

Carbon Dating the Web: Estimating the Age of Web Resources

Hany M. SalahEldeen & Michael L. Nelson

Old Dominion University

Department of Computer Science
Web Science and Digital Libraries Lab.



Motivation

In our research in social media, resource sharing, and user intention a question emerged...

When did a certain resource first appear on the web?



First thought: Last Modified Response Header

```
$ curl -I http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html  
HTTP/1.1 200 OK  
Content-Type: text/html; charset=UTF-8  
Expires: Wed, 08 May 2013 14:18:49 GMT  
Date: Wed, 08 May 2013 14:18:49 GMT  
Cache-Control: private, max-age=0  
Last-Modified: Wed, 08 May 2013 08:03:02 GMT  
ETag: "e419d850-22ae-4fe6-a0f4-8ab9477f0c0d"  
X-Content-Type-Options: nosniff  
X-XSS-Protection: 1; mode=block
```



The server responds with the last modified date ...

```
$ curl -I http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html
```

```
HTTP/1.1 200 OK
```

```
Content-Type: text/html; charset=UTF-8
```

```
Expires: Wed, 08 May 2013 14:18:49 GMT
```

```
Date: Wed, 08 May 2013 14:18:49 GMT → Current Server datetime
```

```
Cache-Control: private, max-age=0
```

```
Last-Modified: Wed, 08 May 2013 08:03:02 GMT → Last modified date (Incorrect)
```

```
ETag: "e419d850-22ae-4fe6-a0f4-8ab9477f0c0d"
```

```
X-Content-Type-Options: nosniff
```

```
X-XSS-Protection: 1; mode=block
```



Lacks accuracy

```
$ curl -I http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html
```

```
HTTP/1.1 200 OK
```

```
Content-Type: text/html; charset=UTF-8
```

```
Expires: Wed, 08 May 2013 14:18:49 GMT
```

```
Date: Wed, 08 May 2013 14:18:49 GMT → Current Server datetime
```

```
Cache-Control: private, max-age=0
```

```
Last-Modified: Wed, 08 May 2013 08:03:02 GMT → Last modified date (Incorrect)
```

```
ETag: "e419d850-22ae-4fe6-a0f4-8ab9477f0c0d"
```

```
X-Content-Type-Options: nosniff
```

```
X-XSS-Protection: 1; mode=block
```

→ Problematic as it is inaccurate in a large percentage of cases.

08 May 2013 ≠ 2012-02-11



Last modified date header is not available

```
% curl -I http://temporalweb.net/
```

```
HTTP/1.1 200 OK
```

```
Set-Cookie: 60gpBAK=R1224192509; path=/; expires=Sat, 11-May-2013 03:45:10 GMT
```

```
Date: Sat, 11 May 2013 02:37:55 GMT
```

```
Content-Type: text/html
```

```
Connection: keep-alive
```

```
Set-Cookie: 60gp=R152135972; path=/; expires=Sat, 11-May-2013 03:36:44 GMT
```

```
Server: Apache/2.2.X (OVH)
```

```
Accept-Ranges: bytes
```

```
Vary: Accept-Encoding
```

→ Sometimes it is not present in the response headers.



Second thought: Timestamp on the page



The screenshot shows the BBC News China website. The top navigation bar includes links for News, Sport, Weather, Travel, Future, and Autos. Below this, the main header reads "NEWS CHINA". A secondary navigation bar lists various regions: Home, US & Canada, Latin America, UK, Africa, Asia, Europe, Mid-East, Business, Health, and Sci/Environn. Under the "China" link, there is a sub-link for "India". A red rectangular box highlights the timestamp "8 May 2013 Last updated at 06:14 ET". To the right of the timestamp, there is a "102" icon, a "Share" button, and social media icons for Facebook, Twitter, and Email.

China poisoning case sparks White House petitions flurry

A 19-year-old poisoning case has sparked a rush of interest in the US White House petitions site from Chinese internet users.

Over 130,000 users have signed a petition demanding the US investigate a woman they call a suspect in the 1994 poisoning incident.



But the timestamp is highly
inconsistent

TIME NewsFeed

NATION

Top 10 U.S Cities with the Worst Traffic

By Courtney Subramanian | May 07, 2013 | 3 Comments



... and dependent on the page's style/scheme.



23 dead in Mexico tanker blast

By Mariano Castillo, CNN

updated 10:53 PM EDT, Tue May 7, 2013

So as its location on the page



The image shows a screenshot of a blog post header. At the top is a dark blue navigation bar containing a search icon, a search input field, a 'G+ Share' button, a counter showing '4', and links for 'More' and 'Next Blog'. Below this bar is the main title 'Web Science and Digital Libraries Research Group' in a large, bold, black serif font. Underneath the title is the date 'Saturday, February 11, 2012', which is enclosed in a red rectangular box. Below the date is the post title '2012-02-11: Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone.' in a smaller, bold, black serif font.

Web Science and Digital Libraries Research Group

Saturday, February 11, 2012

2012-02-11: Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone.



Pages' Timestamps Differ

TIME
NewsFeed

NATION

Top 10 U.S Cities with the Worst T

By Courtney Subramanian | May 07, 2013 | 3 Comments



23 dead in Mexico tanker blast

By Mariano Castillo, CNN

updated 10:53 PM EDT, Tue May 7, 2013

Search  Share 4 More ▾

Web Science and Digital Libraries Research Group

Saturday, February 11, 2012

2012-02-11: Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone.

