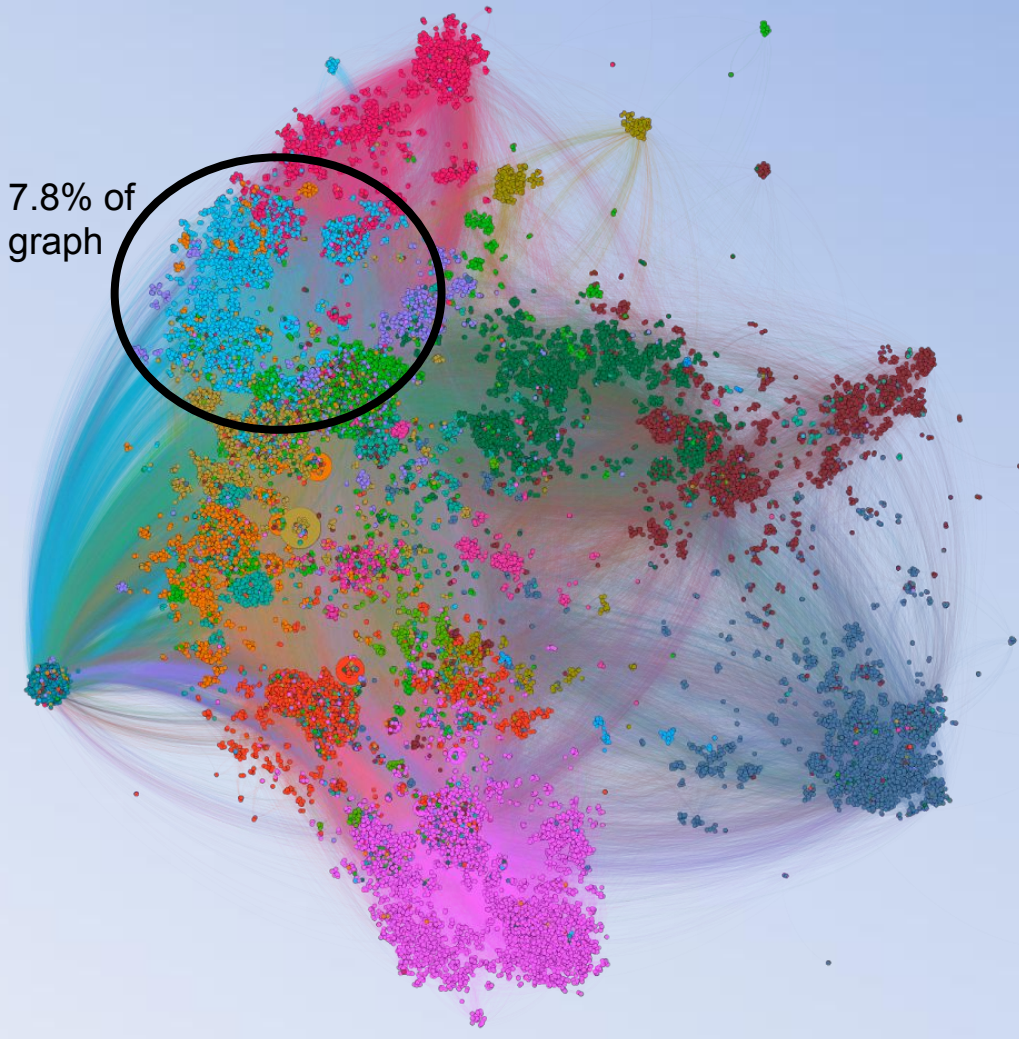
February 27

1941

August 25

December 25

Topic: dates!

