Warcbase

Building a Scalable Platform on HBase and Hadoop

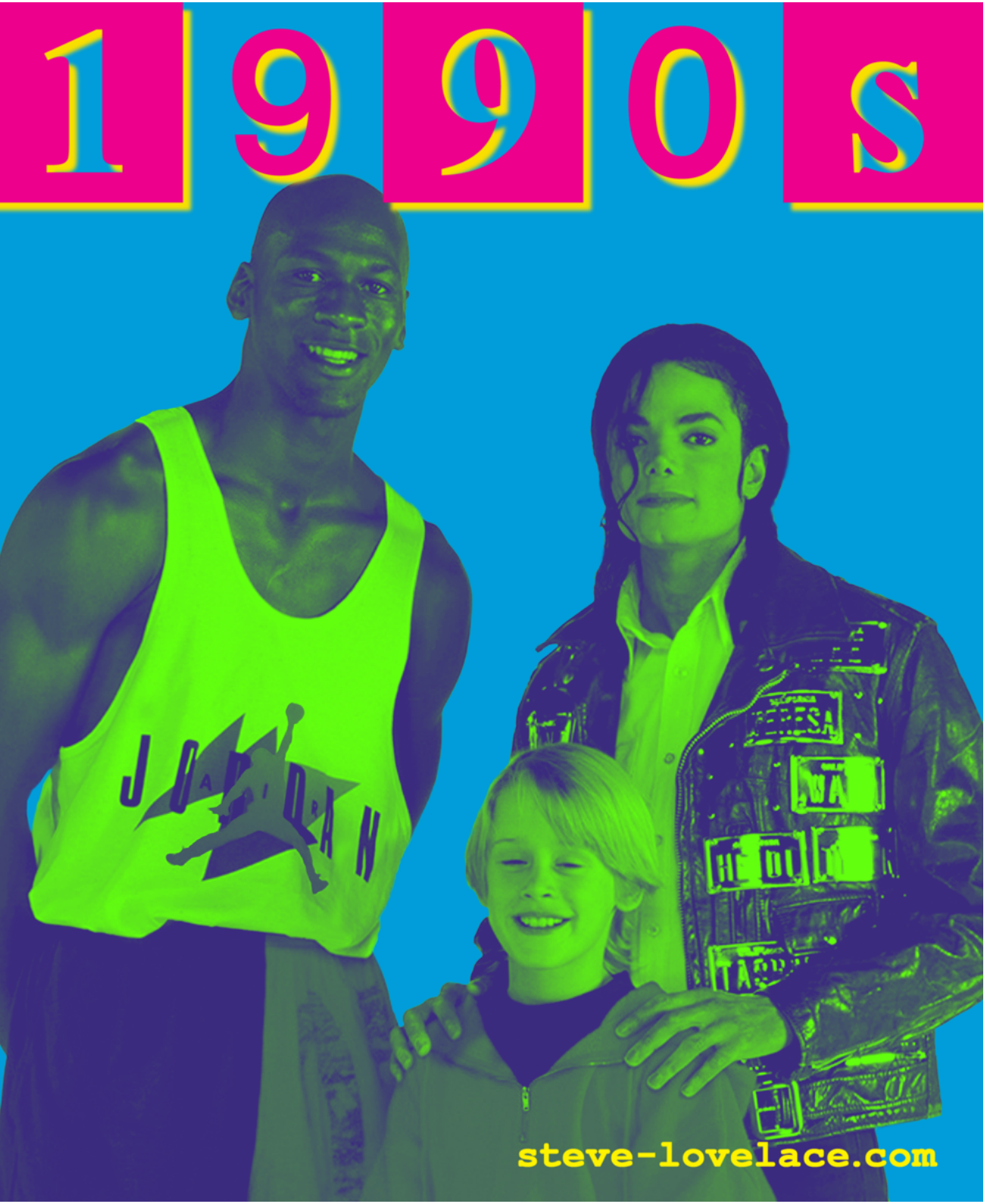
Part Two: Historian Use Case

Jimmy Lin
University of Maryland
College Park, MD

Ian Milligan
University of Waterloo
Waterloo, ON Canada

Why should a historian care?

The sheer amount of social, cultural, and political information generated every day presents new opportunities for historians.



Could one even study the 1990s and beyond **without web archives?**

No.

Historians need to do this now, or we're going to be left behind.