- Very dependent on the page's scheme/style
- Not consistent
- Non-existent sometimes



Shortcomings of using timestamp extraction

- M. Inoue and K. Tajima. Noise robust detection of the emergence and spread of topics on the web. In Proceedings of the 2nd Temporal Web Analytics Workshop, TempWeb '12, pages 9 {16, New York, NY, USA, 2012. ACM
- M. Inoue and K. Tajima developed a technique of extracting creation timestamps on web pages.

Shortcomings:

- Ambiguity (12/07 is it the 12th of July or the 7th of December?).
- Non generalizable.
- Highly dependent on the specific CMS
- Highly dependent on the most prominent timestamp patterns.



**But what if the resource itself
doesn't exist any more?**



Third thought: First existence in public archives



<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing>

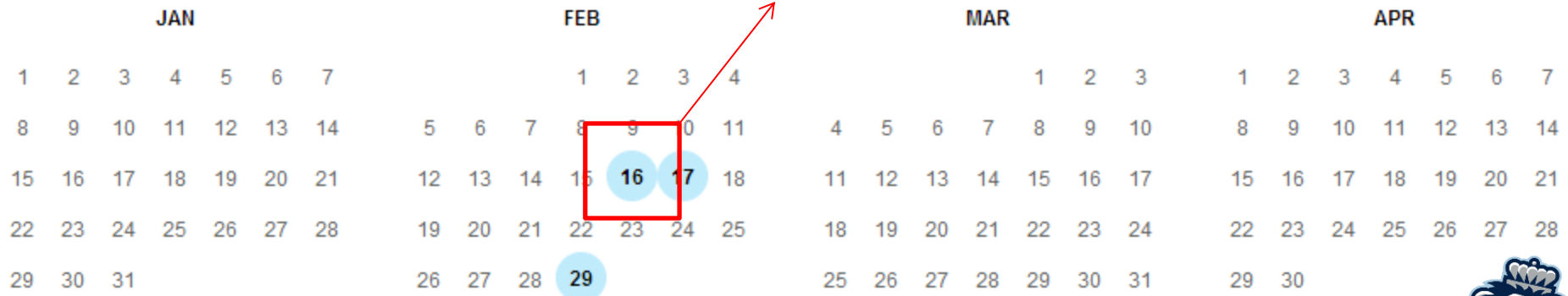
Go Wayback!

<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html> has been crawled 11 times going all the way back to February 16, 2012.

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. [FAQ](#)



Timestamp of the first memento



INTERNET ARCHIVE
WayBackMachine

[Go Wayback!](#)

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. [FAQ](#)

JAN

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

FEB

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29			

MAR

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

APR

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					



Shortcomings:



<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing>

Go Wayback!

<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html> has been crawled 11 times going all the way back to February 16, 2012.

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. [FAQ](#)



JAN

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

FEB

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20
21	22	23	24
25	26	27	28
29			

MAR

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31		

APR

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					



Shortcomings:



<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing>

Go Wayback!

<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html> has been crawled 11 times going all the way back to February 16, 2012.

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. [FAQ](#)



JAN

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

FEB

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20
21	22	23	24
25	26	27	28
29			

MAR

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31		

APR

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					



Shortcomings:



<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing>

Go Wayback!

<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html> has been crawled 11 times going all the way back to February 16, 2012.

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. [FAQ](#)



JAN

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

FEB

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20
21	22	23	24
25	26	27	28
29			

MAR

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31		

APR

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					



Goal

Create a tool that estimates with generality the creation date of the resource without relying on specific infrastructures

