Infrastructure for Supporting Exploration and Discovery in Web Archives

Jimmy Lin
University of Maryland
@lintool

Digital Preservation 2014 Meeting
Wednesday, July 23, 2014

Source: Wikipedia (All Souls College, Oxford)

… to building sh*t that works

Source: Wikipedia (Factory)

Source: https://www.flickr.com/photos/bongtongol/3491316758/

Source: Wikipedia (All Souls College, Oxford)

Source: http://images3.nick.com/nick-assets/shows/images/house-of-anubis/flipbooks/hidden-room/hidden-room-04.jpg

Why?

Restrictive use regimes?
But I don’t think that’s all…

Users can’t do much with current web archives…

Hard to develop tools for non-existent needs…

Source: http://www.flickr.com/photos/cheryne/8417457803/

We need *deep* collaborations between:

Users (e.g., archivists, journalists, historians, digital humanists, etc.)

Technologists (me and my colleagues)

Goal: tools to support exploration and discovery in web archives

Beyond browsing…
Beyond searching…

Source: http://waterloocyclingclub.ca/wp-content/uploads/2013/05/Help-Wanted-Sign.jpg

We Need Infrastructure!

Open-source implementation of Google's MapReduce framework for petabyte-scale analysis.

Open-source implementation of Google's Bigtable, which powers maps, gmail, …

Warcbase

An open-source platform for managing web archives built on hadoop and APACHE HBASE

http://warcbase.org/

Source: Wikipedia (Archive)