Nightmare Scenario

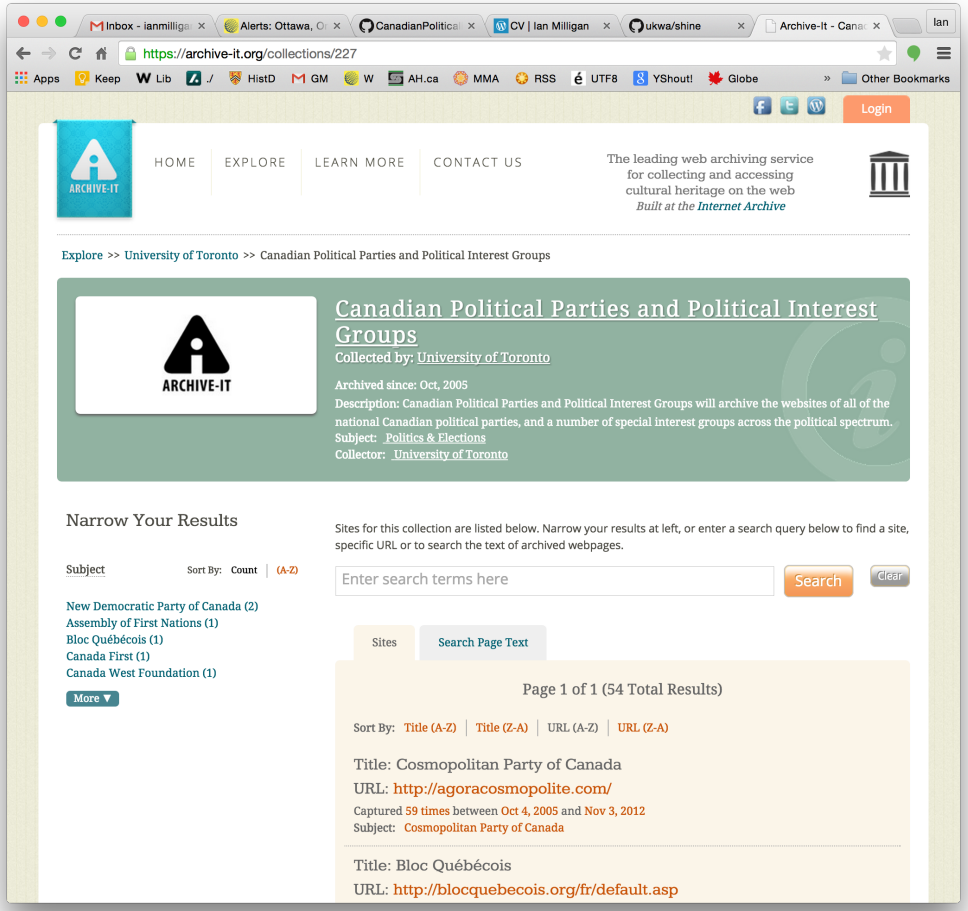
- Wayback Machine won't be enough. We won't use that.
- Historians rely uncritically on **date-ordered keyword search results**, putting them at mercy of search algorithms they do not understand;
- Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/*Nature* magazine article);
- Our profession gets left behind…

Unlocking an Archive-It Collection

- Archive-It has **amazing collections** of social, cultural, political, and economic records generated by **everyday people, leaders, businesses, academics, and beyond.**
- Stories waiting to be hold.
- The data is there, but the problem is **access.**

Example Dataset

- Archive-It Collection 227, **Canadian Political Parties and Political Interest Groups (University of Toronto)**
- October 2005 - Present
  - All major and minor political parties, as well as organized political interest groups (Council of Canadians, Coalition to Oppose the Arms Trade Assembly of First Nations, etc.)
- Started by now-retired librarian, hard to get details on seed list



Two Main Approaches

- Warcbase
  - Link extraction and analytics
  - Full-text extraction and analytics
- Full-text faceted search
  - UK Web Archive's **Shine** solr front end

Using Warcbase to analyze links and full-text

Basic Link Statistics

- Count number of pages per domain
- Count number of links for each crawl so they can be normalized (very important)
- Run on command line using relatively simple pig scripts

Example Script (counting number of links for each crawl)

```
register 'target/warcbase-0.1.0-SNAPSHOT-fatjar.jar';

DEFINE ArcLoader org.warcbase.pig.ArcLoader();
DEFINE ExtractLinks
org.warcbase.pig.piggybank.ExtractLinks();

raw = load '/shared/collections/CanadianPoliticalParties/
arc/' using ArcLoader as
  (url: chararray, date: chararray, mime: chararray,
content: bytearray);

a = filter raw by mime == 'text/html' and date is not null;
b = foreach a generate SUBSTRING(date, 0, 6) as date, url,
FLATTEN(ExtractLinks((chararray) content, url));
c = group b by $0;
d = foreach c generate group, COUNT(b);
```

Social Media Appearances - Twitter

(20080611220246,http://creativecommons.org/,twitter)
(20080711224545,http://www.pm.gc.ca/eng/feature.asp?pageId=105,twitter)
(20080712030632,http://www.pm.gc.ca/fra/feature.asp?pageId=105,twitter)
(20080712142357,http://www.pm.gc.ca/eng/media.asp?category=2&;id=1814,twitter)
(20080930221618,http://www.ndp.ca/home,twitter)
(20080930221618,http://www.ndp.ca/home,twitter)
(20080930221638,http://www.liberal.ca/default_e.aspx,twitter)
(20080930221641,http://www.liberal.ca/story_15081_e.aspx,twitter)
(20080930221714,http://www.liberal.ca/video_e.aspx,twitter)
(20080930221903,http://www.ndp.ca/page/5246,twitter)
(20080930221904,http://www.ndp.ca/twitterblogwidget/ndp-twitter.php?
lang=en,twitter)
(20080930222049,http://greenparty.ca/en/action,twitter)
(20080930222124,http://www.ndp.ca/bloggingtools,twitter)
(20080930222825,http://greenparty.ca/en/campaign/35053,twitter)
(20080930223014,http://greenparty.ca/en/campaign/35068,twitter)
(20080930223240,http://www.liberal.ca/depth_e.aspx,twitter)
(20080930223258,http://www.liberal.ca/enews_e.aspx,twitter)
(20080930223315,http://www.liberal.ca/glance_e.aspx,twitter)
(20080930223320,http://www.liberal.ca/story_15073_e.aspx,twitter)
(20080930223323,http://www.liberal.ca/gallery_e.aspx,twitter)

