WIRE 2014 WORKSHOP

Working with Internet Archives for Research
Welcome

- Opening Session
- Internet Archives & Research Potential
- Building Community: Research Highlights
  - Oxford Internet Institute
  - Centre for Internet Studies & NetLab
  - LS3 & the ALEXANDRIA Project
  - WebScience @ University of Southampton
- Discussion and Challenges
ArchiveHub and Internet Archive Research
AGENDA

1. Large Scale Data
2. Developing New Tools
3. Testing and Building Theory
Opportunity: The Internet Archive contains the largest single record of the history of the World Wide Web from 1995 to the present—a wealth of untapped research data.

Challenge: There is a significant lack of research-ready databases and tools available to the scholarly community.
A sense of scale

The Library of Congress contains approximately 3 PB of data\(^a\)

The Internet Archive contains in excess of 10 PB of archived cultural material

The Wayback Machine contains more than 410 Billion available web pages (as of 2014).

\(^a\)http://blogs.loc.gov/digitalpreservation/2012/03/how-many-libraries-of-congress-does-it-take/
Opportunity: The ArchiveHub project aims to support the creation and dissemination of general guidelines & tools for conducting theoretically and methodologically rigorous longitudinal research using archival Web data.
HistoryTracker Tool

Version 2.0

PIG Scripts in Hadoop Environment

20th Century Collection @ RU
Curated Data Sets

RU High-Speed Computing Cluster

Link Lists & Text Data

Large Scale Data | Developing New Tools | Testing and Building Theory
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Research Potential</th>
<th>Dates</th>
<th>Captures</th>
<th>Unique URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>US Media</td>
<td>Previous studies of news media organizations (Greer &amp; Mensing, 2006; Weber, 2012; Weber &amp; Monge, In Press); focus on evolutionary patterns</td>
<td>2008 - 2012</td>
<td>1,315,132,55</td>
<td>539,184,823</td>
</tr>
<tr>
<td>Occupy Wall Street</td>
<td>Previous research on NGOs in the online environment (Bach &amp; Stark, 2004; Shumate, 2003, 2012; Shumate, Fulk, &amp; Monge, 2005); use of hyperlink data to study the formation and role of alliances between SMOs</td>
<td>2010 - 2012</td>
<td>247,928,272</td>
<td>11,3259,655</td>
</tr>
<tr>
<td>US Senate</td>
<td>Study the growth of political activity in online environments (Adamic &amp; Glance, 2005; Bruns, 2007; Chang &amp; Park, 2012); focus on evolutionary patterns</td>
<td>109 112 Congresses</td>
<td>26,965,770</td>
<td>8,674,397</td>
</tr>
<tr>
<td>US House</td>
<td></td>
<td></td>
<td>51,840,777</td>
<td>12,410,014</td>
</tr>
<tr>
<td>Hurricane Katrina</td>
<td>Online networks and organizational resilience (Chewning, Lai and Doerfel, 2012; Perry, Taylor and Doerfel, 2003) in the wake of disasters; information dissemination</td>
<td>2003 - 2012</td>
<td>1,694,236</td>
<td>663,740</td>
</tr>
<tr>
<td>Superstorm Sandy</td>
<td></td>
<td>2003 - 2012</td>
<td>41,703,112</td>
<td>20,013,455</td>
</tr>
</tbody>
</table>
## What’s in the data?

**Link Data:**

<table>
<thead>
<tr>
<th>Source</th>
<th>Destination</th>
<th>Date</th>
<th>Frequency</th>
<th>Content Type</th>
<th>Bytes</th>
<th>Descriptive Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="http://gawker.com">http://gawker.com</a></td>
<td>2012-10-22</td>
<td></td>
<td></td>
<td></td>
<td>Mitt Romney's Staff Played the Media Covering Them in a Friendly Game of Flag Football</td>
</tr>
</tbody>
</table>

**http://gawker.com**


2012-10-22

Mitt Romney's Staff Played the Media Covering Them in a Friendly Game of Flag Football
Welcome to ArchiveHub!

