

Examining Online Organizations with Longitudinal Network Data from the World Wide Web

Matthew Weber
Peter Monge
University of Southern California

Sunbelt 2010
July 1, 2010



1. Introduce historical web research

2. Introduce a new tool for mining historical Web data

3. Provide a research example

Organizations Online

The logo for the Internet Archive, featuring a pair of hands holding a glowing, translucent sphere. Inside the sphere is an open book with a flame rising from its center. The text "Internet Archive" is overlaid in white on the sphere.

Internet Archive

Internet Archive

- Non-profit founded in 1996
- Mission to preserve the history of the Internet for future generations
 - Archives included Web sites dating back to 1996, and video, audio and images dating back to 1999.
- As of 2010:
 - 150 billion Web pages
 - 2 petabytes of data
 - Adding 20 terabytes a month

Internet Archive

- What Web sites are included?
 - Any Web site that can be found in Alexa.com's directory is likely to be archived
- Except:
 - Sites with a robots.txt file
 - Javascript
 - Server side image maps (data maintained on the host server)
 - Unknown sites
 - Orphan pages (no links!)

Internet Archive: Free Movies, Music, Books & Wayback Machine

http://www.archive.org/index.php

Getting Started Latest Headlines My del.icio.us post to del.icio.us myUSC 2.0 Welcome to Google YouTube Indiana Chicago PM

Internet Archive: Free Movies, Mus... +

Web Moving Images Texts Audio Software Patron Info About IA Projects

Universes to all know

Forums | FAQs | Contributions | Jobs | Donate

Search: All Media Types Anonymous User

Announcements (2008)

Web 150 billion pages

Wayback Machine

Welcome to the Archive

The Internet Archive, a 501(c)(3) non-profit, is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.

Full Text Lending Library

1.1 Million Digital Books Now Available Free to the Print-Disabled

Thousands of documents from over 3500 federal court cases now freely available

Video Images Browse (by keyword)

230 movies

Curator's Choice (more)

New Amigas

There was a Macintosh, the Amiga the computer for the creative community. This program...

Recent Reviews

LAND:ITALIA.CHE

Live Music Archive Browse (by band)

79,465 concerts

Curator's Choice (more)

Umphreys McGee Live at Bonnaroo - Which Stage on...

Umphreys McGee June 11th, 2010 Bonnaroo - Which Stage Manchester, TN Source: Schoeps MK4v (DIN) >...

Recent Reviews

Grateful Dead Live at Oxford Plains

Audio Browse (by keyword)

509,010 recordings

Curator's Choice (more)

propergol y colargol - no particular destination...

French duet Propergol y Colargol comes back after 6 years of silence with No Particular...

Recent Reviews

bim radio 29 junio 2010

Texts Browse (by keyword)

2,388,430 texts

Curator's Choice (more)

The rise and fall of anarchy in America. From its...

Lawrence J. Gutter Collection of Chicagoans

Recent Reviews

Acoustics And Architecture

Los Angeles Times

DECEMBER 21, 1996

WELCOME NEWS ENTERTAINMENT DESTINATION L.A. ARCHIVES

REGISTRATION

Click here Microsoft Office 95 takes workplace by storm

TOP NEWS STORY

U.S. Acts to Ease HMOs' Cost Pressure on Doctors

Automobiles

Start Your Engines

Check out our new Automobiles section, including 'Your Wheels,' Paul Dean's reviews and Shav Glick's racing column.

Surf & Ski! If you live in California, you can: 1) Set yourself up with the best gear for surfing the Net; 2) Check out some of the coolest winter sports Web sites; and 3) Enter our Surf & Ski Sweepstakes, where you could win one of five FREE deluxe Colorado ski vacations for two!

Up-to-the-minute personalized news, including daily Times stories, are delivered direct to your computer screen on the PointCast Network. No subscription fees! Click here to download.

America Online Users! If you use the AOL browser, you may have trouble registering from other pages on our site. You can always REGISTER directly. Also, our HELP section can guide you through downloading and using Netscape Navigator with AOL.

Our site is best viewed with NETSCAPE Navigator versions 1.1 and above for the Macintosh or versions 1.22 and above for Windows. We also support MICROSOFT Internet Explorer version 2.0 or higher—download it now and enter our contest!

By visiting this site, you are agreeing to the terms of our user agreement. Copyright 1996 Los Angeles Times.

