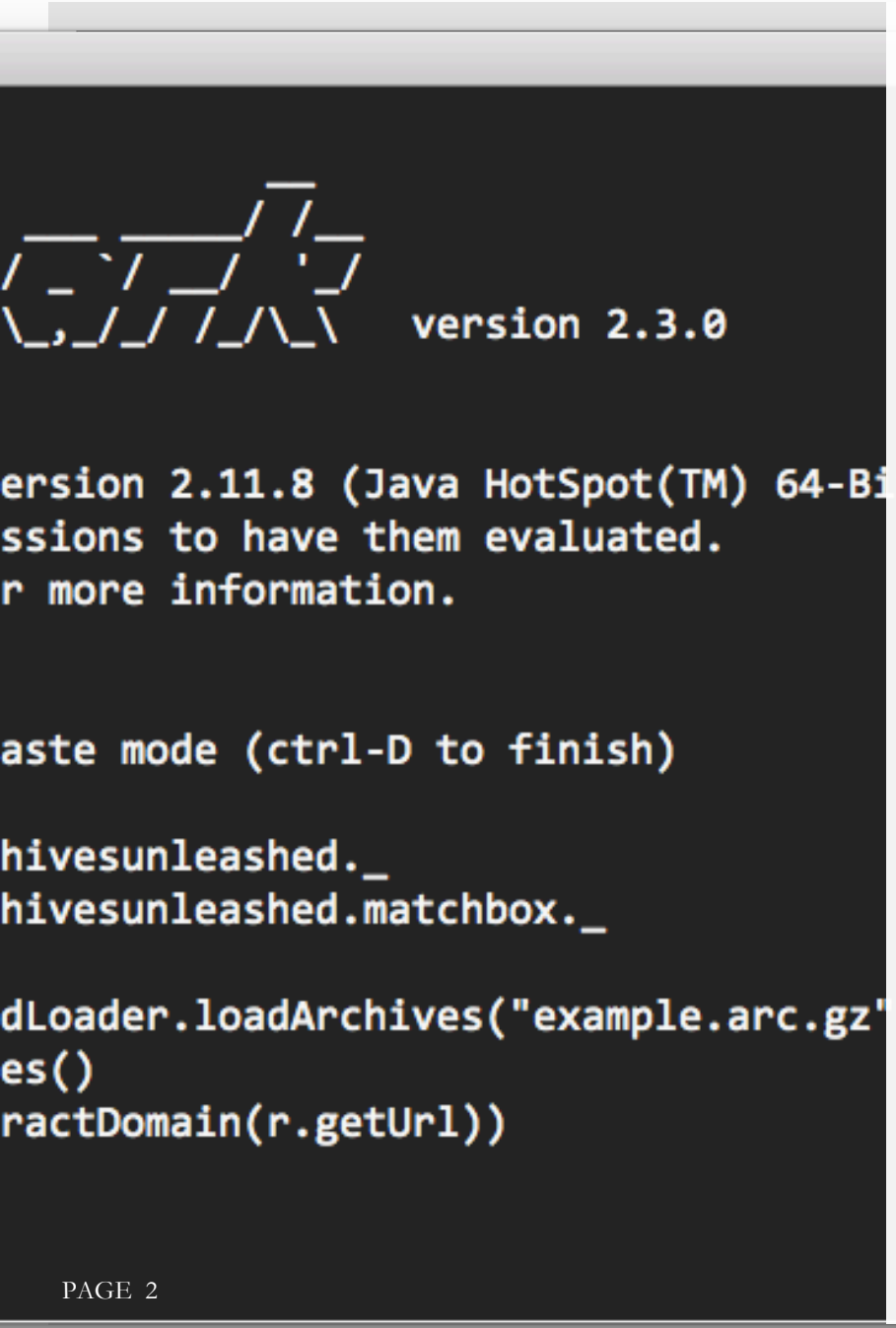
YOU'VE GOT WEB ARCHIVES.. WHAT TO DO WITH THEM?