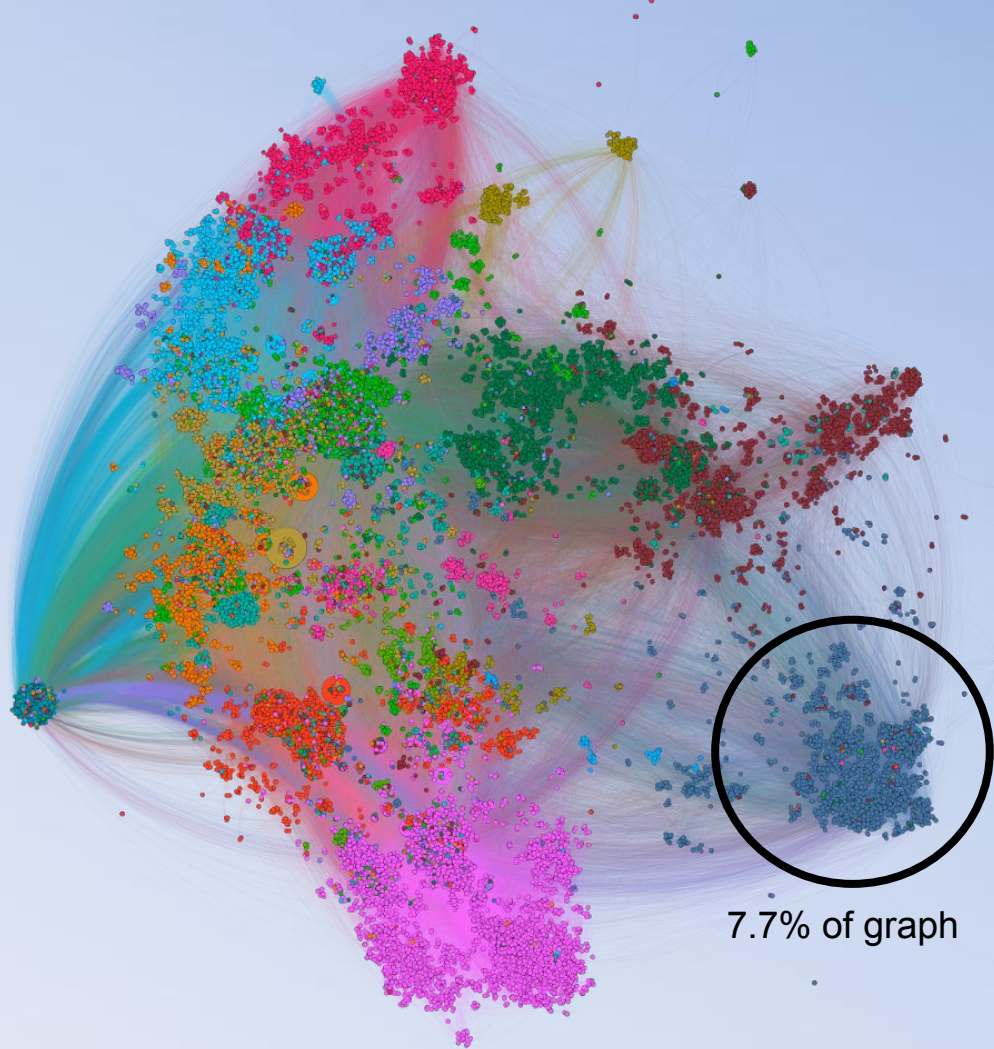


7.8% of
graph

Top Nodes:

Italy
Roman Empire
Christianity
Middle Ages
Judaism
Bible
Jerusalem
Constantinople
New Testament
Venice
Charlemagne
Pope
Alexandria
Mesopotamia

Topic: History/Religion?



Top Nodes:

Microsoft

Internet

Unicode

C (programming language)

Apple Inc.

Stanford University

MIT

World Wide Web

Intel

Java (programming language)