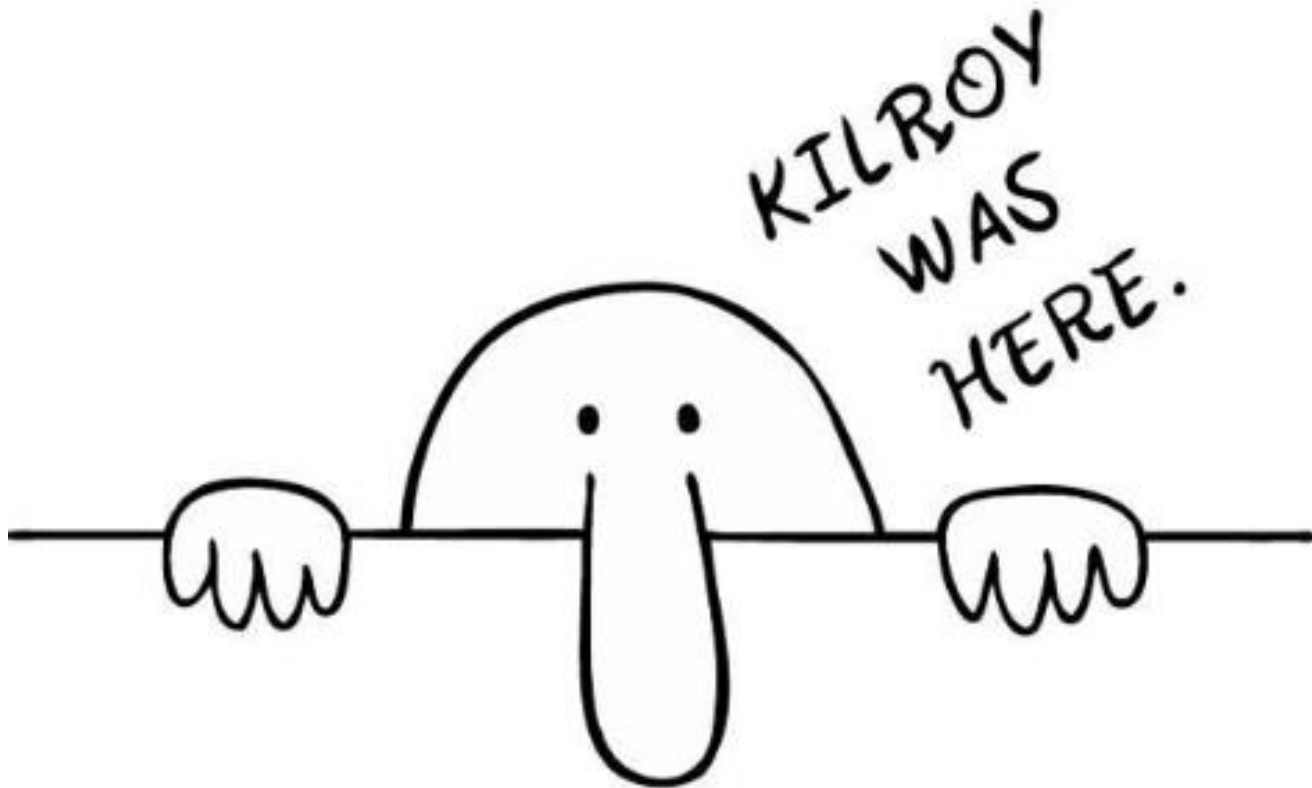
Target Specification

- Doesn't rely on the infrastructure of the hosting web server.
- Doesn't rely on the state and template of the resource.
- Highly generic.
- Fast response with no quarantine periods.
- High accuracy, getting close estimates to real creation date.



Idea

Moving objects leave trails...



Idea

Moving objects leave trails...

Or:

Foo → If you were Aussie

Chad → if you were British



Idea

Web pages leave trails as well since the day they were created...



Web Trails

A web page could leave a trail of one of the following denoting its existence:

- References
- Links (anchors)
- Social media likes and interactions.
- URL shortening.
- Backlinks



The Assumptions

We can propose reasonable assumptions that:

1. We have no prior knowledge of the resource or its hosting web server.
2. The creation date and the publishing date of a resource coincide.

→ **Ex.:** When you write a blog, you publish it as soon as you create it.



Idea

The creation date of any of the associated events/trails could be an estimate of the creation date.



Scenario

Let's consider the following scenario, on Saturday night on the 11th of February of last year I wrote a blog post about my work on the research group's blog page.



After creating the post I tweeted about it ...



<https://twitter.com/hansalaheldeen/status/168704224488730625>



Then it picked up some speed on Twitter and Facebook ...

TOPSY

Search the Social Web

Search

documentation egyptian libraries losing revolution

289

posts

TOP ★5K

tweet

Language

► All languages

Português

English

Français

Deutsch

Network

Google Plus

► Twitter

[Web Science and Digital Libraries Research Group: 2012-02-11: Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone.](http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html)

 ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html – view page – cached page

2012-02-11: Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone.