Hacking the News Workshop



WEB ARCHIVE ANALYSIS

- The **Archives Unleashed Toolkit** was designed to tackle this problem.
- Allows a user to take WARCs and:
  - Determine elemental statistics about a collection;
  - Extract particular images, domains, URLs, pages with keywords, etc.
  - Do sophisticated Apache Spark-powered network analysis; and
  - Write custom Scala scripts to do almost anything you imagine with our set of custom web archiving User Defined Functions

LET'S SEE HOW THIS WORKS.

HACKING THE NEWS

Safari File Edit View History Bookmarks Develop Window Help

archivesunleashed.org

Welcome to the Archives Unleashed Project

Welcome to the Archives Unleashed Project

Welcome

Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Supported by a grant from the Andrew W. Mellon Foundation, we are developing web archive search and data analysis tools to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the World Wide Web.

Interested in the project? Subscribe to our newsletter! Or you can follow the links at left for information about the project, the Archives Unleashed Cloud, Archives Unleashed Toolkit, Archives Unleashed Jupyter Notebooks, or our events.

We're always looking for ways to engage archivists, librarians, researchers, developers, or any others interested in born-digital heritage!

Contact Us

Questions? Comments? Please contact us, either by leaving an issue on one of our GitHub projects or by sending us an e-mail. Are you a Slack user? Join our Slack team!