Warcbase Application Lifecycle

WARC/ARC

Ingestion

Processing & Analytics

APACHE HBASE

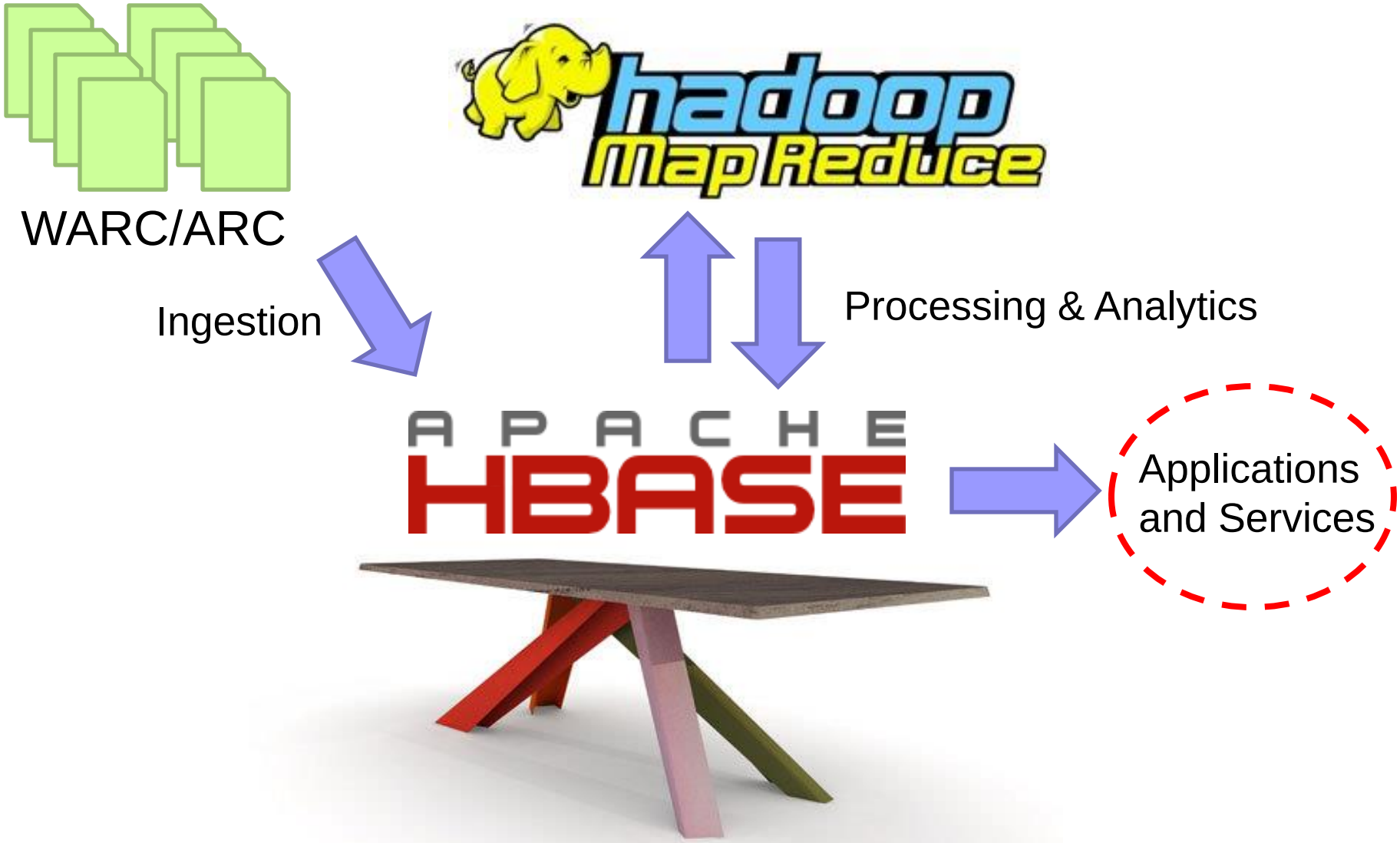
Applications and Services

Scalability?

We got 99 problems but scalability ain't one…

– Jay-Z

Demo Applications



Sample dataset: crawl of the 108th U.S. Congress

- Monthly snapshots, January 2003 to January 2005
- 1.15 TB gzipped ARC files
- 29 million captures of 7.8 million unique URLs
- 23.8 million captures are HTML files

Hadoop/HBase cluster:

- 16 nodes, dual quad-core processors, 3 × 2TB disks each

Warcbase Browser = Wayback Machine

- Implementation
  - Lightweight HBase client embedded in Jetty app
  - Reuses Wayback code for rendering

Senator Hillary Rodham Clinton

New York

ABOUT NEW YORK
ABOUT SENATOR CLINTON
CONSTITUENT SERVICES
CONTACTING MY OFFICES
COMMITTEES & LEGISLATION
ISSUES
NEWS & SPEECHES
SUPPORT OUR TROOPS
USEFUL LINKS

Privacy Policy

Information on Sending Items to the Senate

Watch a Video Welcome Message from

Dear Friend,

Thank you for visiting my on-line office! I appreciate your interest in the issues before the United States Senate.

Please let me hear from you about your views on the issues that matter to your family and your community.

Sincerely,

Hillary Rodham Clinton

Senator Clinton's Thanksgiving Trip to Afghanistan and Iraq

Senator Clinton and Senator Reed with U.S Marines outside of the U.S. Embassy in Kabul, Afghanistan

edworkforce.house.gov
Committee on Education and the Workforce
2181 Rayburn House Office Building
U.S. House of Representatives
Washington, D.C. 20515
202-225-4527
John A. Boehner, Chairman

Congress Online
Silver Mouse
Award Winner
2003

GOP Education Leaders Unveil "College Cost Central" Website to Seek Input from Parents & Students

Washington, D.C. – Providing a new resource for parents, students, and taxpayers troubled by dramatically increasing higher education prices, Republicans on the U.S. House Education & the Workforce Committee today announced the launch of the College Cost Central website. Read more.

Click here for the College Cost Central website

House GOP Education Committee Leaders Release Report on College Cost Crisis

WASHINGTON, D.C. – On September 4, 2003 the U.S. House Education & the Workforce Committee Chairman John Boehner (R-OH) and 21st Century Competitiveness Subcommittee Chairman Howard P. "Buck" McKeon (R-CA) introduced a congressional report declaring that the nation's higher education system is in crisis as a result of exploding cost increases that threaten to put college out of reach for low and middle income students and

Click Here for the Latest Information on the *No Child Left Behind Act*

Latest News . . .

Issues:

Click here to give your input on the Reauthorization of IDEA

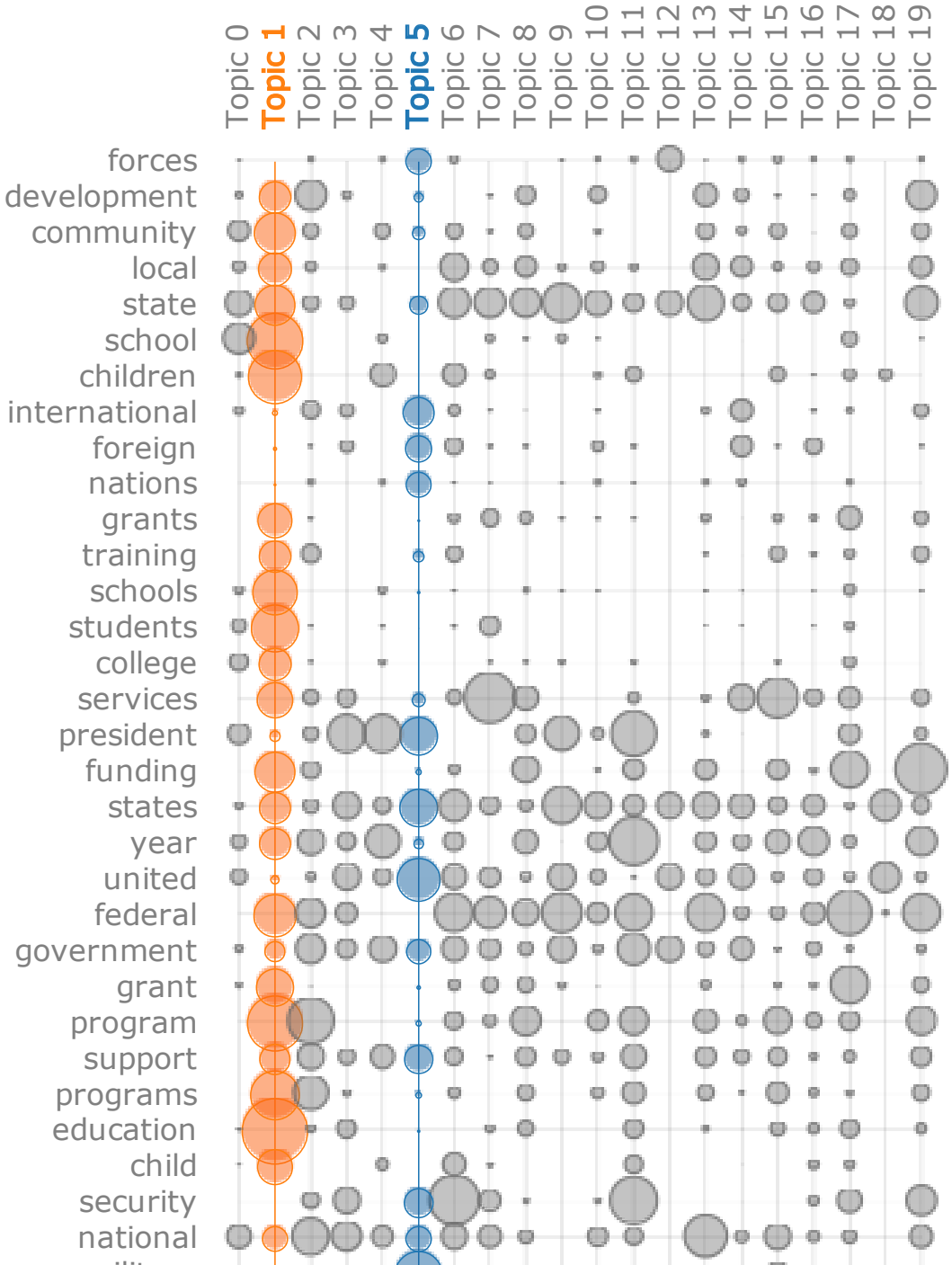
Press:

Statement by Education & the Workforce Committee Chairman John Boehner (R-OH) on 30-Year Treasury Rate Pension Fix - 12/10/03

House Education Committee Members Praise Education Department for New No Child Left Behind Rule Providing Flexibility for Local Schools on Testing Children with Disabilities - 12/9/03