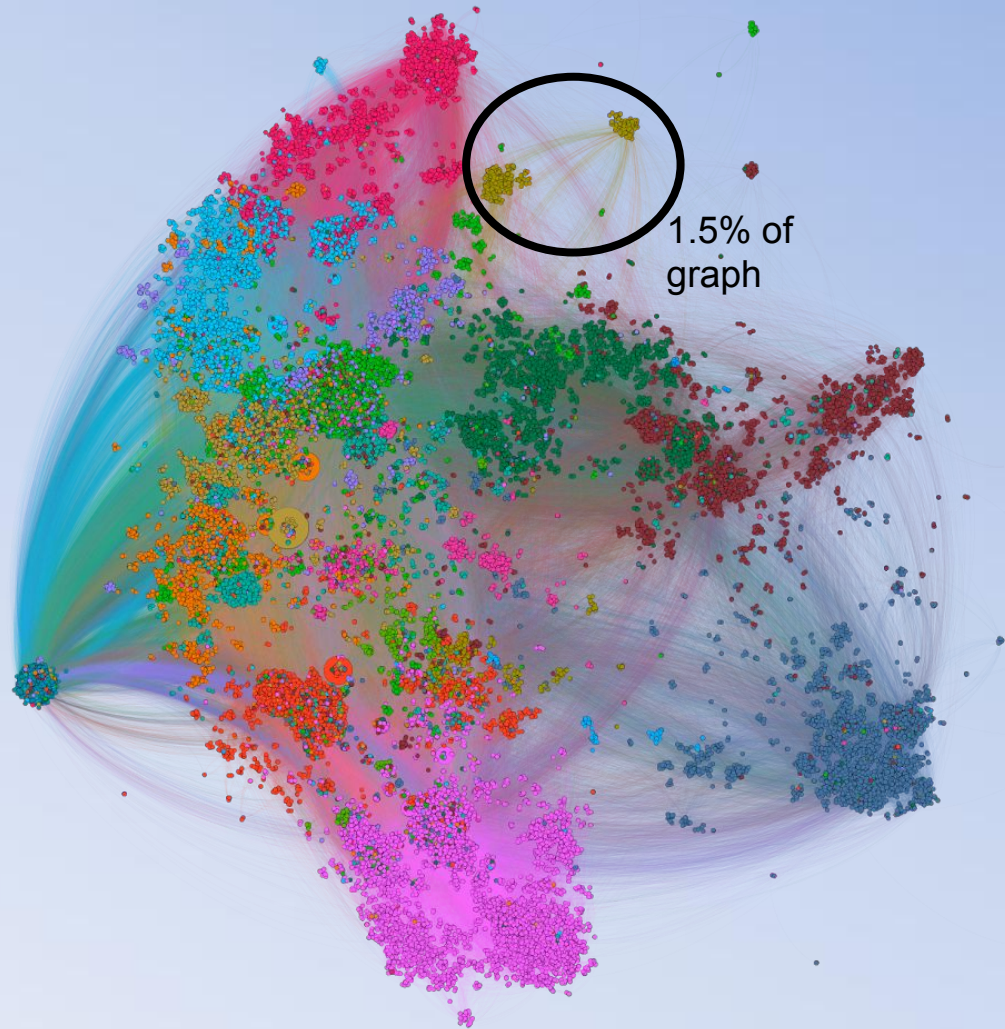
Portable Document Format

C++

Wired (magazine)

Python (programming language)