A community of scholars focused on tackling next generational questions of Internet research using vast repositories of archival data.

http://archivehub.rutgers.edu
Resources: All

- **Full Occupy Wall Street Data**
  - 22 Feb 2014
  - Downloads: [Details]
  - Contributor(s): Matthew Weber
  - The following datasets provides access to the full Occupy Wall Street data, from 2010 through 2013.
  - [Link](http://linuxweb2.scilsnet.rutgers.edu/data/weber/OWSMAIN.zip)

- **Superstorm Sandy dataset**
  - 18 Feb 2014
  - Downloads: [Details]
  - Contributor(s): Allie Kosterich
  - Superstorm Sandy dataset: [Link](http://linuxweb2.scilsnet.rutgers.edu/data/weber/AsburyData.zip)

- **OWS dataset**
  - 05 Feb 2014
  - Downloads: [Details]
  - Contributor(s): Allie Kosterich
  - OWS dataset: [Link](http://linuxweb2.scilsnet.rutgers.edu/data/weber/OWSBothMatching.zip)
Congressional DataSet
Work In Progress
PUTTING BIG THEORY INTO BIG DATA

[or]

moving from observing the Web to observing new phenomenon on the Web
Tracing the Emergence of Organizational Forms

**Environment:**
Organizations compete for scarce resources; during rapid periods of disruption, new entrants seek “protected” niches (Weber & Monge, 2014)

**Population:**
In digital spaces, online connections provide communicative representations of information flows (Weber & Monge, 2012)
Formation of ties (e.g. hyperlinks) can positively impact long-term likelihood of organization survival (Weber, 2012)

**Organization:**
Organizations adapt internally, reconfiguring team structures and developing new routines for knowledge sharing (Ellison, Gibbs & Weber, In Press; Weber & Kim, Under Review)
WHAT IS OUR ONE DEMAND?

#OCCUPYWALLSTREET
SEPTEMBER 17TH.
BRING TENT.
#TRUTH
Posted 1 day ago on Feb. 17, 2014, 2:50 p.m. EST by OccupyWallSt
Tags: memewar

NEWS: RICH PEOPLE PAYING RICH PEOPLE TO TELL MIDDLECLASS PEOPLE TO BLAME POOR PEOPLE

The best way to support Occupy right now is to buy prints of this fabulous poster for yourself and your friends this holiday season.

Please Join Our Mailing List

Occupy Wall Street is a leaderless resistance movement with people of many colors, genders and political persuasions. The one thing we all have in common is that We Are The 99% that will no longer tolerate the greed and
On Our Way To Brooklyn

Posted Oct. 3, 2011, 12:37 a.m. EST by OccupyWallSt

Where: Liberty Square (How to get there)
Help & Directions: +1 (877) 881-3020
General inquiries: general@occupywallst.org
Press inquiries: press@occupywallst.org
Watch: This 10-Min Documentary
Read: Latest NYCGA Minutes

Mail
The UPS Store
Re: Occupy Wall Street
118A Fulton St. #205
New York, NY 10038
Money orders only please, cannot cash checks yet.
Non-perishable goods only. We can accept packages of any size. We're currently low on food.

Occupy Wall Street is leaderless resistance movement with people of many colors, genders and political persuasions. The one thing we all have in common is that We Are The 99% that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary Arab Spring tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

Occupy Wall Street Agenda
Tuesday, October 4, 2011
The 1% have Addresses. The 99% have Messages.

Posted Oct. 25, 2011, 6:43 p.m. EST by OccupyWallSt

Occupy the Boardroom!

Life gets awfully lonely for those at the top. What can we do to let them know someone’s thinking of them? Maybe they need some new friends! We’ve thought of two ways we can help them with that:

Option 1: Pen Pals

Make your voice heard by the Wall Street elites who wrecked the economy and made the rest of us pay. Click on someone below and tell them a story that you think they should listen to. Just got a college degree and nothing to show for it? Just got evicted while your banker gets bonuses? Share your special story with someone who ought to know.

Find a pen pal!

Option 2: Best Friends Forever

Occupy Wall Street is leaderless resistance movement with people of many colors, genders and political persuasions. The one thing we all have in common is that We Are The 99% that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary Arab Spring tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

This #ows movement empowers real people to create real change from the bottom up. We want to see a general assembly in every backyard, on every street corner because we don’t need Wall Street and we don’t need politicians to build a better society.

the only solution is World Revolution

Occupy Wall Street Agenda
Wednesday, October 26, 2011
All of you are missing the point... (User Submitted)

Posted Oct. 14, 2011, 12:24 a.m. EST (12 days ago) by turak

All of you are haggling about money.