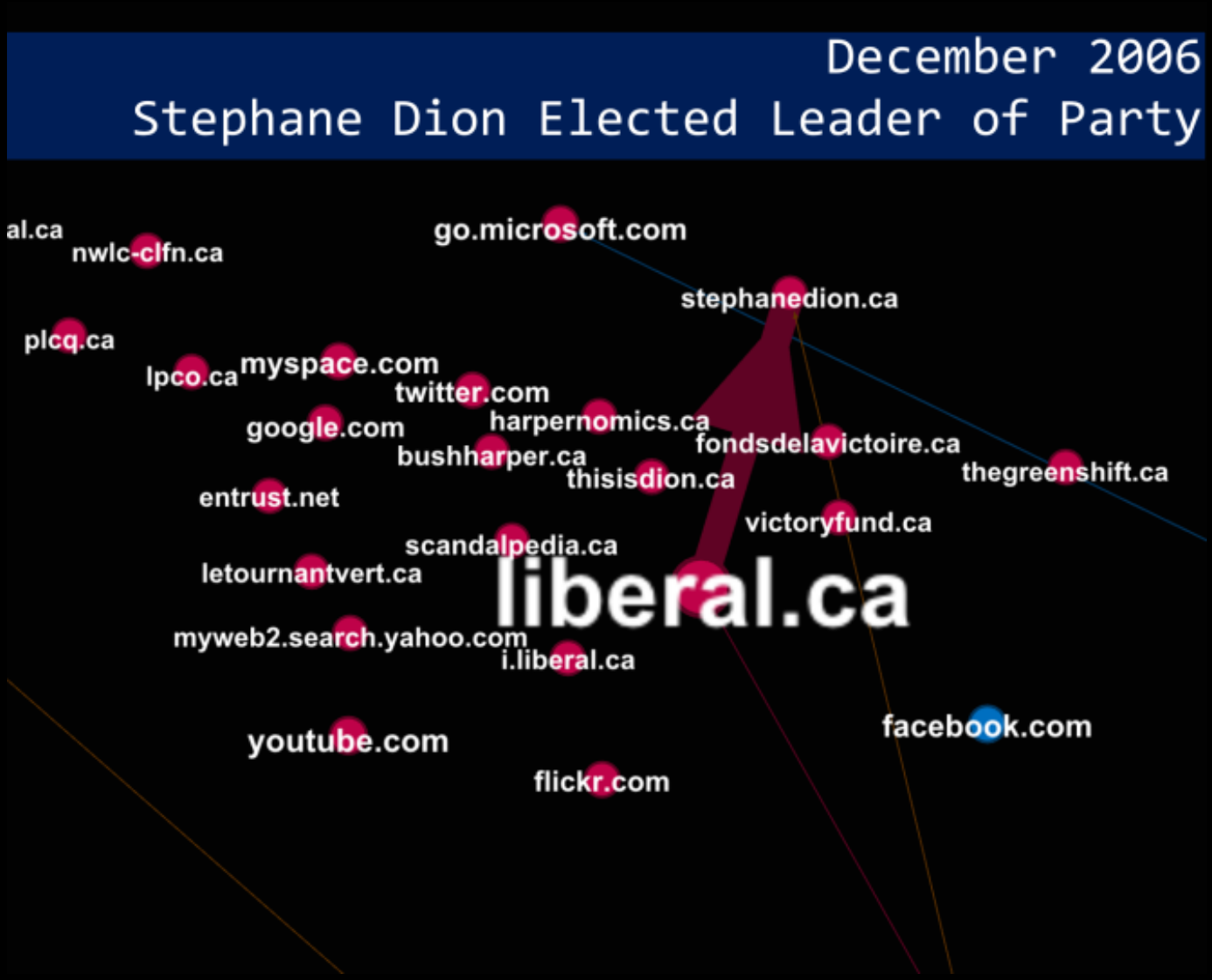
Social Media Appearances - Facebook

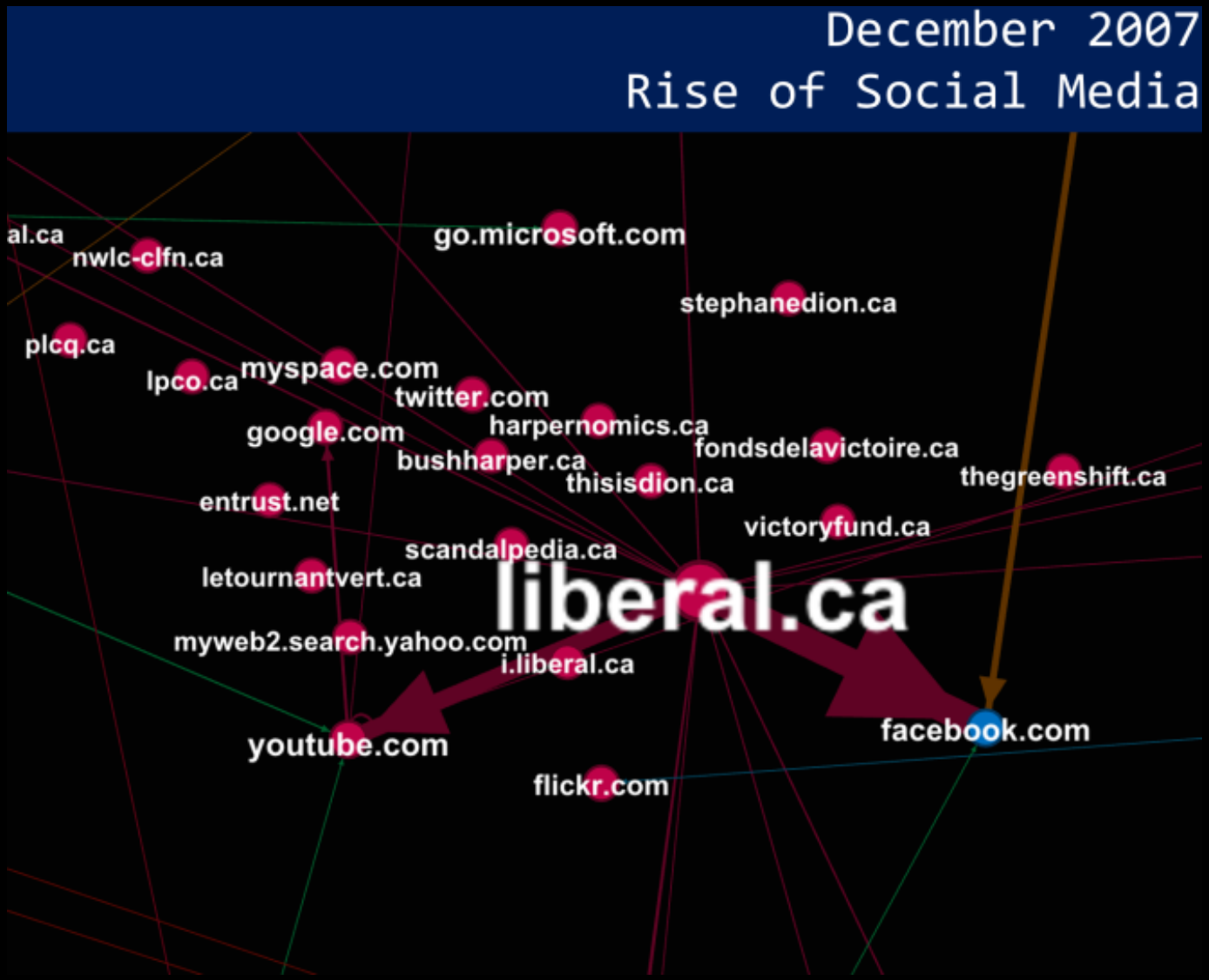
(20070418135140,http://www.liberal.ca/glance_e.aspx,facebook)
(20070418135947,http://greenparty.ca/en/blog/activemenu/menu?page=2,facebook)
(20070418140056,http://greenparty.ca/en/blog/activemenu/book?page=2,facebook)
(20070418140511,http://greenparty.ca/en/blog/popular?page=3,facebook)
(20070418140516,http://www.liberal.ca/glance_f.aspx,facebook)
(20070418141139,http://greenparty.ca/en/blog/431,facebook)
(20070418141930,http://greenparty.ca/en/blog?page=2,facebook)
(20070418143749,http://greenparty.ca/en/node/1280,facebook)
(20070418143900,http://greenparty.ca/en/blog/activemenu/activemenu/book?page=2,facebook)
(20070418144002,http://greenparty.ca/en/blog/activemenu/activemenu/menu?page=2,facebook)
(20070418151727,http://www.equalvoice.ca/youth/,facebook)
(20070418151734,http://www.equalvoice.ca/youth/index.htm,facebook)