NETSCAPE Navigator Microsoft Internet Explorer

1996

NEWS CONTACT INFO CALIFORNIA EVENTS CONTENTS HELP ABOUT THE TIMES



2000



2004



Internet Archive

- Validity as a tool for research
 - Applied to look at why Web sites create hyperlinks to journal publishers over time (Vaughn & Thelwall, 2003)
 - Reliable tool for measuring the age of a Web site, and for viewing changes to content over time (Murphy, Hashim and O'Conner, 2008)
- Accurate representation of a given page at a single point in time
- Allows for the potential capture of links that exist between Web sites at that point in time
- BUT, due to the numerous challenges facing this type of research, it has not often been utilized (Arms, Aya et. al., 2006)

LATimes (2004)

<http://web.archive.org/web/20040626082124/http://www.latimes.com/>



20040626



June 26, 2004

Daily Pilot (2003)
Oct. 1, 2003

Time Mirror Foundation
(2003)
Oct. 8, 2003

LATimes (2004)
June 26, 2004

LATimes Contact (2004)
August 28, 2004

KTLA (2004)
June 27, 2004





Multiple levels of data

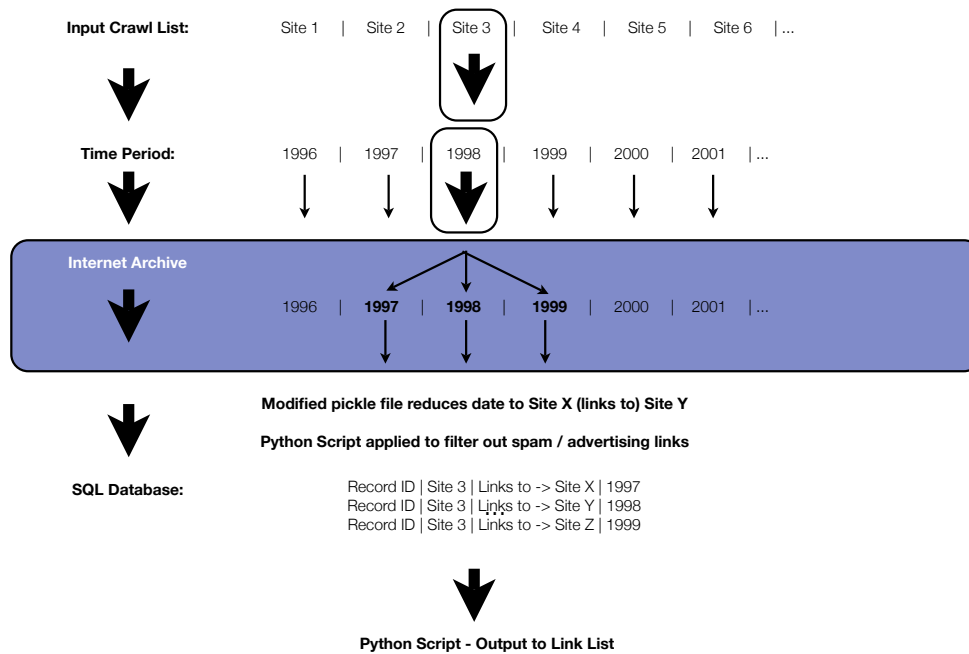
**Great,
but that's a lot of work.**

History Crawl

- Python-based crawler for extracting relational data from the Internet Archive
- Can access Wayback Machine and scrape links, or direct access to IA servers

- History Crawl (Beta)
 - Python 2.6.2 & Navicat MySQL
 - Filters data to remove spam, unwanted sites
 - Exports link lists of data, sortable to month-level
 - Includes URL, date, count

History Crawl (Beta)



Research Example: Newspaper Orgs Online

Newspaper Organizations Online

- Study examining the transition of newspaper organizations from print to online
 - Looked at the development of online news over time (1996 - 2007)
 - Examined effect of hyperlinking strategies
 - Examined interaction with competing forms of media (social media, social networking, online only news, tv, radio, etc.)
- History Crawl
 - Seed organizations - 100 (76)
 - Captured 25,628 Web sites - filtered to 2,977
 - 410 newspaper organizations