Get rid of it.

Instead of a money economy: based upon corruption, opposition, selfishness and greed: create your own self-sufficient communities and economies based upon cooperation, trust and fun.

If you give someone a million dollars: are you helping them? No. They just gave you some pieces of worthless paper...

If someone wants to help you by giving their time and energy and resources, and sharing what they have with you, and you share your time and energy and help them: neither of you need to exchange paper to do this. Human society functions because it is based upon cooperation.

Instead of a paper economy: create an economy based upon hugs & kisses. Have everyone pay for whatever they receive with hugs and kisses. Do you want to be rich in paper money, or rich in hugs and kisses? Create your own self-sufficient communities based upon gratitude: instead of entitlement... communities based upon generosity and giving instead of selfishness and taking.

You're all missing the point and missing the boat. Let the 1% have all the money in the world: then do not
775 Comments

wweddingMadeintheUSA 7 points 7 days ago
Your Idea is cute though
reply permalink

booshington 4 points 5 days ago
Yeah I'm sure 99% of Americans would love to just give up all their money. Because we're known around the world for being a people that doesn't really care much about money.
reply permalink

booshington 1 points 1 day ago
How the hell did this post get so big?!
reply permalink

DKAtoday 1 points 15 hours ago
I don't imagine it's very nice, but when you run into an idiot like turdak, it's just so fun to take a swing and then watch em froth.
reply permalink

republicofolancho 1 points 2 days ago
Olancho is not "cute" but it is spectacular! Embrace change.

https://www.facebook.com/pages/Republic-of-Olancho/149833091753264
www.republicofolancho.com
Big Data… Big Theory?

• Networks are central to social movements in that links between nodes can be influential in collective action

• Examples of nodes includes participants, organizations, media and communications technologies
  • Social networks and social movements (Diani, 2003)

• The interaction between actors, and between actors and hashtags, collectively represent a networked form of organization
  • Network form of organization (Powell, 1990)
Data

- Triangulation of data insulates against false readings from large-scale data (see Lazer, Kennedy, King and Vespignani, 2014)

- Internet Archive:
  - 335 OWS related websites; ~330 million edges over a 2-year period

- Lexis Nexis:
  - Search conducted to assess U.S. newspaper coverage of OWS from the early stages of the movement in September 2011 through Sept. 2012
  - Search OWS keywords, e.g. “Occupy Wall Street,” “Occupy Oakland”

- Twitter
  - Gnip PowerTrack
    - Search by keywords; captures a larger volume of Twitter data than other options
  - Sample includes October 17, 2011, through January 5, 2012. Initial study focused on the critical two-month period from November 1 through December 31, 2011,
  - 750,816 tweets across the two-month period.
Large Scale Data  |  Developing New Tools  |  Testing and Building Theory
OWS News Coverage

Number of Articles in Major US Newspapers

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>July 3-9</td>
<td>100</td>
</tr>
<tr>
<td>July 10-16</td>
<td>200</td>
</tr>
<tr>
<td>July 17-23</td>
<td>300</td>
</tr>
<tr>
<td>July 24-30</td>
<td>400</td>
</tr>
<tr>
<td>Aug 1-7</td>
<td>500</td>
</tr>
<tr>
<td>Aug 8-14</td>
<td>600</td>
</tr>
<tr>
<td>Aug 15-21</td>
<td>700</td>
</tr>
<tr>
<td>Aug 22-28</td>
<td>800</td>
</tr>
<tr>
<td>Aug 29-Sep 4</td>
<td>900</td>
</tr>
<tr>
<td>Sep 5-11</td>
<td>1000</td>
</tr>
<tr>
<td>Sep 12-18</td>
<td>900</td>
</tr>
<tr>
<td>Sep 19-25</td>
<td>800</td>
</tr>
<tr>
<td>Sep 26-Oct 2</td>
<td>700</td>
</tr>
<tr>
<td>Oct 3-9</td>
<td>600</td>
</tr>
<tr>
<td>Oct 10-16</td>
<td>500</td>
</tr>
<tr>
<td>Oct 17-23</td>
<td>400</td>
</tr>
<tr>
<td>Oct 24-30</td>
<td>300</td>
</tr>
<tr>
<td>Nov 1-7</td>
<td>200</td>
</tr>
<tr>
<td>Nov 8-14</td>
<td>100</td>
</tr>
<tr>
<td>Nov 15-21</td>
<td>0</td>
</tr>
<tr>
<td>Nov 22-28</td>
<td>0</td>
</tr>
<tr>
<td>Dec 1-7</td>
<td>0</td>
</tr>
<tr>
<td>Dec 8-14</td>
<td>0</td>
</tr>
<tr>
<td>Dec 15-21</td>
<td>0</td>
</tr>
<tr>
<td>Dec 22-28</td>
<td>0</td>
</tr>
<tr>
<td>Jan 1-7</td>
<td>0</td>
</tr>
<tr>
<td>Jan 8-14</td>
<td>0</td>
</tr>
<tr>
<td>Jan 15-21</td>
<td>0</td>
</tr>
<tr>
<td>Jan 22-28</td>
<td>0</td>
</tr>
<tr>
<td>Feb 1-7</td>
<td>0</td>
</tr>
<tr>
<td>Feb 8-14</td>
<td>0</td>
</tr>
<tr>
<td>Feb 15-21</td>
<td>0</td>
</tr>
<tr>
<td>Feb 22-28</td>
<td>0</td>
</tr>
</tbody>
</table>