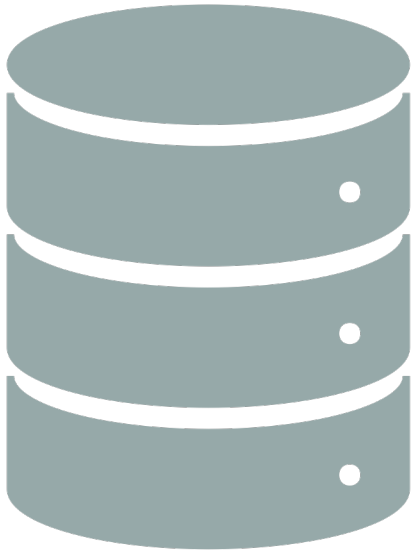
2. ssh

ssh ⌘1

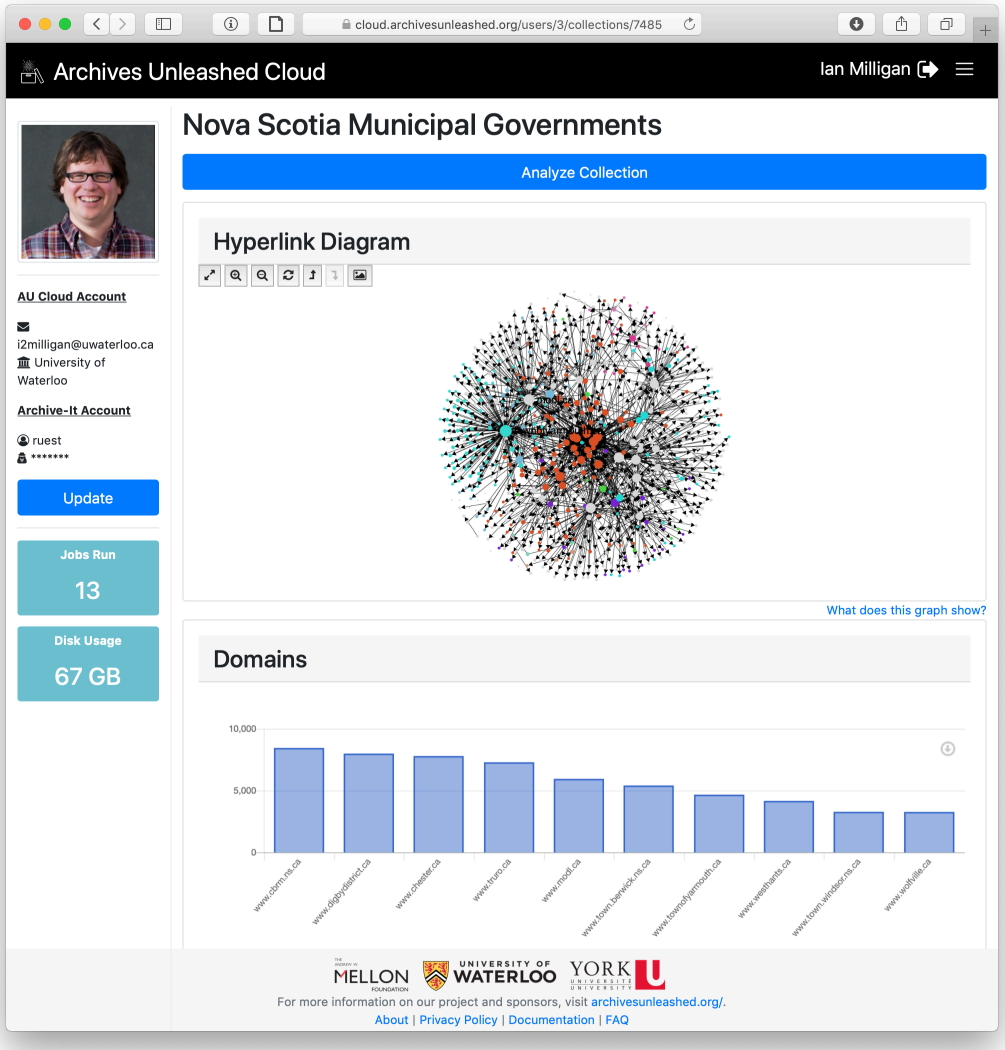
i2milligan@tuna:~$

THIS IS DIFFICULT (REQUIRES COMMAND LINE)

HACKING THE NEWS



THE DATASETS IN "HACKING-THE-NEWS.ZIP" WERE GENERATED BY THIS



AN EASIER WAY

- The **Archives Unleashed Cloud** was designed to make this easier..
- Allows a user to take WARCs and:
  - Use a modern UI to sync their collections from the provider;
  - Run basic analyses in the browser to find major sites of interest;
  - Download derivative file formats that can integrate with standard workflows.
- In other words: let's get the WARC out of the equation and *translate* it into a standard file format.

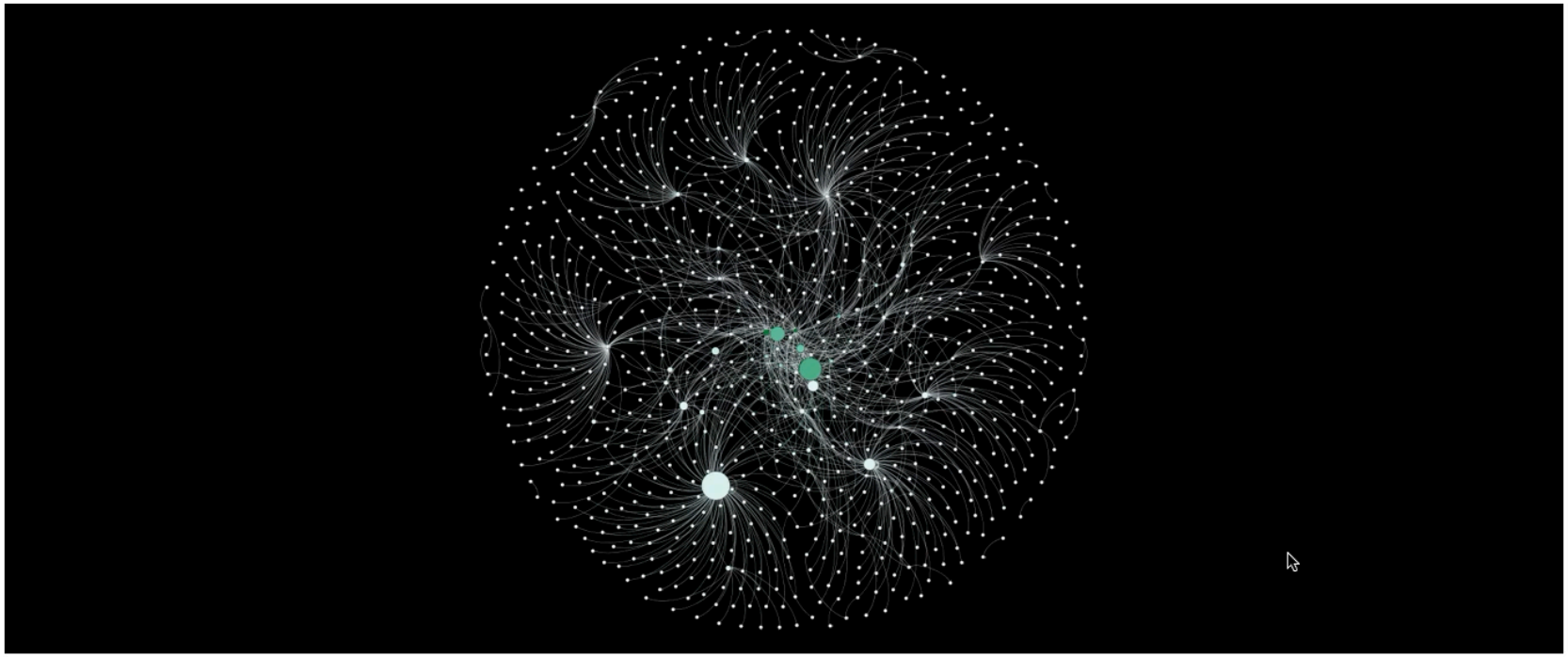
IF YOU HAVE AN ARCHIVE-IT ACCOUNT, TRY IT OUT YOURSELF AT **CLOUD.ARCHIVESUNLEASHED.ORG**