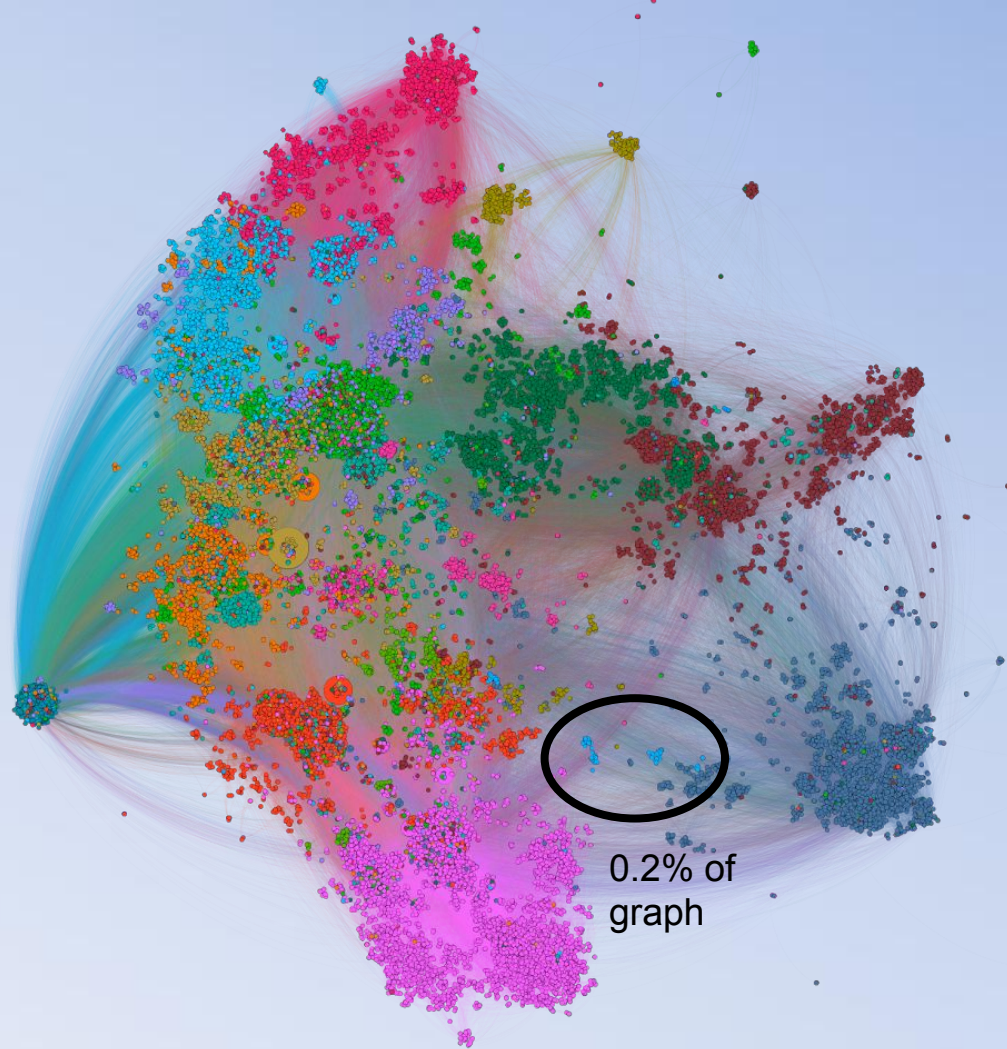
Topic: Technology



Top Nodes:

Old Norse
J. R. R. Tolkien
Isaac Asimov
Robert A. Heinlein
Prose Edda
Science fiction
The Lord of the Rings
Poetic Edda
Beowulf
Dungeons & Dragons
The Hobbit
Terry Pratchett