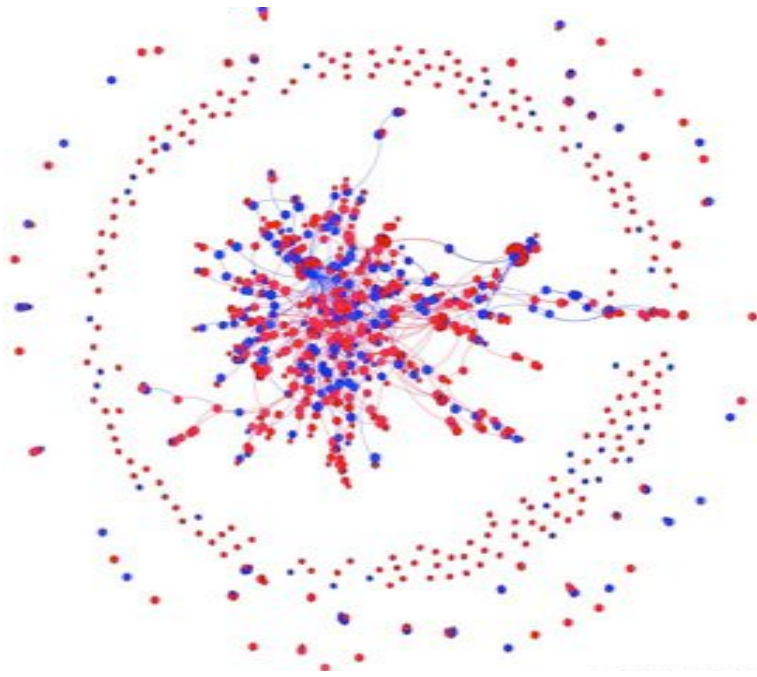
History Crawl Accuracy

link capture analysis

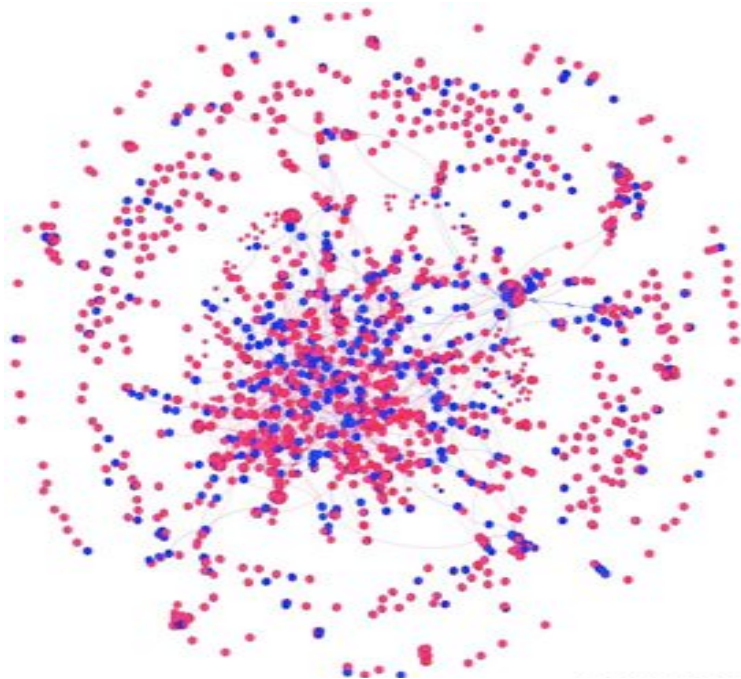
Year	History Crawl	Hand Crawl	Accuracy
January 2000	955	1028	93%
January 2002	1102	1295	85%
January 2004	902	968	93%

Data: History Crawl

Organization Type	Number of Organizations	Average Alexa Pageview Rank	Average Number of Interorganizational Links
Blogs	269	2,480,112	6
Online Only News Source	60	2,731,232	4
Newspapers	410	1,929,415	3
Radio Stations	46	1,003,886	2
Television Stations	15	1,023,308	2
Online Communities (Forums, SNS, etc.)	192	2,476,646	5
Magazines	60	989,512	3



Data: 1999



Data: 2006

Hypotheses - Strategy and Evolution

- H_{1A}: Established organizations that adopt an aggressive link strategy early on during periods of change will continue to form a high proportion of linkages to new populations over time.
- H_{2A}: Established organizations that adopt an aggressive link strategy early on during periods of change will receive a high proportion of linkages from new populations over time.

Methods

- Network Analysis
 - Longitudinal network analysis - RSiena
 - Gephi - visualization

Results

	Attribute-Based Parameter	98-99	99-00	00-01	01-02	02-03	03-04	04-05	05-06	06-07
H1A	<i>CovEgo.blink</i>	1.023 (0.609)	1.095 (0.590)	1.468* (0.500)	2.010* (0.510)	1.980* (0.509)	2.011* (0.512)	2.001* (0.511)	1.950* (0.507)	1.655* (0.697)
	<i>CovEgo.dlink</i>	0.758 (0.609)	0.655 (0.602)	0.608 (0.610)	0.698 (0.603)	0.710 (0.699)	0.707 (0.640)	0.709 (0.632)	0.712 (0.621)	0.766 (0.623)
H2A	<i>CovAlter.blink</i>	0.650 (0.359)	0.880* (0.361)	0.751* (0.329)	0.712* (0.362)	0.810* (0.389)	0.823* (0.409)	0.708 (0.420)	0.785 (0.419)	0.860* (0.411)
	<i>CovAlter.dlink</i>	0.542 (0.361)	0.562 (0.349)	0.583 (0.319)	0.559 (0.310)	0.561 (0.341)	0.563 (0.327)	0.552 (0.320)	0.505 (0.319)	0.490 (0.311)

* indicates significance at $p < 0.05$

Implications

- Strategy is an important factor for understanding how existing organizations adapt to new technological innovations
 - Particularly in an online environment, organizational strategy can play a critical role in position and development over time
 - A clear link exists between immediate choices in online linking and long term organizational performance

History Crawl: Future Directions

- History Crawl v1.0
 - GUI Overlay (allowing for easier URL and variable input)
 - Continuing to test reliability with new research questions
- Evolutionary & ecological research
 - Continuing to examine the development of online news media
 - Increased focus on the effect of social networking sites and user generated news
 - Examining the nature of the link economy through longitudinal analysis of the news community

Annenberg Networks Network:
ascnetworksnetwork.org

My Research:
mediareinvented.com
twitter: [mediareinvented](https://twitter.com/mediareinvented)
matthew.weber@usc.edu