cloud.archivesunleashed.org

Archives Unleashed Cloud

Ian Milligan

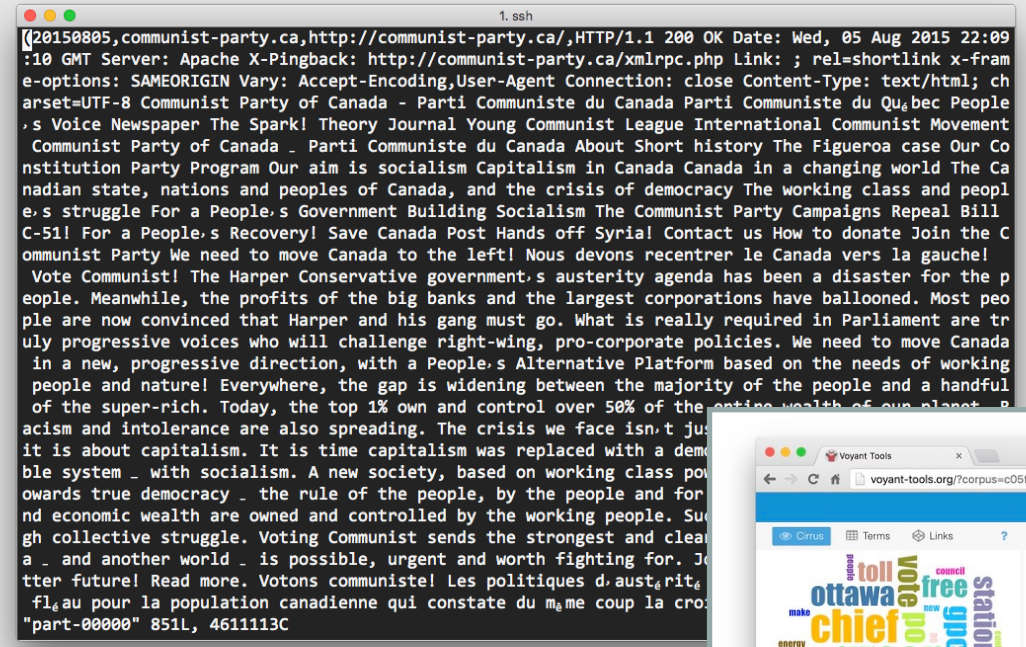
New user? You currently need Archive-It credentials to use our service. Please read our FAQ.