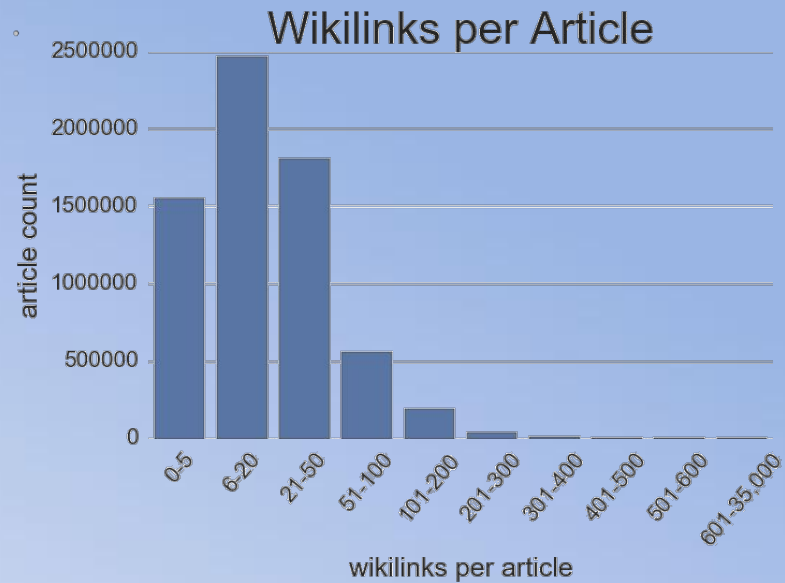
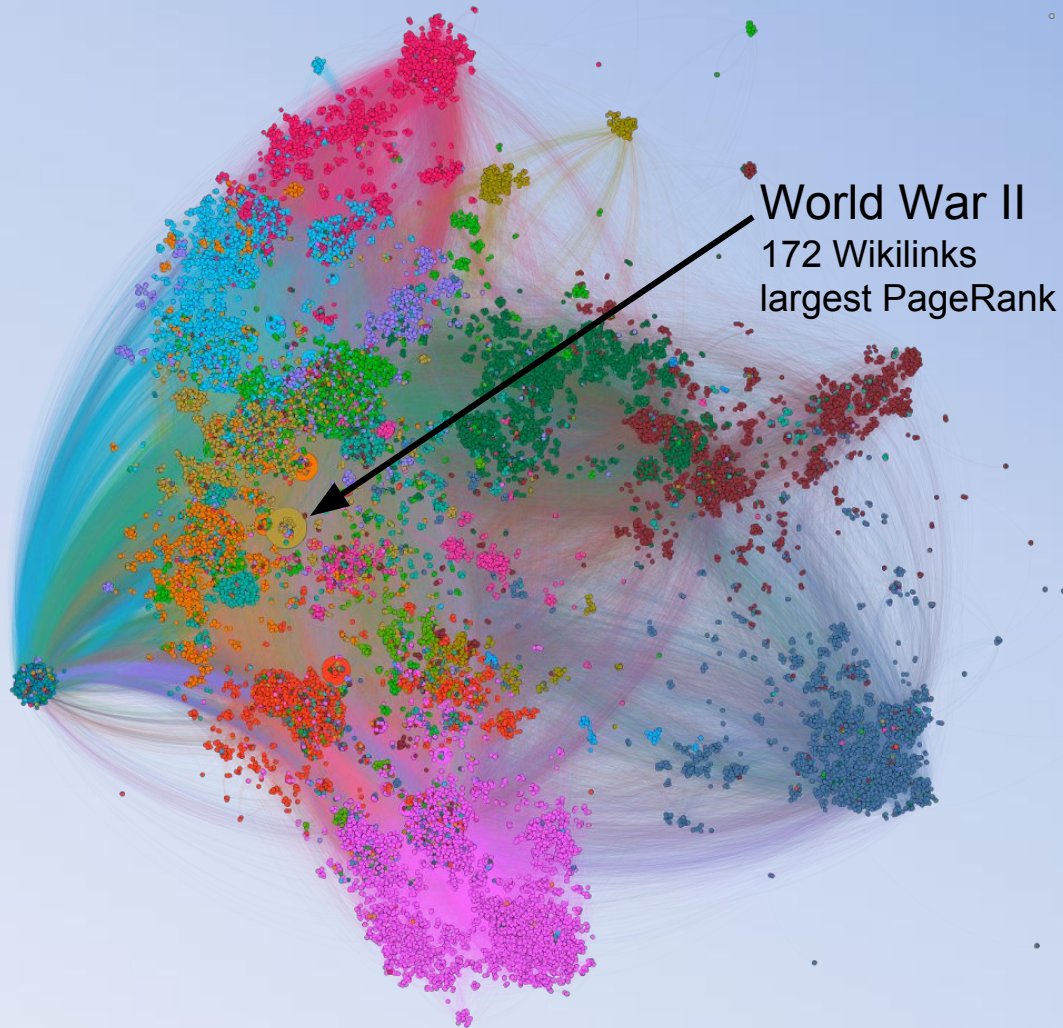
Topic: science fiction!



Top Nodes:

Texas hold 'em
Omaha hold 'em
Declaration (poker)
Bluff (poker)
World Series of Poker
List of poker variants
Draw (poker)
Out (poker)
Playing card
High-low split
Precision Club
Aggression (poker)

Topic: Poker!



