

Infrastructure for Supporting Exploration and Discovery in Web Archives

Jimmy Lin

University of Maryland
@lintool



Working with Internet
Archives for Research
Wednesday, June 18, 2014



From the Ivory Tower...



... to building sh*t that works

Mishne et al. Fast Data in the Era of Big Data: Twitter's Real-Time Related Query Suggestion Architecture. SIGMOD 2013.



Tweets

Struggling with complex data of Data Science 2/20 to rethi
Promoted by Cloudera

1,641 Tweets
Compose new Tweet...

Leibert et al. Automatic Management of Partitioned, Replicated Search Services. SoCC 2011

Who to follow · Refresh · View all

- plotly @plotlygraphs Follow Promoted
- Brad Anderson @boorad Followed by Florian Leibert ... Follow
- Sheila Morrissey @sheilaMorr Follow

Popular accounts · Find friends

Clinton Paquin @clintonpaquin Simply stated, "The only problem with muscle memory" @TheChan View conversation

The Hill @thehill · 1h Republicans take debt ceiling View summary

Retweeted by Alex Feinberg Popehat @Popehat · 10h In a world in which few things feed does. Expand

sochi

- #Sochi2014
- #SochiProblems
- Sochi
- #SochiFail

- Sochi 2014 @Sochi2014
- Sochi Olympics 2014 @2014Sochi
- Игры Сочи 2014 @sochi2014_ru
- Sochi Problems @SochiProblem
- NYT Olympics @SochiNYT
- Sochi Problems @SochiProblems

Search all people for sochi

Trends · Change

- #Olymp
- Ukraine
- #Conf
- Venny
- #Premi

I worked on...

- data products to surface relevant content to users

- analytics infrastructure to support data science

Gupta et al. WTF: The Who to Follow Service at Twitter. WWW 2013
Lin and Kolcz. Large-Scale Machine Learning at Twitter. SIGMOD 2012



circa ~2010

~150 people total

~60 Hadoop nodes

~6 people use analytics stack daily

circa ~2012

~1400 people total

10s of Ks of Hadoop nodes, multiple DCs

10s of PBs total Hadoop DW capacity

~100 TB ingest daily

dozens of teams use Hadoop daily

10s of Ks of Hadoop jobs daily



And back!



Web archives are an important part of our cultural heritage...

... but they're underused

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

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...



Desiderata

HDFS

Scalable storage of archived data

Efficient random access HBase

Scalable processing and analytics Hadoop MapReduce

HBase
Scalable storage and access of derived data



(Existing work by IA, IIPC, Common Crawl, etc.) (Not much previous work here...)



Warcbase

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

<http://warcbase.org/>

APACHE HBASE

A collection of tables, each of which represents a sparse, distributed, persistent multidimensional sorted map

(Open-source implementation of Google's Bigtable, the system behind Gmail, Google Maps, etc.)



Warcbase Application Lifecycle

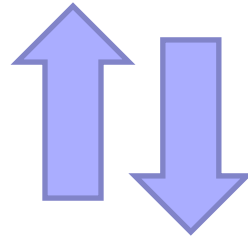


WARC/ARC

Ingestion

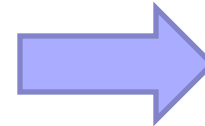


Processing & Analytics



A P A C H E
HBASE

Applications
and Services



Processing and Analytics

text analysis, webgraph analysis, ...



Derived data stored alongside raw source!



Scalability?

We got 99 problems but scalability ain't one...

– Jay-Z

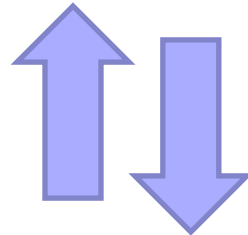
Scalability of Warcbase limited by Hadoop/HBase
Applications are lightweight clients

Demo Applications



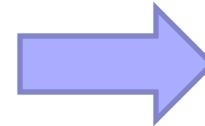
WARC/ARC

Ingestion



Processing & Analytics

A P A C H E
HBASE



Applications
and Services





LIBRARY OF
CONGRESS

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

Key features

- Offloads storage management to HBase
- Transparently scales out with HBase

A P A C H E
HBASE





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
Senator Clinton

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



[Home](#)