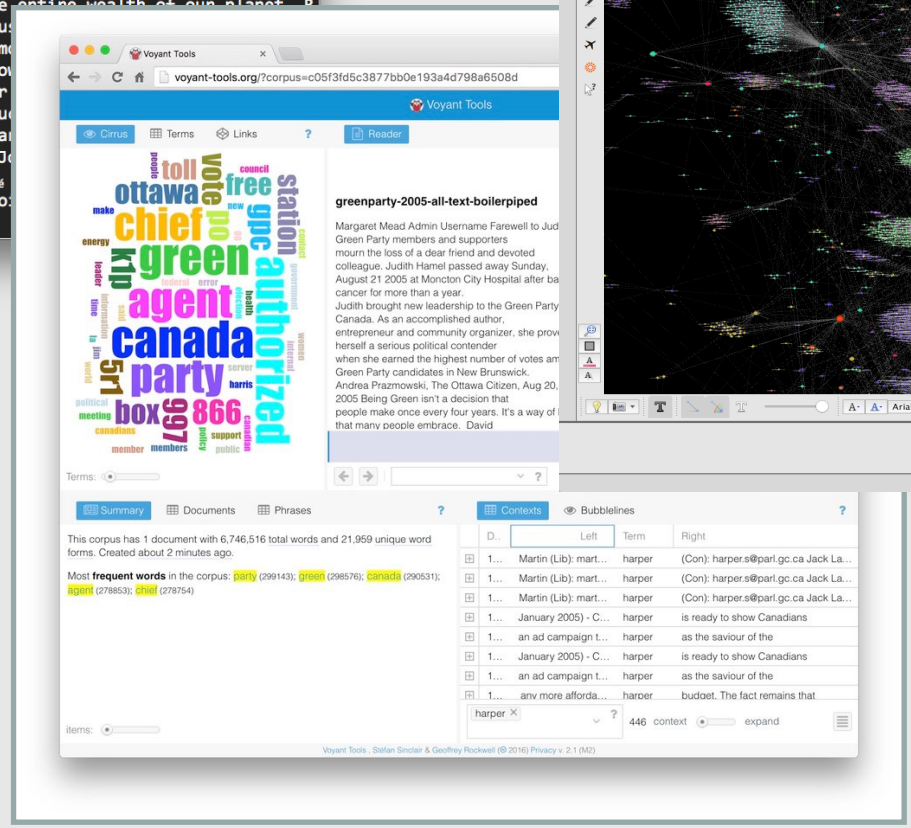


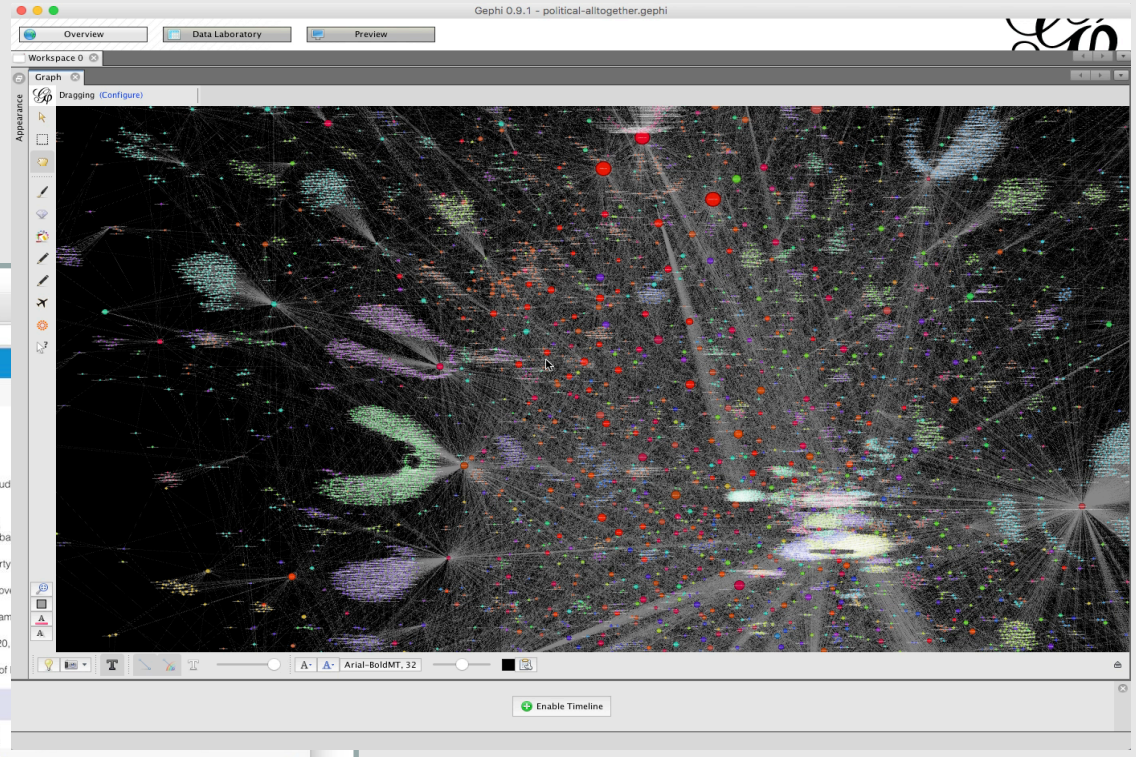
MELLON FOUNDATION · UNIVERSITY OF WATERLOO · YORK UNIVERSITÉ UNIVERSITY