Interesting posts about this link



sunra sunra

Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone
<http://t.co/n4oxcst8>

02/12/2012 Reply Retweet Favorite 85 similar tweets



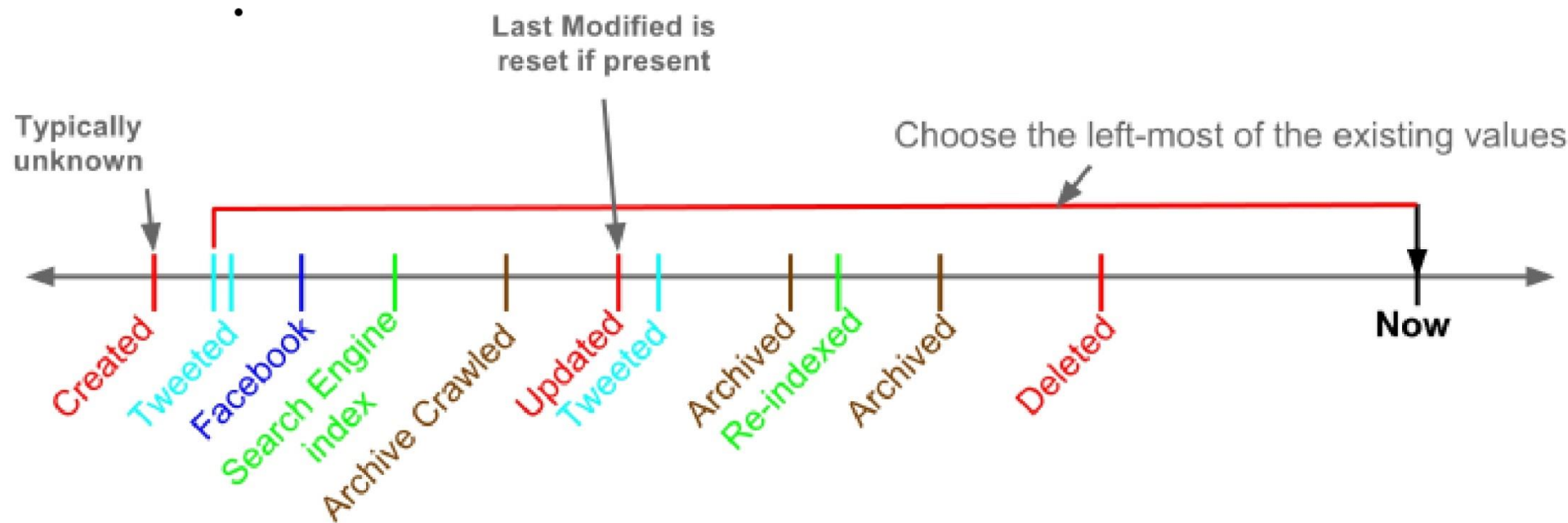
theotypes Theodore Kim **Influential**

@acarvin You may have seen this already. Arab Spring digital content is apparently being lost

<http://topsy.com/http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html>



The timeline of the resource



**Given the events linked to the
existence of the resource we will
examine ways to extract first
observations**



Age Estimation Methods

1. *Resource and server analysis.*
2. Backlinks analysis.
 - a) Web page backlinks.
 - b) Social media backlinks.
3. Archiving analysis.
4. Search engine indexing analysis



Resource and Server Analysis

Examine the server response and extract the last modified date from the header if exists.

```
$ curl -I http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html  
HTTP/1.1 200 OK  
Content-Type: text/html; charset=UTF-8  
Expires: Wed, 08 May 2013 14:18:49 GMT  
Date: Wed, 08 May 2013 14:18:49 GMT  
Cache-Control: private, max-age=0  
Last-Modified: Wed, 08 May 2013 08:03:02 GMT  
ETag: "e419d850-22ae-4fe6-a0f4-8ab9477f0c0d"  
X-Content-Type-Options: nosniff  
X-XSS-Protection: 1; mode=block
```



Observations recorded:

1. Last modified date from the response header.

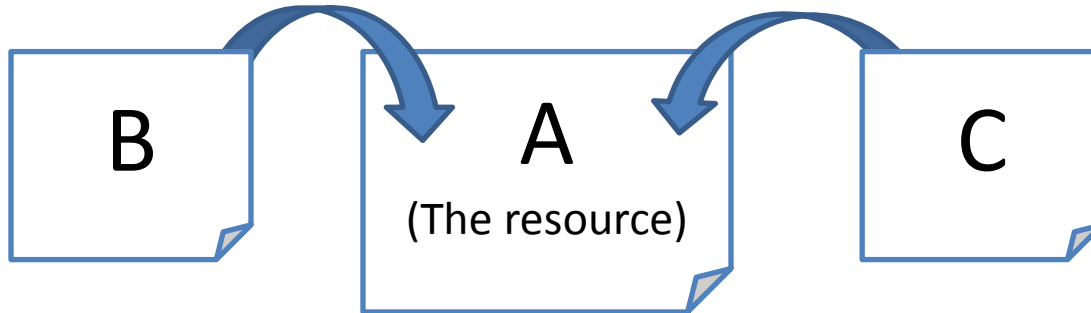


Age Estimation Methods

1. Resource and server analysis.
- 2. *Backlinks analysis.***
 - a) Web page backlinks.
 - b) Social media backlinks.
3. Archiving analysis.
4. Search engine indexing analysis



Backlinks Analysis



- We use Google search API to discover backlinks of A.
- B & C were created after A was created.
- But this assumption is *not completely true*.
- Page B or C could be modified later to its creation of A

Time Magazine

Ex.: If the front page of Time magazine decided to finally feature me as “Person of the Year”

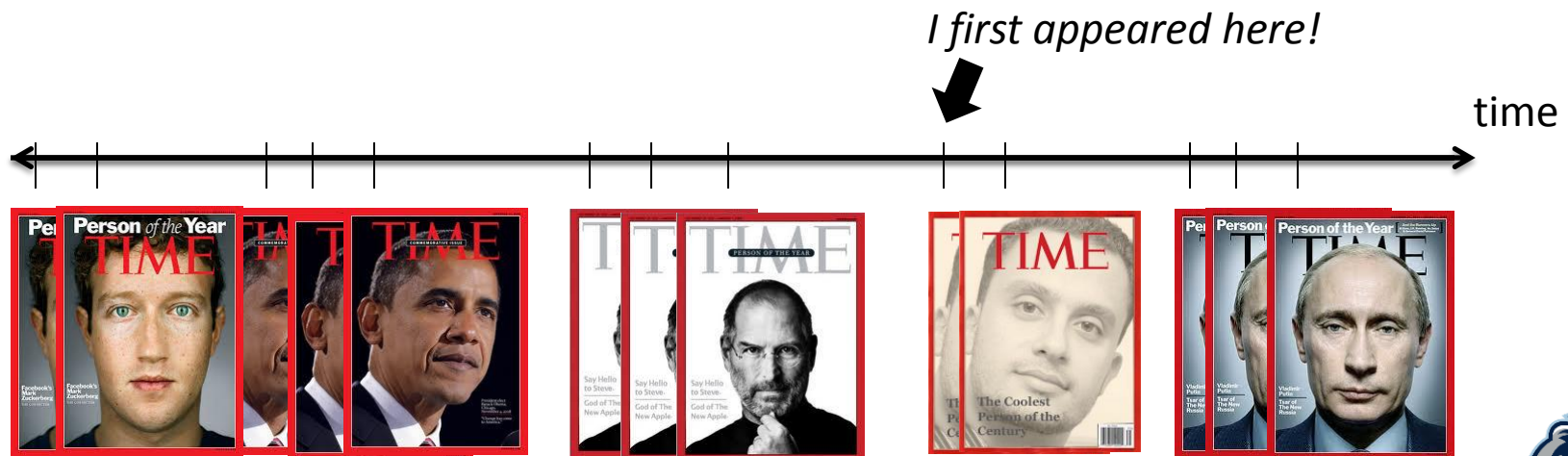
In this case page B (Time magazine’s front page) was modified to point to my page A