(20070418151843,http://www.equalvoice.ca/youth/Bios.htm,facebook)
(20070418153832,http://greenparty.ca/fr/node/1280,facebook)
(20070418154008,http://greenparty.ca/en/blog/activemenu/activemenu/activemenu/menu?page=2,facebook)
(20070418154112,http://greenparty.ca/en/blog/activemenu/activemenu/activemenu/book?page=2,facebook)
(20070518134656,http://www.liberal.ca/glance_e.aspx,facebook)
(20070518134918,http://www.liberal.ca/conversation_e.aspx,facebook)
(20070518134918,http://www.liberal.ca/conversation_e.aspx,facebook)
(20070518134941,http://www.ndp.ca/page/4733,facebook)

Link Analysis

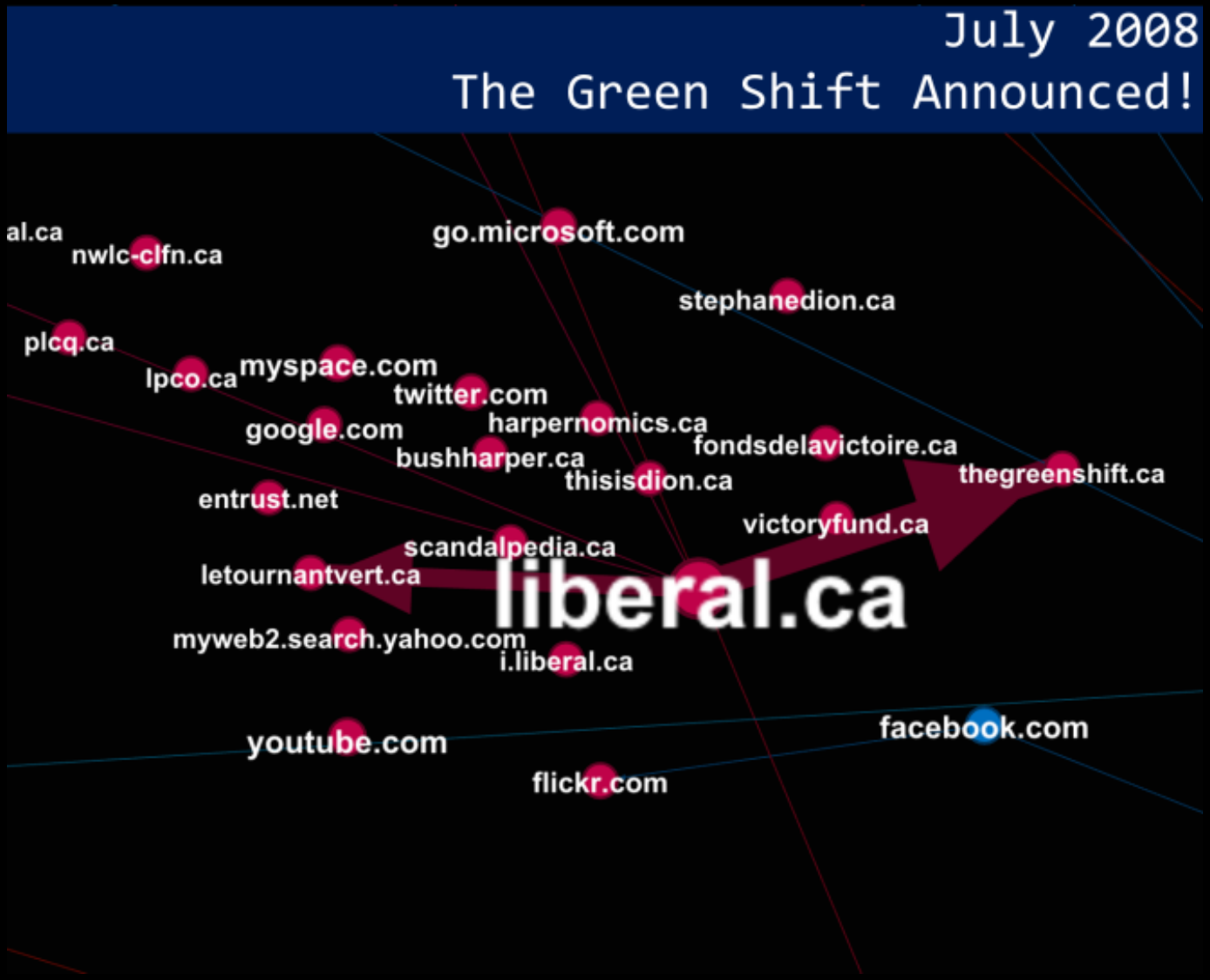
- Extracting links by domain (tab-separated values):

```
200810 conservative.ca   digg.com   2325
200810 conservative.ca   facebook.com   2325
200810 conservative.ca   mycampaign.conservative.ca   7902

[..]

200902 liberal.ca ctv.ca 16
200902 liberal.ca del.icio.us   1118
200902 liberal.ca digg.com   1118
```