Large Scale Data | Developing New Tools | Testing and Building Theory
OWS on the Web

- 335 seed organizations based on records from #OccupyResearch
- Data extracted for 2011 & 2012, based on “both matching”
Maximal Cores (k Coreness)
Large Scale Data | Developing New Tools | Testing and Building Theory
Density


Large Scale Data | Developing New Tools | Testing and Building Theory
Clusters


Large Scale Data  |  Developing New Tools  |  Testing and Building Theory
<table>
<thead>
<tr>
<th>Week</th>
<th>Twitter Mentions as Dyads</th>
<th>Website Hyperlinks as Dyads</th>
<th>Twitter: Clustering Coefficient</th>
<th>Website Clustering Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov. 3, 2011</td>
<td>112889</td>
<td>317</td>
<td>0.0037</td>
<td>0.545</td>
</tr>
<tr>
<td>Nov. 10, 2011</td>
<td>151114*</td>
<td>416</td>
<td>0.0068*</td>
<td>0.495</td>
</tr>
<tr>
<td>Nov. 17, 2011</td>
<td>93249</td>
<td>418</td>
<td>0.0043</td>
<td>0.672*</td>
</tr>
<tr>
<td>Nov. 24, 2011</td>
<td>251406*</td>
<td>496</td>
<td>0.0043</td>
<td>0.689*</td>
</tr>
<tr>
<td>Dec. 1, 2011</td>
<td>78935</td>
<td>430</td>
<td>0.0026</td>
<td>0.677*</td>
</tr>
<tr>
<td>Dec. 8, 2011</td>
<td>66290</td>
<td>531</td>
<td>0.0043</td>
<td>0.661*</td>
</tr>
<tr>
<td>Dec. 15, 2011</td>
<td>35516</td>
<td>601*</td>
<td>0.0071*</td>
<td>0.629</td>
</tr>
<tr>
<td>Dec. 22, 2011</td>
<td>30902</td>
<td>580*</td>
<td>0.0048</td>
<td>0.629</td>
</tr>
<tr>
<td>Dec. 29, 2011</td>
<td>34980</td>
<td>715*</td>
<td>0.0052</td>
<td>0.428</td>
</tr>
</tbody>
</table>

* = p < 0.05 with regards to respective statistical test; t-test was utilized for dyads, and bootstrap t-test was utilized for clustering coefficient.
Challenges:

- **Access Challenges:**
  - Scaling access to the data

- **Data Challenges:**
  - Moving from access to researchable data

- **Research Challenges:**
  - Bridging “big data” to “big theory”
  - Potential for use as a historical research tool
• Want data?
  - Email me! matthew.weber@rutgers.edu
  - ArchiveHub: http://archivehub.rutgers.edu

• The Team
  - Kris Carpenter, Vinay Goel, Internet Archive
  - David Lazer, Katherine Ognyanova, Northeastern University
  - Allie Kosterich, Hai Nguyen, Luan Nguyen, Marya Doerfel, Rutgers University
  - Peter Monge, Ayushman Datta, Kristen Guth, USC

Research supported by NSF Award #1244727 and the NetSCI Lab @ Rutgers