Avg. Wikilinks per article: 23

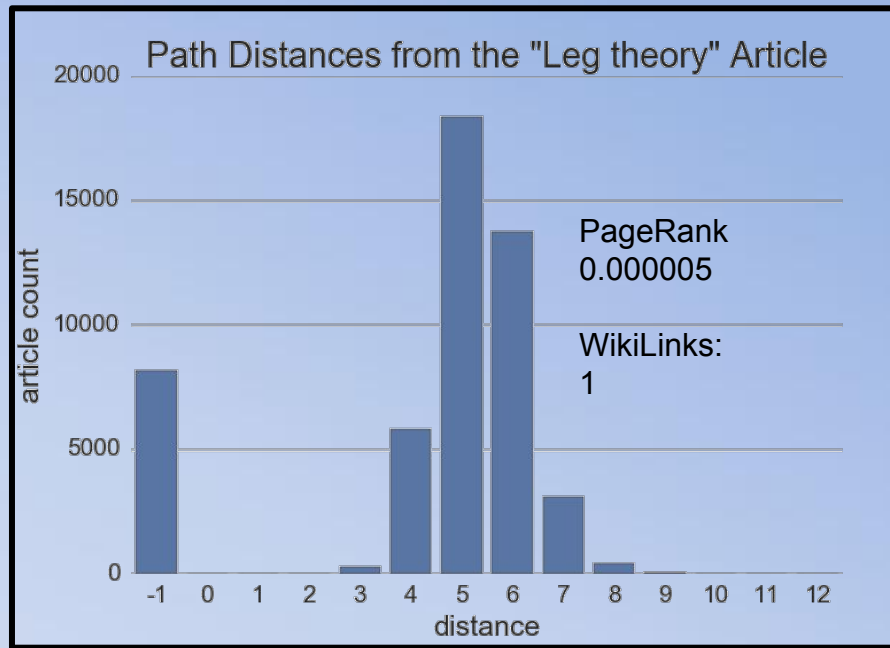
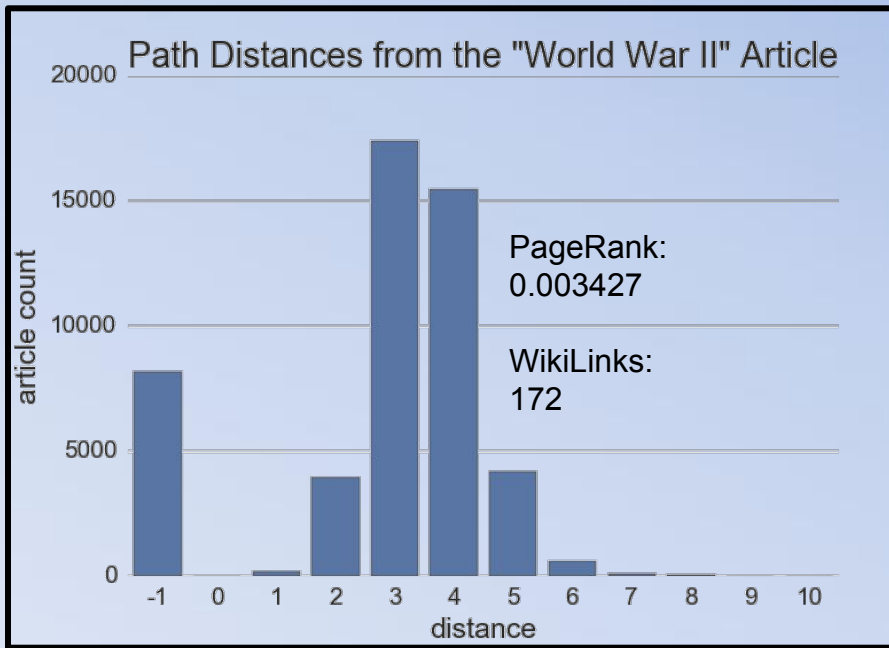
$$23^0 = 1$$

$$+ 23^1 = 24$$

$$+ 23^2 = 553$$

$$+ 23^{\mathbf{3}} = \mathbf{12,720}$$

$$+ 23^{\mathbf{4}} = \mathbf{292,561}$$



Most important nodes out of WWII:

Romani people	5313
Charles de Gaulle	2412
Netherlands	2309
Second Sino-Japanese War	1866
Mongolia	1842
Manhattan Project	1734
...	

No path exists between WWII and...:

Cliffhanger (film)...

Reticulated Python...

Economy of San Marino...