December 2006
Stephane Dion Elected Leader of Party
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com


December 2007
Rise of Social Media
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com


April 2008
Fundraising with the Victory Fund/
Fonds de la Victoire
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com

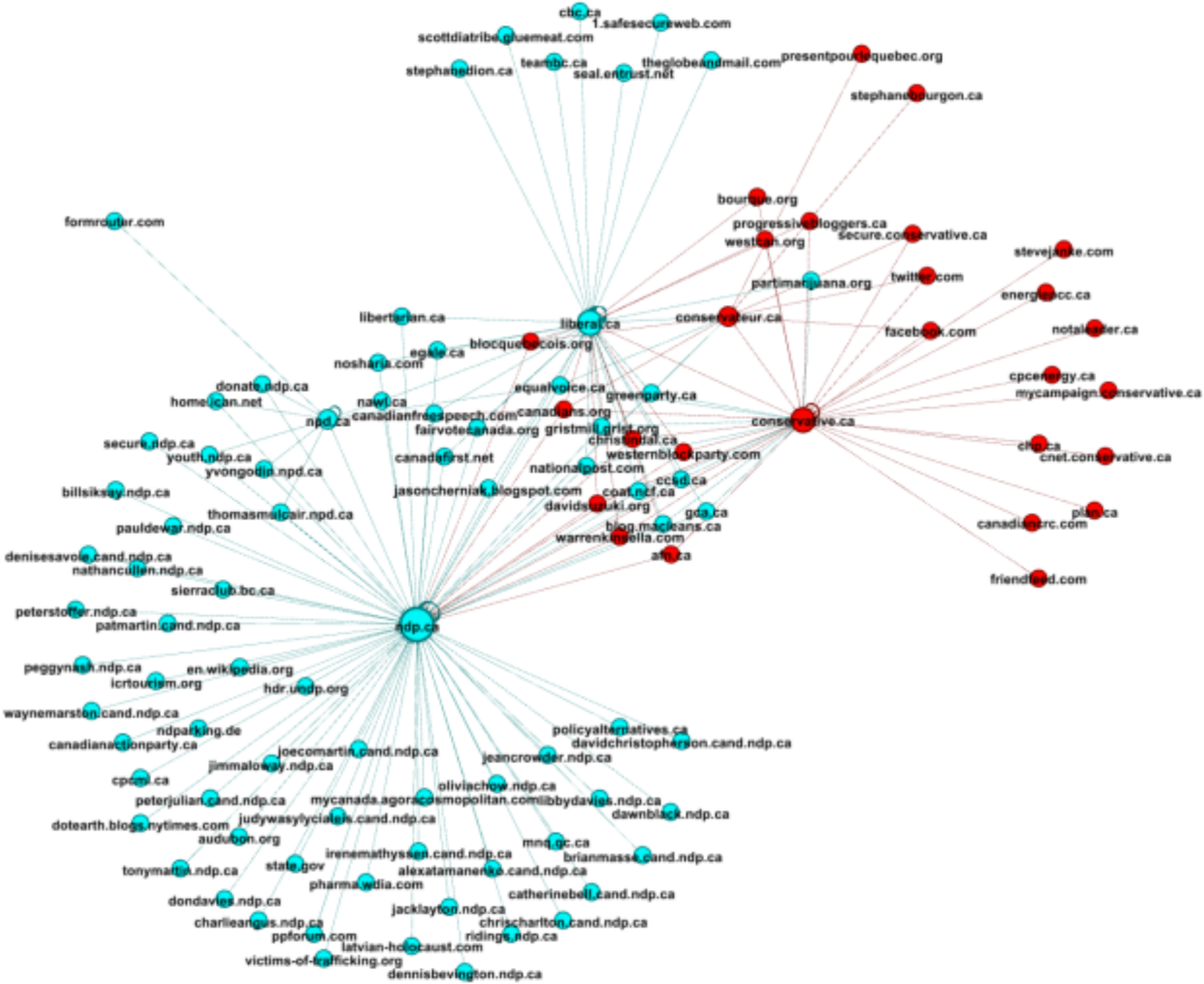

July 2008
The Green Shift Announced!
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com