[What's New!](#)

[Schedule](#)

[Hearings](#)

[Markups](#)

[Press](#)

[Issues](#)

[Legislation](#)

Search our site:

Search

[Webcasting](#)



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](#)

About the Committee

[Chairman's Welcome](#)

[Contact the Committee](#)

[Committee History](#)

[Internships & Fellowships](#)

[Members & Jurisdiction](#)

[Publications](#)

[Links to Additional Resources](#)

[Democrat Views](#)

[SITE INDEX](#)

Congress Online

Silver
Mouse
Award Winner

2003

Topic Model Explorer

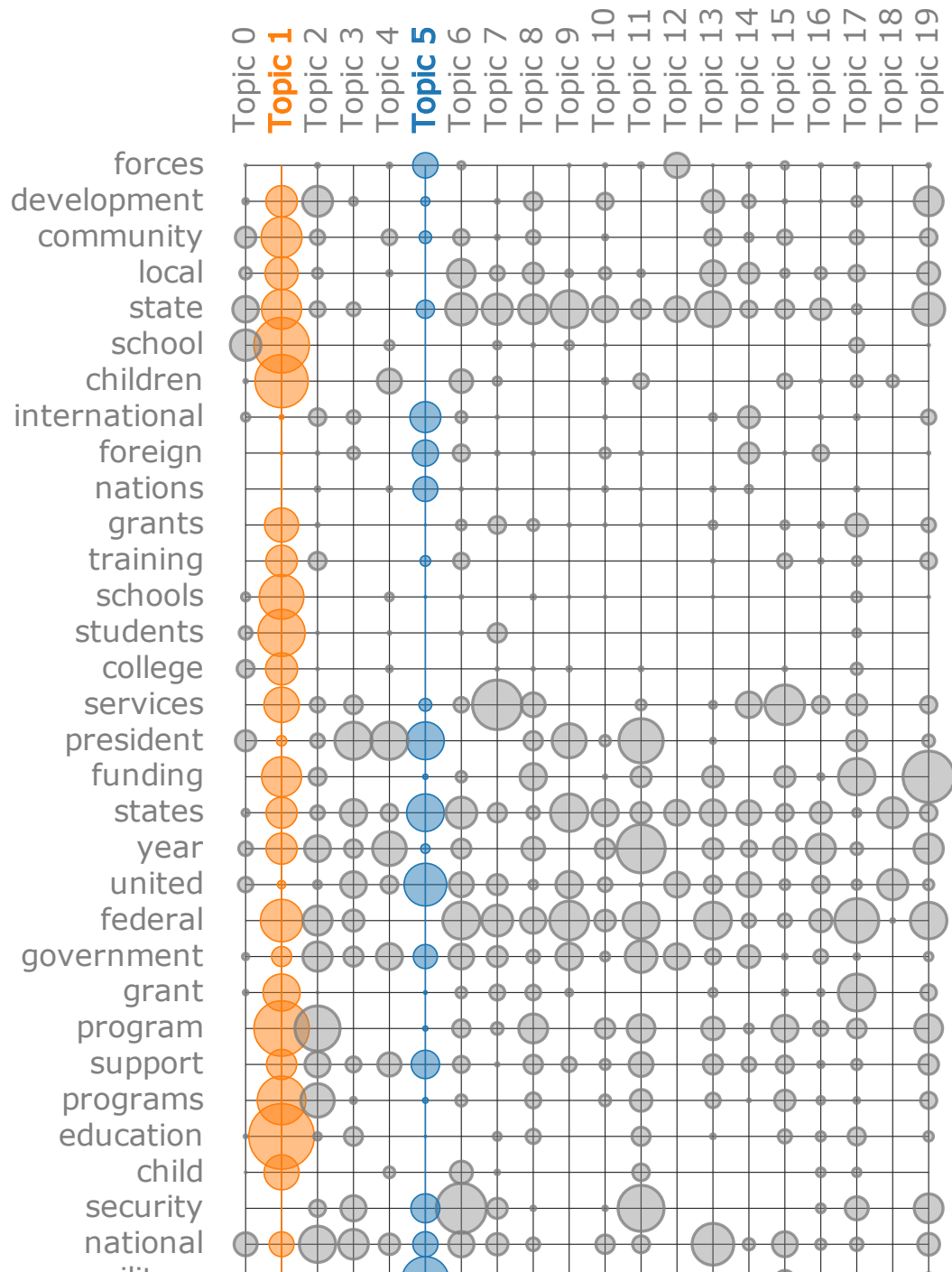
Implementation

LDA on each temporal slice

Adaptation of Termite visualization

A P A C H E
HBASE





Webgraph Explorer

Implementation

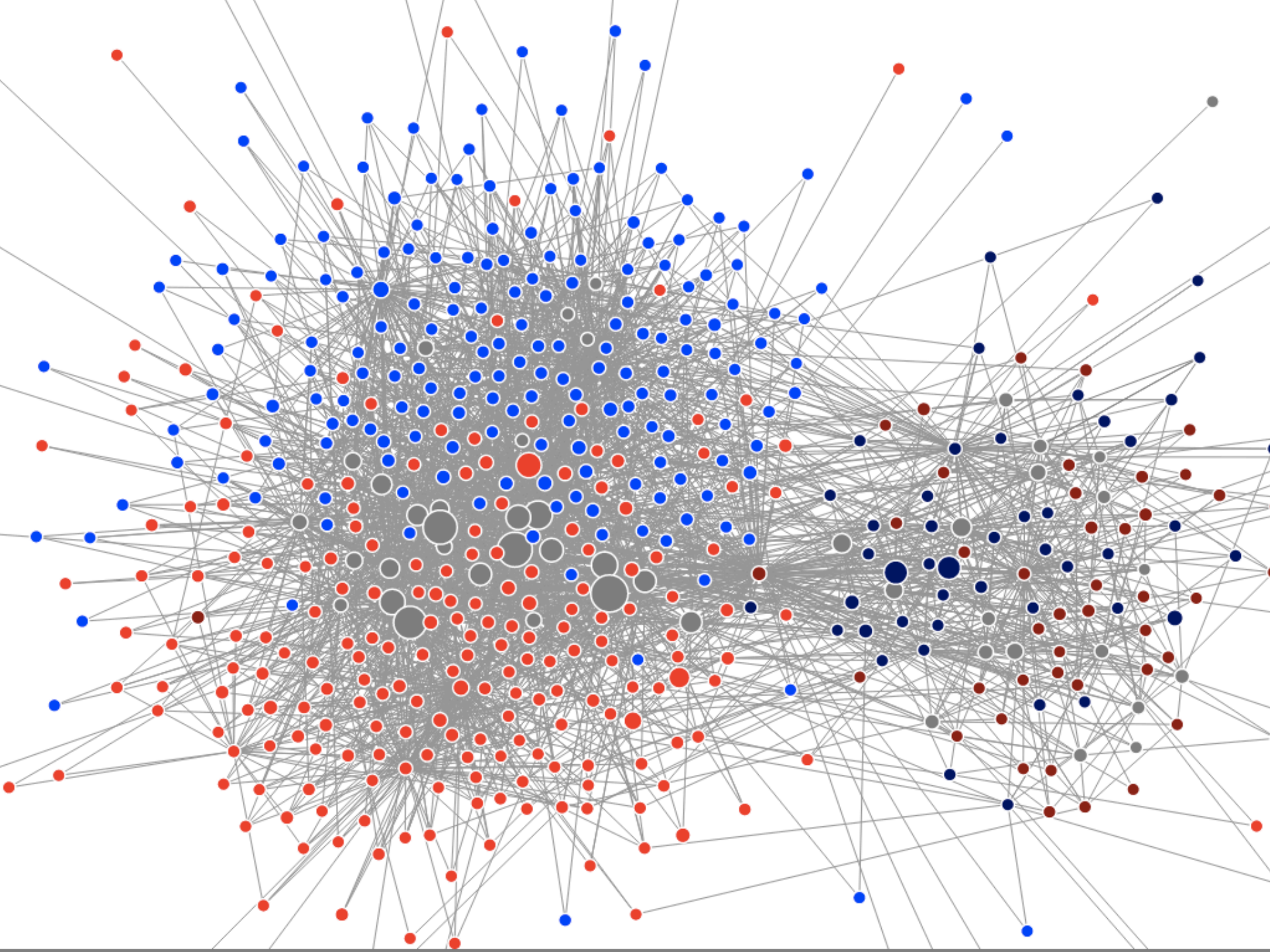
Link extraction with Hadoop, site-level aggregation