When did the link first appear?

To solve this problem:

1. We extract the timemap of the archived mementos of B.
2. Perform binary search to allocate the **first appearance** of the link to A in B.
3. Get the timestamp of that first memento.



Observations recorded:

1. Last modified date from the response header.
2. First Appearance of a backlink.



Social Media Backlinks

- Similarly, you create a social backlink when you tweet about a page



Topsy Otter API



sunra sunra

Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone <http://t.co/n4oxcst8>

02/12/2012 Reply Retweet Favorite 85 similar tweets



theotypes Theodore Kim **Influential**

@acarvin You may have seen this already. Arab Spring digital content is apparently being lost

02/15/2012 Reply Retweet Favorite 25 similar tweets



blakehounshell Blake Hounshell **Highly Influential**

Social media revolutions are bad news for historians RT @NiemanLab: 404: Revolution Not Found <http://t.co/U5T01Kur>

02/15/2012 Reply Retweet Favorite 23 similar tweets



hanyssalaheldeen Hany SalahEldeen

After less than a year, more than 10% of the social media about the Egyptian revolution is gone forever! <https://t.co/sDawQZxa> #Egypt

02/12/2012 Reply Retweet Favorite 12 similar tweets



pomeranian99 Clive Thompson **Influential**

After only one year, 10% of social media about Egyptian revolution is gone: <http://t.co/rJexe1yp> (As per @anildash: <http://t.co/ZhkwZzj9>)

02/15/2012 Reply Retweet Favorite 11 similar tweets



ndsa2 NDSA

RT @ndiipp: Losing My Revolution: 10% of Egyptian Revolution social media documentation lost. <http://t.co/CkGyT7ih> #digitalpreservation

07/30/2012 Reply Retweet Favorite 9 similar tweets

Up to 500 Tweets



Topsy Otter API



sunra sunra

Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone
<http://t.co/n4oxcst8>

02/12/2012 Reply Retweet Favorite 85 similar tweets



theotypes Theodore Kim **Influential**

@acarvin You may have seen this already. Arab Spring digital content is apparently being lost

02/15/2012 Reply Retweet Favorite 25 similar tweets



blakehounshell Blake Hounshell **Highly Influential**

Social media revolutions are bad news for historians RT @NiemanLab: 404: Revolution Not Found <http://t.co/U5T01Kur>

02/15/2012 Reply Retweet Favorite 23 similar tweets



hanyssalaheldeen Hany SalahEldeen

After less than a year, more than 10% of the social media about the Egyptian revolution is gone forever!

<https://t.co/sDawQZxa> #Egypt

02/12/2012 Reply Retweet Favorite 12 similar tweets



pomeranian99 Clive Thompson **Influential**

After only one year, 10% of social media about Egyptian revolution is gone: <http://t.co/rJexe1yp> (As per @anildash: <http://t.co/ZhkwZzj9>)

02/15/2012 Reply Retweet Favorite 11 similar tweets



ndsa2 NDSA

RT @ndiipp: Losing My Revolution: 10% of Egyptian Revolution social media documentation lost. <http://t.co/CkGyT7ih>
#digitalpreservation

07/30/2012 Reply Retweet Favorite 9 similar tweets

Different shortened versions



Topsy Otter API



sunra sunra

Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone
<http://t.co/n40xcst8>

02/12/2012 Reply Retweet Favorite 85 similar tweets



theotypes Theodore Kim **Influential**

@carvin You may have seen this already. Arab Spring digital content is apparently being lost

02/15/2012 Reply Retweet Favorite 25 similar tweets



blakehounshell Blake Hounshell **Highly Influential**

Social media revolutions are bad news for historians RT @NiemanLab: 404: Revolution Not Found <http://t.co/U5T01Kur>

02/15/2012 Reply Retweet Favorite 23 similar tweets



hanyalaheldeen Hany SalahEldeen

After less than a year, more than 10% of the social media about the Egyptian revolution is gone forever!

<https://t.co/sDawQZxa> #Egypt

02/12/2012 Reply Retweet Favorite 12 similar tweets



pomeranian99 Clive Thompson **Influential**

After only one year, 10% of social media about Egyptian revolution is gone: <http://t.co/rJexe1yp> (As per @anildash: <http://t.co/ZhkwZzj9>)

02/15/2012 Reply Retweet Favorite 11 similar tweets



ndsa2 NDSA

RT @ndiipp: Losing My Revolution: 10% of Egyptian Revolution social media documentation lost. <http://t.co/CkGyT7ih>
#digitalpreservation

07/30/2012 Reply Retweet Favorite 9 similar tweets

Break ties via the API epoch



Observations recorded:

1. Last modified date from the response header.
2. First Appearance of a backlink.
3. First Tweet published.



URL Shortening

http://bit.ly/losing_revolution

<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html>

bitly

Your stuff

Your network

+ Paste a link here...