October 2008
Election Campaign - Advertisement Sites
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com

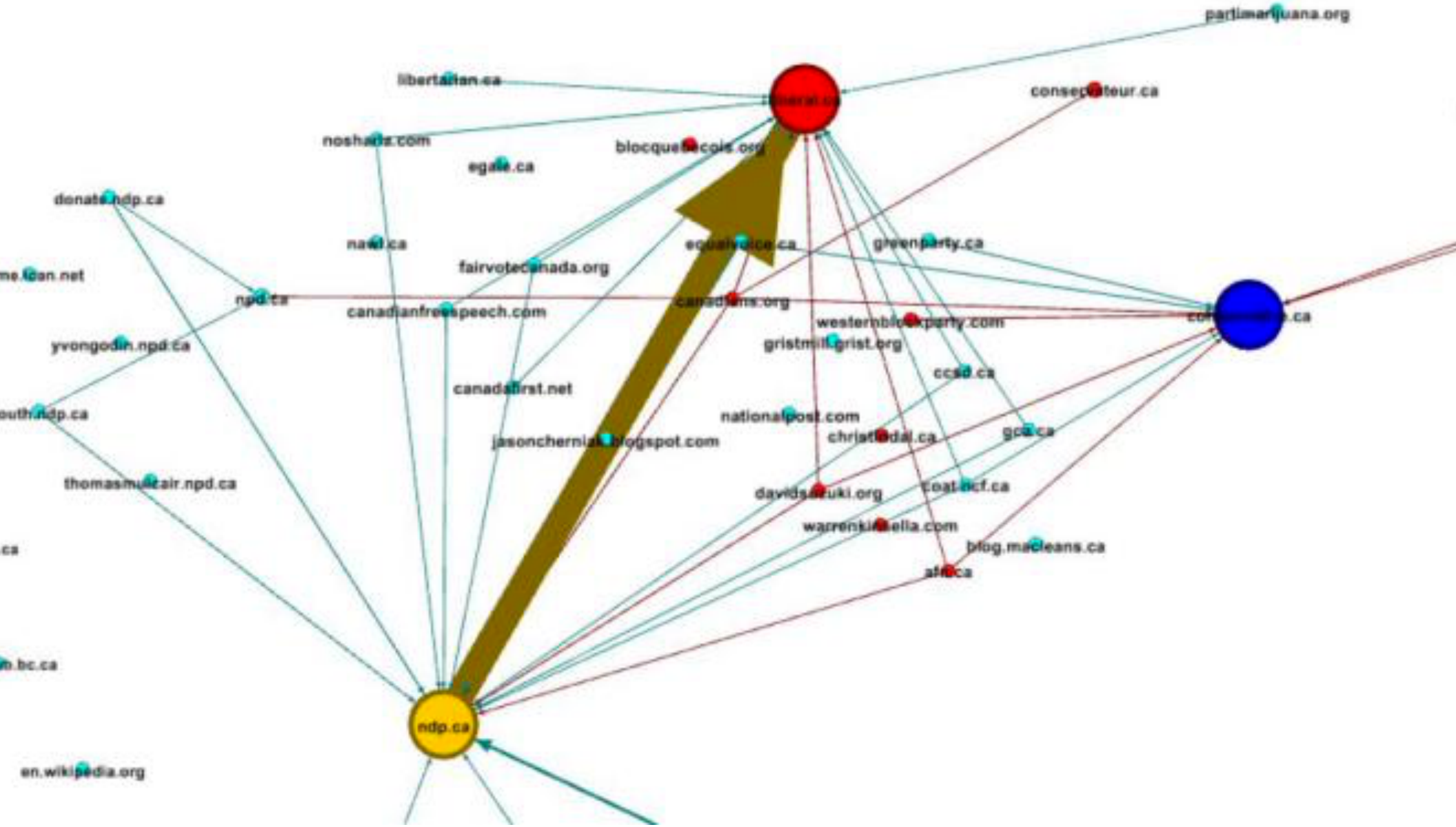

December 2008
Election campaign Ends; Attacking Harper
on Anti-American Grounds (bushharper)
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com

Other Cases

- Extracting all links to the mainstream media, or thinktanks, or other political parties