For more information on our project and sponsors, visit archivesunleashed.org/.

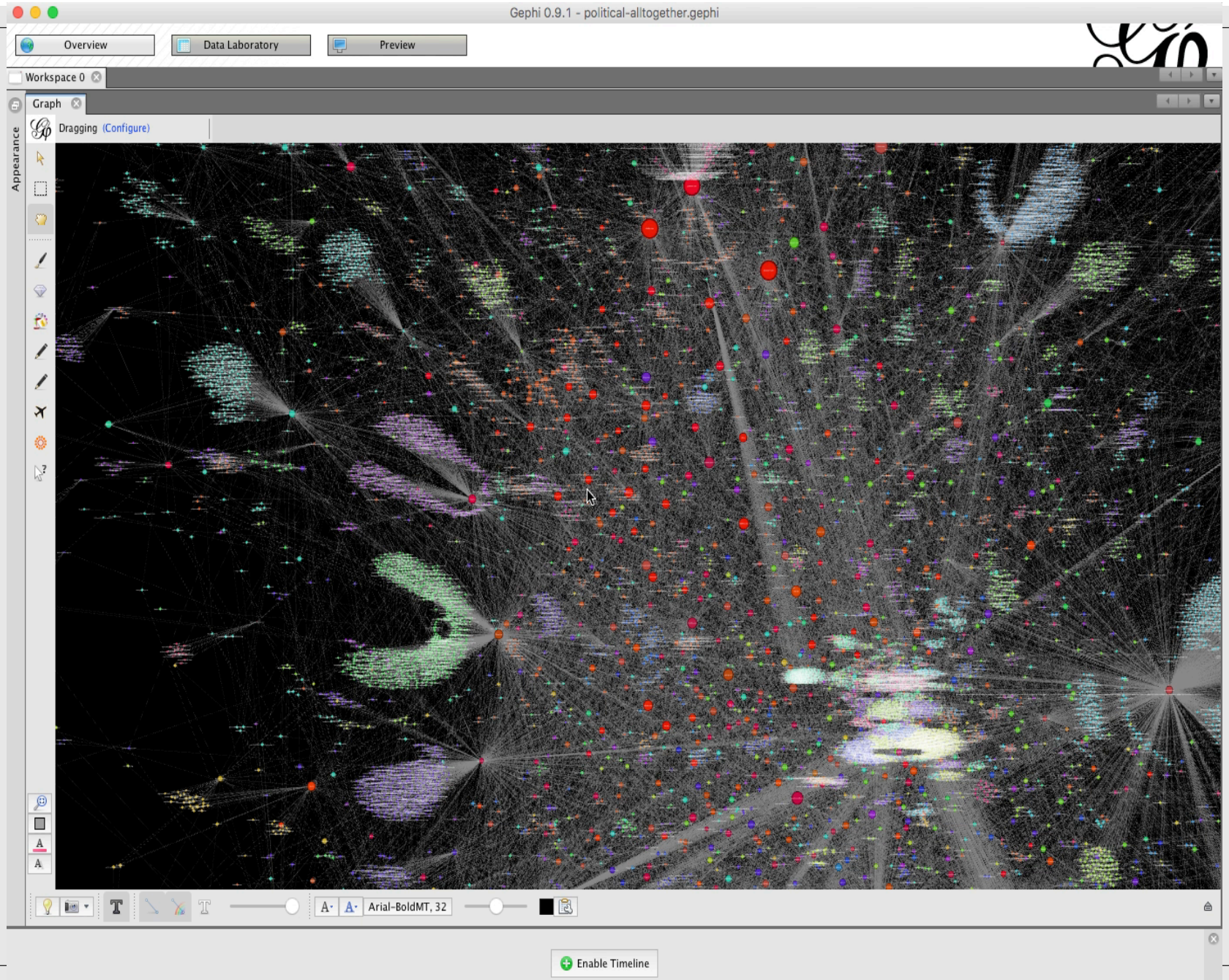
About | Privacy Policy | Documentation | FAQ