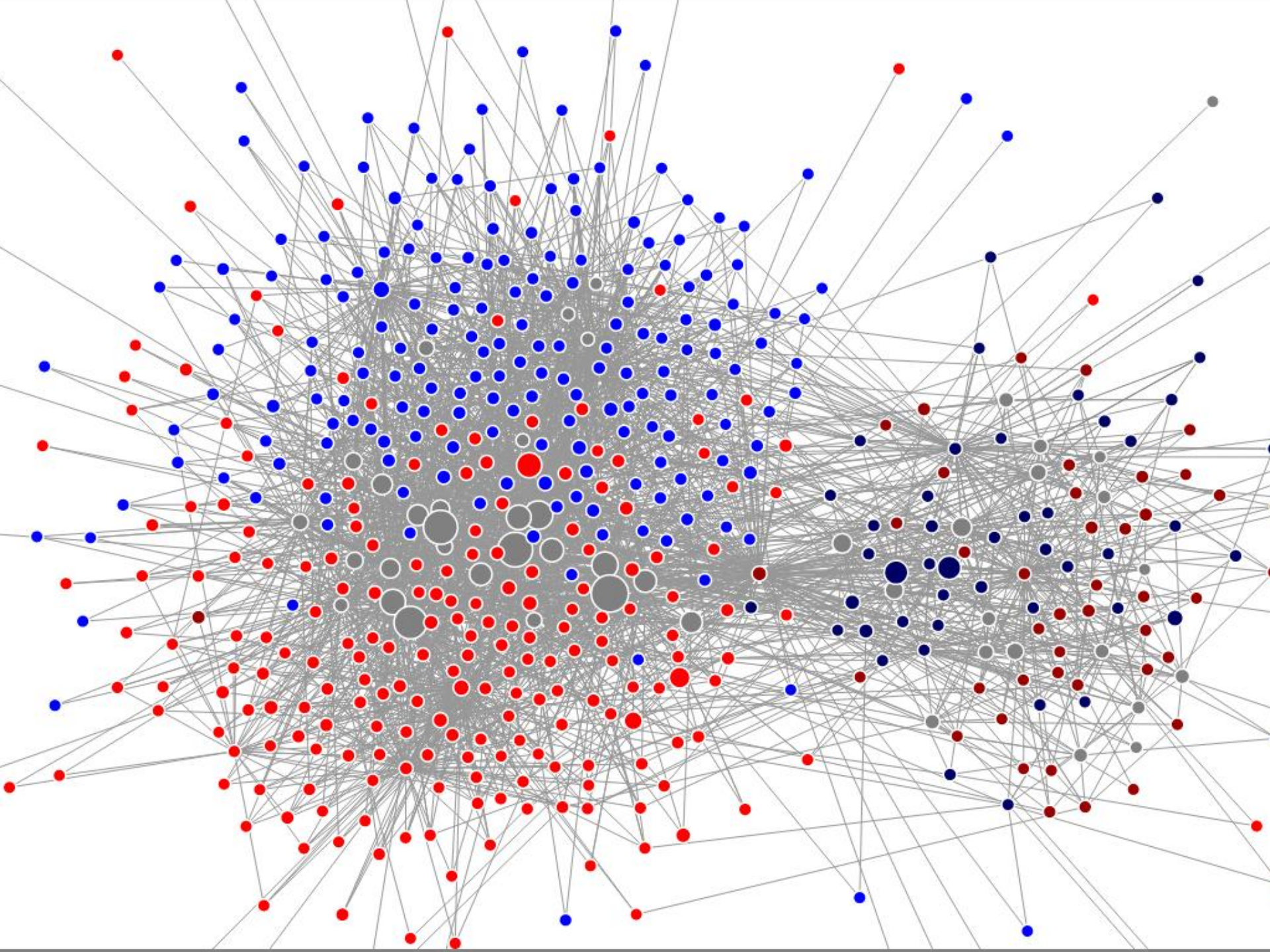
Topic Model Explorer

- Implementation
  - Topic modeling (LDA) on each temporal slice
  - Adaptation of Termite visualization


Topic 0
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8
Topic 9
Topic 10
Topic 11
Topic 12
Topic 13
Topic 14
Topic 15
Topic 16
Topic 17
Topic 18
Topic 19
forces
development
community
local
state
school
children
international
foreign
nations
grants
training
schools
students
college
services
president
funding
states
year
united
federal
government
grant
program
support
programs
education
child
security
national