cbc.ca
1.safesecureweb.com
scottdiatribe.gluemeat.com
stephanedion.ca
teambc.ca
seal.entrust.net
theglobeandmail.com
presentpourlequebec.org
stephanebourgon.ca
bourque.org
progressivebloggers.ca
secure.conservative.ca
westcoh.org
stevejanke.com
twitter.com
partimarijuana.org
energiepcc.ca
formrouter.com
liberty.ca
liberal.ca
conservateur.ca
facebook.com
notaleader.ca
blocquebecois.org
egale.ca
nosharia.com
cpcenergy.ca
mycampaign.conservative.ca
donate.ndp.ca
equalvoice.ca
greenparty.ca
home.ican.net
nawl.ca
canadians.org
canadianfreespeech.com
npd.ca
fairvotecanada.org
gristmill.grist.org
conservative.ca
christendal.ca
secure.ndp.ca
youth.ndp.ca
canadafirst.net
westernblockparty.com
chp.ca
cnet.conservative.ca
yvongodin.npd.ca
nationalpost.com
billsiksay.ndp.ca
ccsd.ca
jasoncherniak.blogspot.com
coat.ncf.ca
davidsuzuki.org
gda.ca
thomasmulcair.npd.ca
blog.macleans.ca
pauldewar.ndp.ca
warrenkinsella.com
plan.ca
canadiancrc.com
denisesavoie.cand.ndp.ca
atn.ca
nathancullen.ndp.ca
friendfeed.com
sierraclub.bc.ca
peterstoffer.ndp.ca
patmartin.cand.ndp.ca
ndp.ca
peggynash.ndp.ca
en.wikipedia.org
icrtourism.org
hdr.undp.org
waynemarston.cand.ndp.ca
policyalternatives.ca
ndparking.de
davidchristopherson.cand.ndp.ca
canadianactionparty.ca
joecomartin.cand.ndp.ca
jeancrowder.ndp.ca
jimmaloway.ndp.ca
cpcml.ca
oliviachow.ndp.ca
peterjulian.cand.ndp.ca
mycanada.agoracosmopolitan.com
libbydavies.ndp.ca
dawnblack.ndp.ca
dotearth.blogs.nytimes.com
judywasylycia-leis.cand.ndp.ca
audubon.org
mnq.gc.ca
irenemathyssen.cand.ndp.ca
brianmasse.cand.ndp.ca
tonymartin.ndp.ca
state.gov
alexatamanenko.cand.ndp.ca
pharmawtdia.com
catherinebell.cand.ndp.ca
dondavies.ndp.ca
jacklayton.ndp.ca
charlieangus.ndp.ca
chrischarlton.cand.ndp.ca
ppforum.com
ridings.ndp.ca
latvian-holocaust.com
victims-of-trafficking.org
dennisbevington.ndp.ca

2005 Canadian Federal Election



Text Analysis

```
register 'target/warcbase-0.1.0-SNAPSHOT-fatjar.jar';

DEFINE ArcLoader org.warcbase.pig.ArcLoader();
DEFINE ExtractRawText org.warcbase.pig.piggybank.ExtractRawText();
DEFINE ExtractTopLevelDomain
org.warcbase.pig.piggybank.ExtractTopLevelDomain();

raw = load '/shared/collections/CanadianPoliticalParties/arc/' using
ArcLoader as
  (url: chararray, date: chararray, mime: chararray, content: bytearray);

a = filter raw by mime == 'text/html' and date is not null;
b = foreach a generate SUBSTRING(date, 0, 6) as date,
                       REPLACE(ExtractTopLevelDomain(url), '^\\s*www\\.',
'') as url, content;
c = filter b by url == 'greenparty.ca';
d = foreach c generate date, url, ExtractRawText((chararray) content) as
text;

store d into 'cpp.text-greenparty';
```

Text Analysis

- Now have circumscribed corpus for specified query (i.e. liberal.ca, or ndp.ca, or conservative.ca)
- Can now use standard text analysis tools, etc. to extract meaning
  - LDA (topic modeling)
  - NER (named entity recognition)