1. ssh
(20150805,communist-party.ca,http://communist-party.ca/,HTTP/1.1 200 OK Date: Wed, 05 Aug 2015 22:09
:10 GMT Server: Apache X-Pingback: http://communist-party.ca/xmlrpc.php Link: ; rel=shortlink x-fram
e-options: SAMEORIGIN Vary: Accept-Encoding,User-Agent Connection: close Content-Type: text/html; ch
arset=UTF-8 Communist Party of Canada - Parti Communiste du Canada Parti Communiste du Qu¿bec People
›s Voice Newspaper The Spark! Theory Journal Young Communist League International Communist Movement
Communist Party of Canada _ Parti Communiste du Canada About Short history The Figueroa case Our Co
nstitution Party Program Our aim is socialism Capitalism in Canada Canada in a changing world The Ca
nadian state, nations and peoples of Canada, and the crisis of democracy The working class and peopl
e›s struggle For a People›s Government Building Socialism The Communist Party Campaigns Repeal Bill
C-51! For a People›s Recovery! Save Canada Post Hands off Syria! Contact us How to donate Join the C
ommunist Party We need to move Canada to the left! Nous devons recentrer le Canada vers la gauche!
Vote Communist! The Harper Conservative government›s austerity agenda has been a disaster for the p
eople. Meanwhile, the profits of the big banks and the largest corporations have ballooned. Most peo
ple are now convinced that Harper and his gang must go. What is really required in Parliament are tr
uly progressive voices who will challenge right-wing, pro-corporate policies. We need to move Canada
in a new, progressive direction, with a People›s Alternative Platform based on the needs of working
people and nature! Everywhere, the gap is widening between the majority of the people and a handful
of the super-rich. Today, the top 1% own and control over 50% of the
acism and intolerance are also spreading. The crisis we face isn›t ju
it is about capitalism. It is time capitalism was replaced with a dem
ble system _ with socialism. A new society, based on working class po
owards true democracy _ the rule of the people, by the people and for
nd economic wealth are owned and controlled by the working people. Su
gh collective struggle. Voting Communist sends the strongest and clea
a _ and another world _ is possible, urgent and worth fighting for. J
tter future! Read more. Votons communiste! Les politiques d›aust¿rit¿
fl¿au pour la population canadienne qui constate du m¿me coup la cro
"part-00000" 851L, 4611113C


Voyant Tools
voyant-tools.org/?corpus=c05f3fd5c3877bb0e193a4d798a6508d
Voyant Tools
Cirrus
Terms
Links
Reader
greenparty-2005-all-text-boilerpiped
Summary
Documents
Phrases
Contexts
Bubblelines
This corpus has 1 document with 6,746,516 total words and 21,959 unique word forms. Created about 2 minutes ago.
Most frequent words in the corpus: party (299143); green (298576); canada (290531); agent (278853); chief (278754)
Left
Term
Right
harper
446 context
expand


Gephi 0.9.1 - political-alltogether.gephi
Overview
Data Laboratory
Preview
Workspace 0
Graph
Dragging (Configure)
Appearance
Arial-BoldMT, 32
Enable Timeline

SO WHAT CAN YOU DO WITH LINKS?


Gephi 0.9.1 - political-alitogether.gephi
Overview
Data Laboratory
Preview
Workspace 0
Graph
Dragging (Configure)
Appearance
Arial-BoldMT, 32
Enable Timeline

LET'S DO A QUICK GEPHI WALKTHROUGH.

HACKING THE NEWS

WANT TO KNOW MORE?

HTTPS://CLOUD.ARCHIVESUNLEASHED.ORG/DERIVATIVES

HACKING THE NEWS