Computation of standard graph statistics

d3 interactive visualization

A P A C H E
HBASE







Name: Craig, Larry

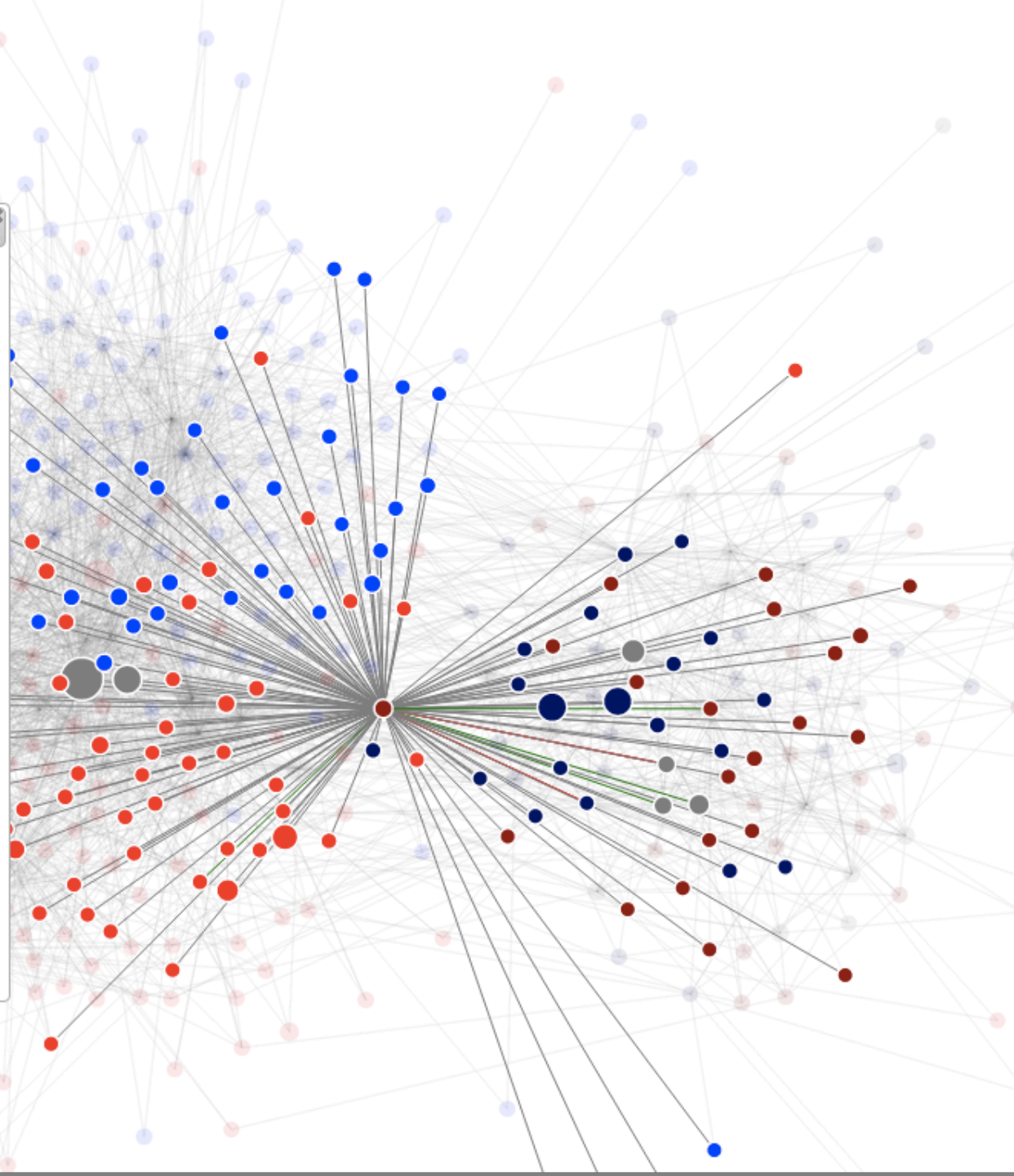
Party: Republican

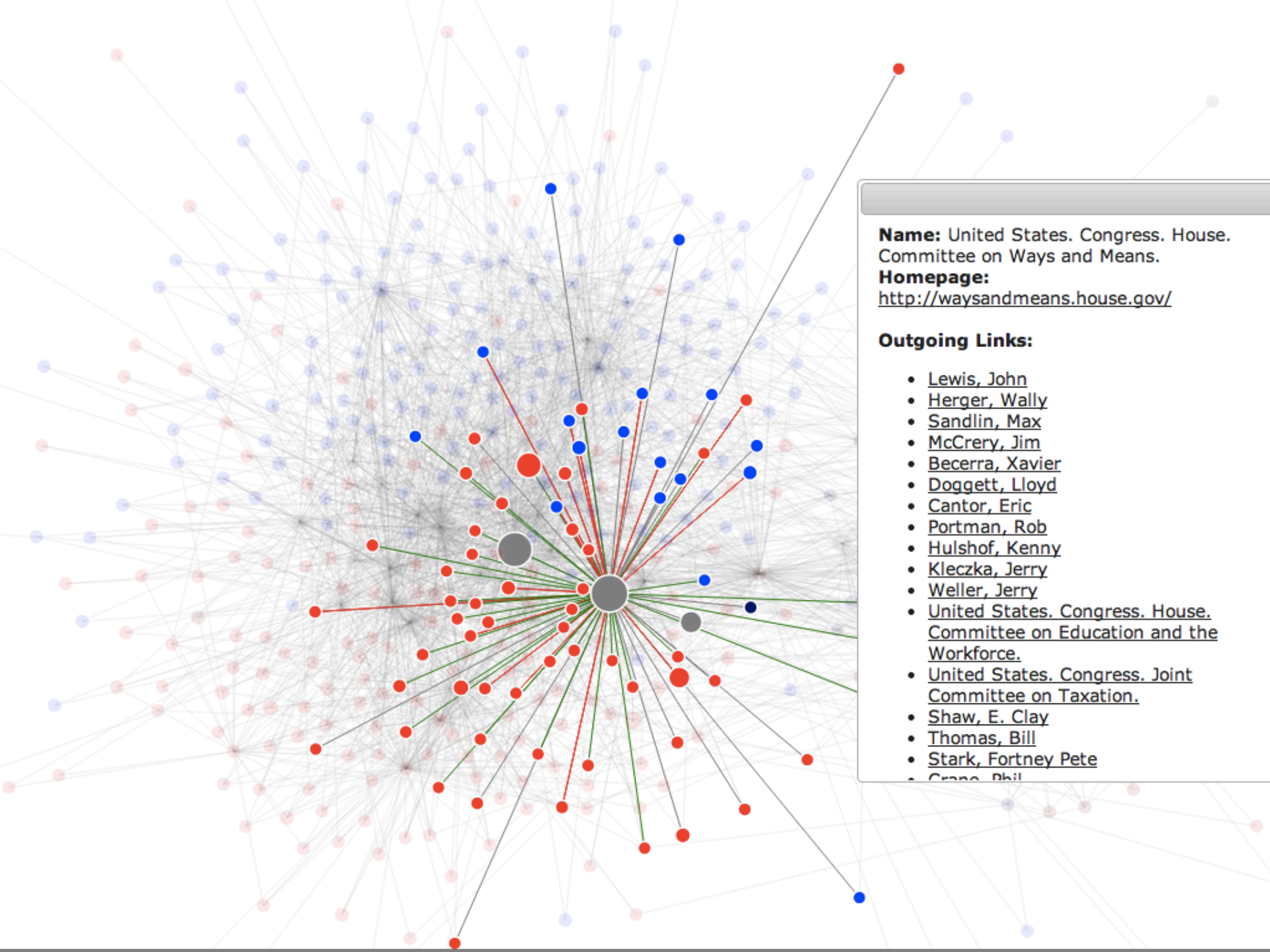
State: Idaho

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](#)
- [Crapo, Phil](#)

Goal: tools to support exploration and discovery

We need *deep* collaborations between:

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

Technologists (me and my colleagues)

Warcbase is just the first step...





Questions?