Webgraph Explorer

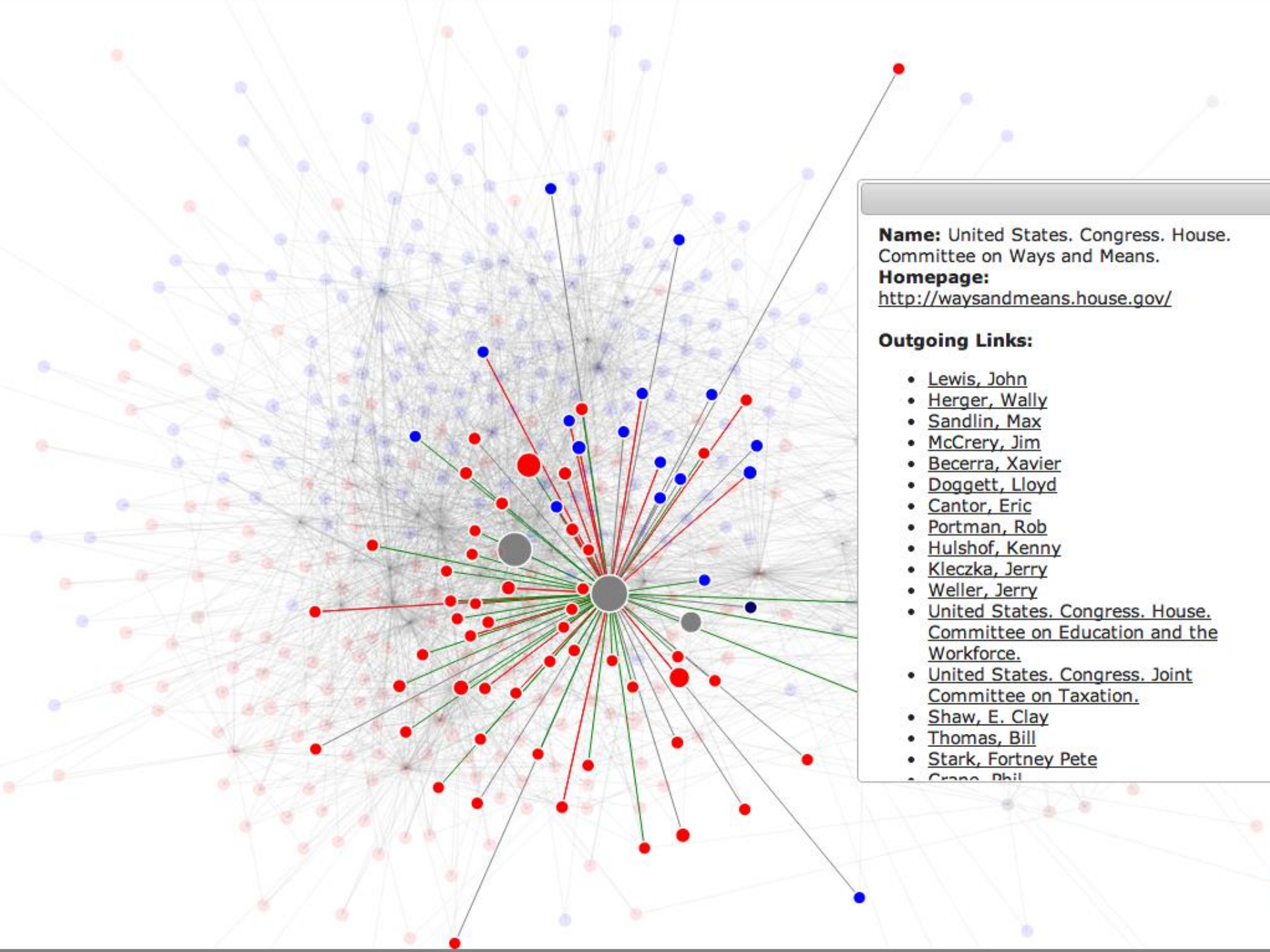
- Implementation
  - Link extraction with Hadoop, site-level aggregation
  - Computation of standard graph statistics
  - d3 interactive visualization



Name: Craig, Larry
Party: Republican
State: Idaho
Homepage: http://craig.senate.gov/

Outgoing Links:

- Simpson, Mike
- Cramer, Bud
- Hunter, Duncan
- Ensign, John
- Cunningham, Randy "Duke"
- Manzullo, Don
- Levin, Carl
- Rogers, Mike
- Walden, Greg
- Lincoln, Blanche
- Hatch, Orrin G.
- Landrieu, Mary L.
- Lowey, Nita M.
- Davis, Jo Ann



Name: United States. Congress. House. Committee on Ways and Means.
Homepage:
http://waysandmeans.house.gov/

Outgoing Links:

- Lewis, John
- Herger, Wally
- Sandlin, Max
- McCrery, Jim
- Becerra, Xavier
- Doggett, Lloyd
- Cantor, Eric
- Portman, Rob
- Hulshof, Kenny
- Kleczka, Jerry
- Weller, Jerry
- United States. Congress. House. Committee on Education and the Workforce.
- United States. Congress. Joint Committee on Taxation.
- Shaw, E. Clay
- Thomas, Bill
- Stark, Fortney Pete
- Crane, Phil

We need *deep* collaborations between:

- Users (e.g., archivists, journalists, historians, digital humanists, etc.)
- Technologists (me and my colleagues)

Warcbase is just the first step…

Source: Wikipedia (Hammer)