NER

October 2005

62476 Stephen Harper
30234 Michael Chong
30109 Gwynne Dyer
28011 ami Entrez
26238 Paul Martin
22303 Harper

NER

November 2008

3188 Stéphane Dion

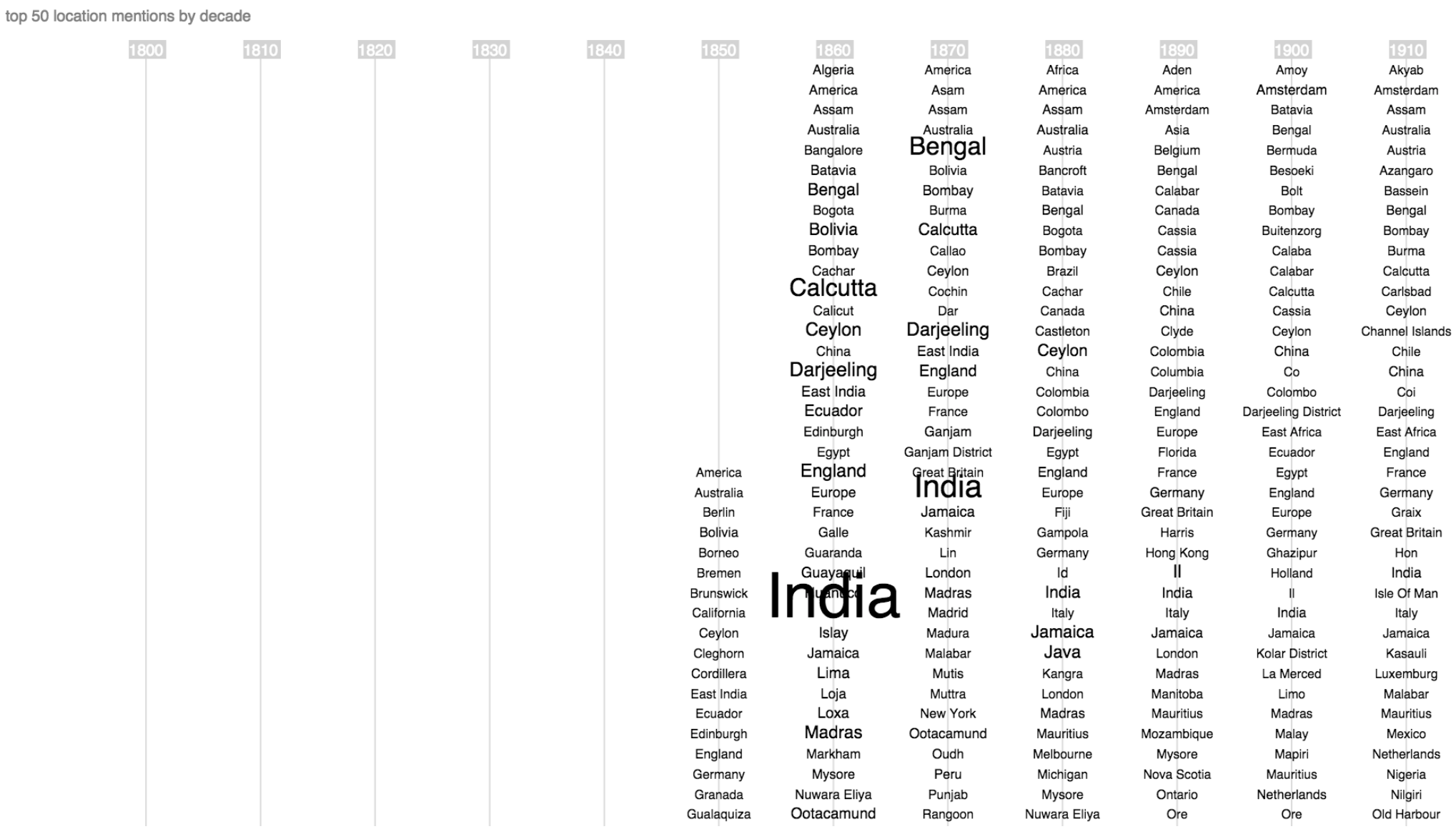
2557 Stephen Harper

2471 Stephen HarperLaureen

2410 Dion

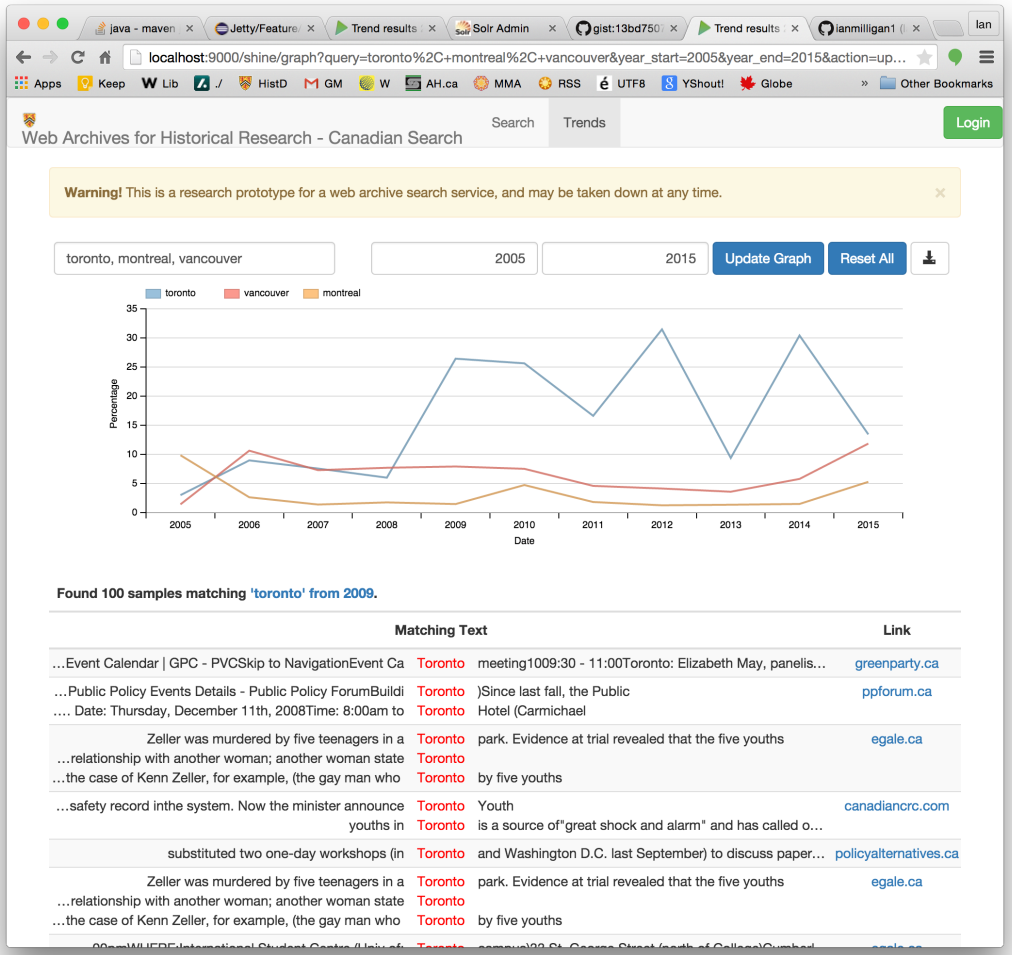
2356 Harper

Visualizing Interface Next Step?



Shine

- UK Web Archive's Shine (https://github.com/ukwa/shine)
- Indexing as bottleneck
  - ~ 250GB of WARCs takes ~ 5 days on a single machine
  - Hadoop indexer available if data in HFDS
- ~ 90GB index size



Examples

Shine

- **Advantages**: accessible to the general public, easy to use, interactive trend diagram allows digging down for context, can move down to level of document itself.

- **Disadvantage**: keyword searching requires you know what to look for; random sampling misleading when tens of thousands of records; etc.

- Doesn’t take advantage of what makes web sources so powerful: hyperlinks

Building connections between Warcbase and Shine

Conclusions & Thanks

Jimmy Lin
University of Maryland
College Park, MD

Ian Milligan
University of Waterloo
Waterloo, ON Canada