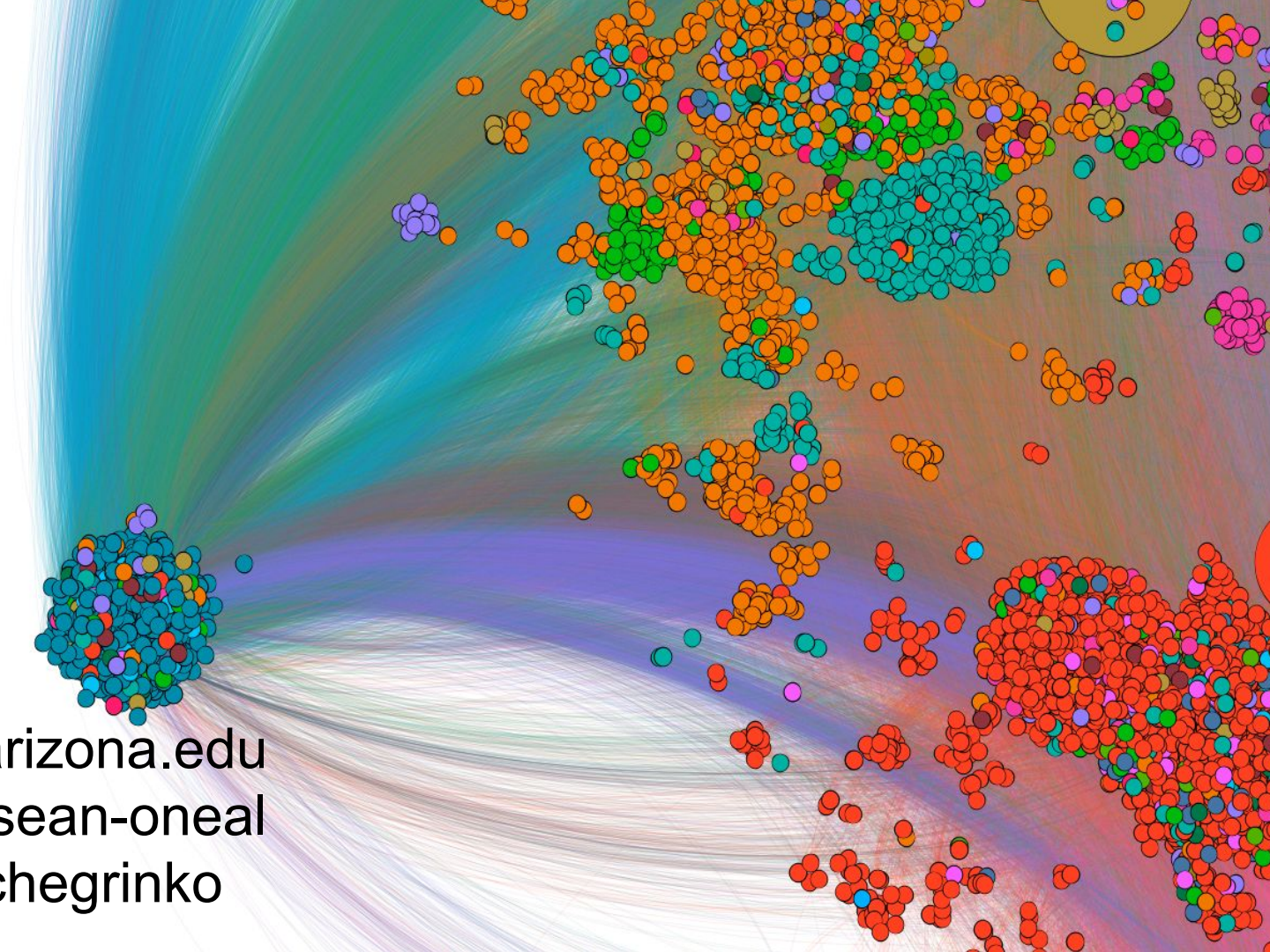
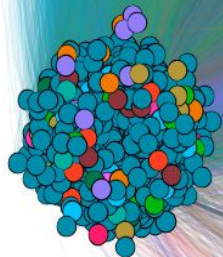
...

Recommender System:

(WWII is not, but should be connected to...)

<u>Article:</u>	<u># of articles that link to both:</u>
Germany	569
The New York Times	499
American Civil War	404
London	391
Russia	390
Japan	363
...	

Sean O'Neal



soneal@email.arizona.edu
linkedin.com/in/sean-oneal
github.com/vyachegrinko