Web Science and Digital Libraries Research Group: 2013-04-19: Carbon Dating...

<http://ws-dl.blogspot.com/2013/04/2013-04-19-carbon-dating-web.html>



heinstein

created this bitly link on Apr 19, 2013.



20

clicks on
your bitly link

100%

of clicks from
your bitly link

bit.ly/CarbonDating...

Copy



All 20 clicks to this content came from this bitly link.

Creation Date of the Bitly

Extract number of clicks



Observations recorded:

1. Last modified date from the response header.
2. First Appearance of a backlink.
3. First Tweet published.
4. First Bitly Shortened URL created.



Age Estimation Methods

1. Resource and server analysis.
2. Backlinks analysis.
 - a) Web page backlinks.
 - b) Social media backlinks.
- 3. *Archiving analysis.***
4. Search engine indexing analysis



Archives Analysis

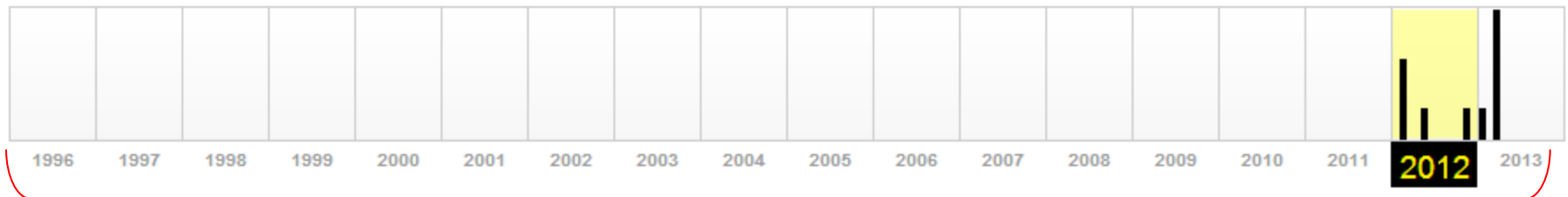


<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing>

Go Wayback!

<http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html> has been crawled **11** times going all the way back to February 16, 2012.

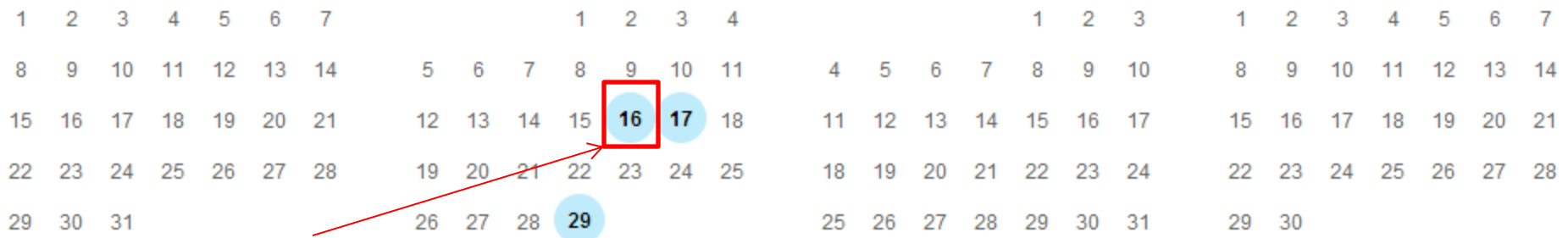
A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. [FAQ](#)



JAN

FEB

Download the memento timemaps of the resource



Get timestamp of first memento

- Furthermore, if the original headers exist for the first memento we extract the original last modified date.



Observations recorded:

1. Last modified date from the response header.
2. First Appearance of a backlink.
3. First Tweet published.
4. First Bitly Shortened URL created.
5. Time stamp of first memento in the archives.



Age Estimation Methods

1. Resource and server analysis.
2. Backlinks analysis.
 - a) Web page backlinks.
 - b) Social media backlinks.
3. Archiving analysis.
4. *Search engine indexing analysis*



Search Engine Index Analysis

Google

http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html

Web Images Maps Shopping More Search tools

About 60,600 results (0.52 seconds)

Losing My Revolution - Web Science and Digital Libraries Research ...
ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html ▾
Feb 11, 2012 ← 2012-02-11: Losing My Revolution: A year after the Egyptian Revolution, 10% ... As an Egyptian in the WS-DL research group at ODU, web preservation of the Revolution is of particular interest. ... Using curl, we see that the web site returns an HTTP response of "503 ... Content-Type: text/html; charset=utf-8 ... Carlton Northern shared this

TpdI Doctoral consortium 2012 - SlideShare
www.slideshare.net/heinestien/tpdl-doctoral-consortium-2012 ▾
Nov 6, 2012 ← ... 10% of the social media documentation is gone.
http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html –

Last crawled dates

- We use Google's search API to extract the **last crawled date**
- Relatively short time between resource creation and search engine discovery.
- **Drawback:** Granularity is by day not by time.



Observations recorded:

1. Last modified date from the response header.
2. First Appearance of a backlink.
3. First Tweet published.
4. First Bitly Shortened URL created.
5. Time stamp of first memento in the archives.
6. Date of the last crawl by the search engine.

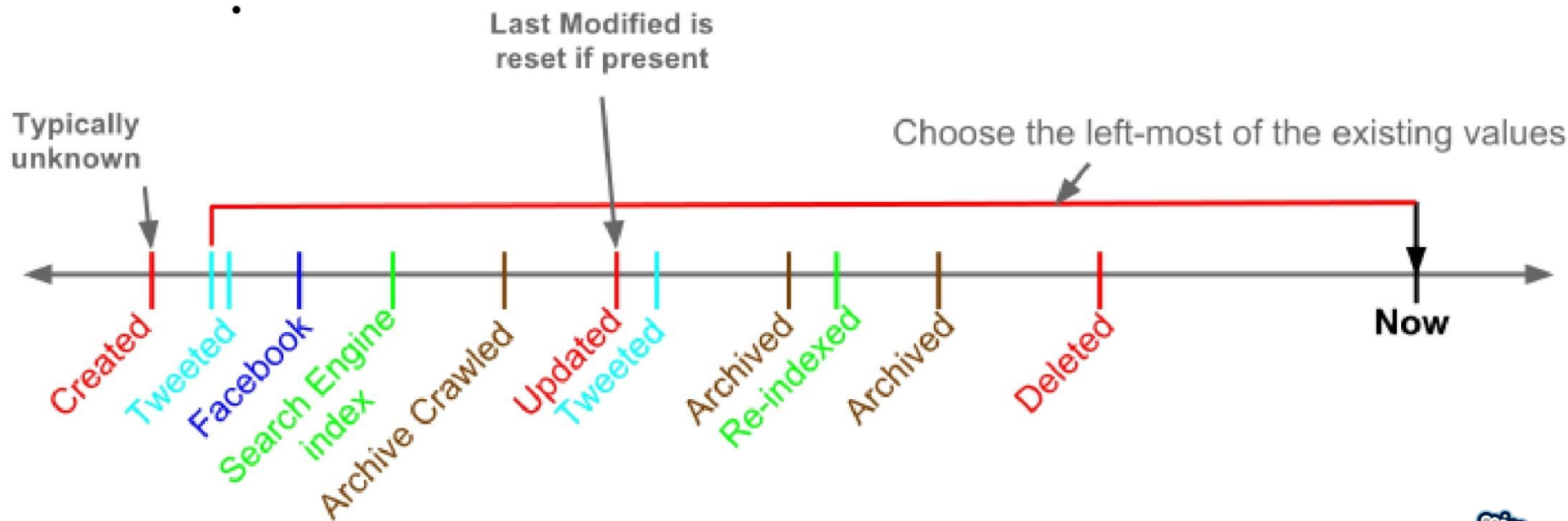


Ok, now we have a collection of sources that return creation dates, what will we do next?



Timestamps Accumulation

- We collect the obtained dates and get the leftmost creation date recorded.



Timestamps Accumulation

```
"URI": "http://www.mementoweb.org",  
"Estimated Creation Date": "2009-09-30T11:58:25",  
"Last Modified": "2012-04-20T21:52:07",  
"Bitly": "2011-03-24T10:44:12",  
"Topsy.com": "2009-11-09T20:53:20",  
"Backlinks": "2011-01-16T21:42:12",  
"Google.com": "2009-11-16",  
"Archives": {  
  "Earliest": "2009-09-30T11:58:25",  
  "By Archive": {  
    "wayback.archive-it.org": "2009-09-30T11:58:25",  
    "api.wayback.archive.org": "2009-09-30T11:58:25",  
    "webarchive.nationalarchives.gov.uk": "2010-04-02T00:00:00"  
  }  
}
```



Next step: Verifying our methods



Estimated Age Verification

1. **Collect** a dataset of webpages of known creation/publishing date.
2. **Compare** the estimated results from our method and the actual dates recorded.



Gold Standard Data Collection

We collect the pages from 4 difference categories of collections to ensure variation.

1. News Sites.
2. Social Media and Blogs.
3. Long Standing Domains.
4. Manual Random Extraction.



News Sites

Using RSS and Atom feeds or XML sitemaps we extracted numerous pages along with their respective publishing dates.

1. Google News (29,154 pages)
2. BBC (3,703 pages)
3. CNN (18,519 pages)
4. Yahoo News (34,588 pages)
5. The Hollywood Gossip (6,859 pages)



Social Sites

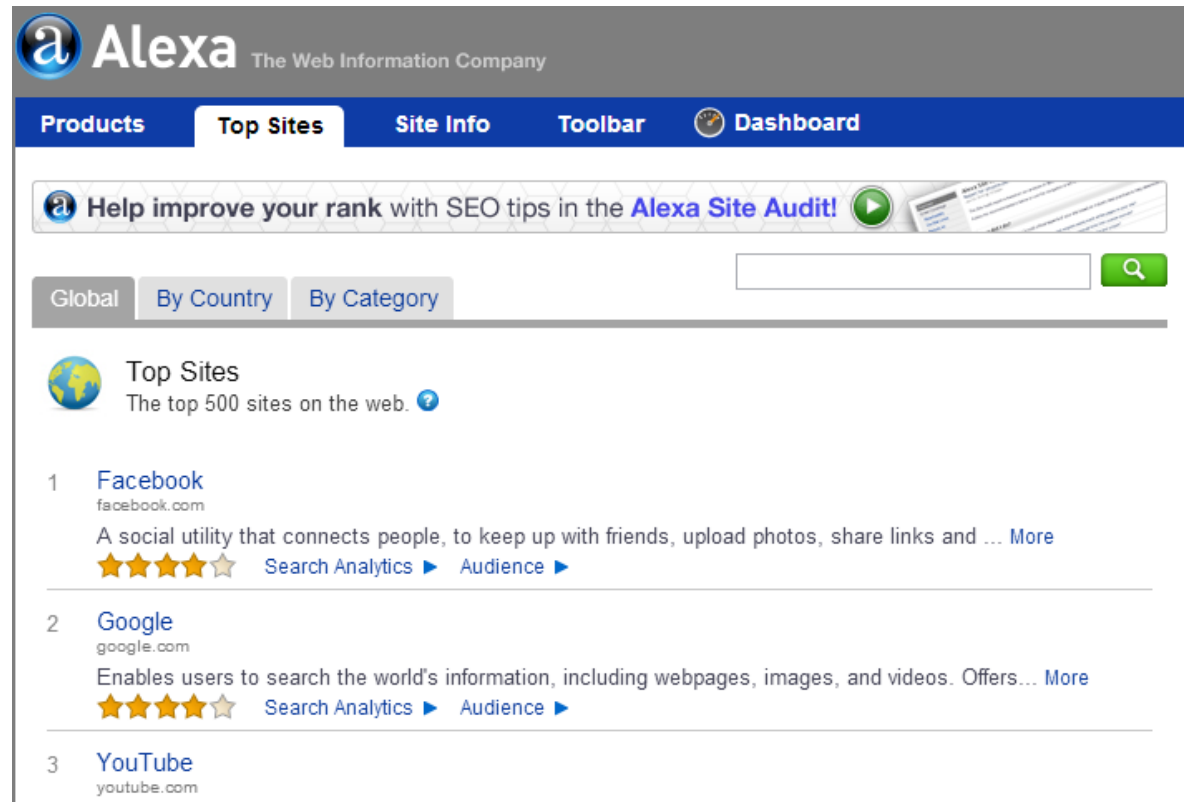
We randomly selected different resources with no regard to popularity to avoid the inherent bias:

1. [Pinterest](#) (55,463 posts)
2. [Tumblr](#) (52,513 posts)
3. [Youtube](#) (78,000 posts)
4. [Word Press](#) (2,405,901 posts)
5. [Blogger](#) (32,417 posts)



Long Standing Domains

- Extract the top 500 domains from Alexa.com
- Query their DNS registry dates.
- Were able to extract 167 dates.



The screenshot shows the Alexa website interface. At the top, the Alexa logo and "The Web Information Company" are displayed. Below this is a navigation bar with tabs for "Products", "Top Sites", "Site Info", "Toolbar", and "Dashboard". A banner below the navigation bar promotes the "Alexa Site Audit" tool. Underneath, there are tabs for "Global", "By Country", and "By Category", followed by a search bar and a magnifying glass icon. The main content area is titled "Top Sites" with a globe icon and the subtitle "The top 500 sites on the web." It lists the top three sites: 1. Facebook (facebook.com), 2. Google (google.com), and 3. YouTube (youtube.com). Each entry includes a brief description, a star rating, and links for "Search Analytics" and "Audience".

Rank	Site Name	Domain	Description	Rating	Search Analytics	Audience
1	Facebook	facebook.com	A social utility that connects people, to keep up with friends, upload photos, share links and ...	★★★★★	Search Analytics	Audience
2	Google	google.com	Enables users to search the world's information, including webpages, images, and videos. Offers...	★★★★★	Search Analytics	Audience
3	YouTube	youtube.com				

Manual Random Extraction

- We extracted 90 different random URLs obtained from random walks on the web, visually inspected them to extract the creation date.
- The 10 URLs analyzed by Jatowt et al.*

* A. Jatowt, Y. Kawai, and K. Tanaka. Detecting age of page content. In Proceedings of the 9th annual ACM international workshop on Web information and data management, WIDM '07, pages 137--144, New York, NY, USA, 2007. ACM.



Gold Standard Data Collection

	Data Sources	Resources Collected	Sampled Resources	Timestamp Allocation Method
News Sites	news.Google.com	29,154	100	XML sitemap
	BBC.co.uk	3,703	100	Page Scraping
	CNN.com	18,519	100	Page Scraping
	news.Yahoo.com	34,588	100	XML sitemap
	theHollywoodGossip.com	6,859	100	Page Scraping
Social Sites	Pinterest.com	55,463	100	RSS feed
	Tumblr.com	52,513	100	RSS feed
	Youtube.com	78,000	100	Search API
	WordPress.com	2,405,901	100	Atom feed
	Blogger.com	32,417	100	Atom feed
	Alexa.com Top Domains	167	100	Page Scraping & Who.is service
	Manual Extraction	100	100	Manual inspection
	Total:	2,717,384	1,200	

→ From each we randomly selected 100 unique URLs to create our gold standard dataset



Evaluation

- Applied our 6 methods on 1200 resources.
- Get leftmost estimation.

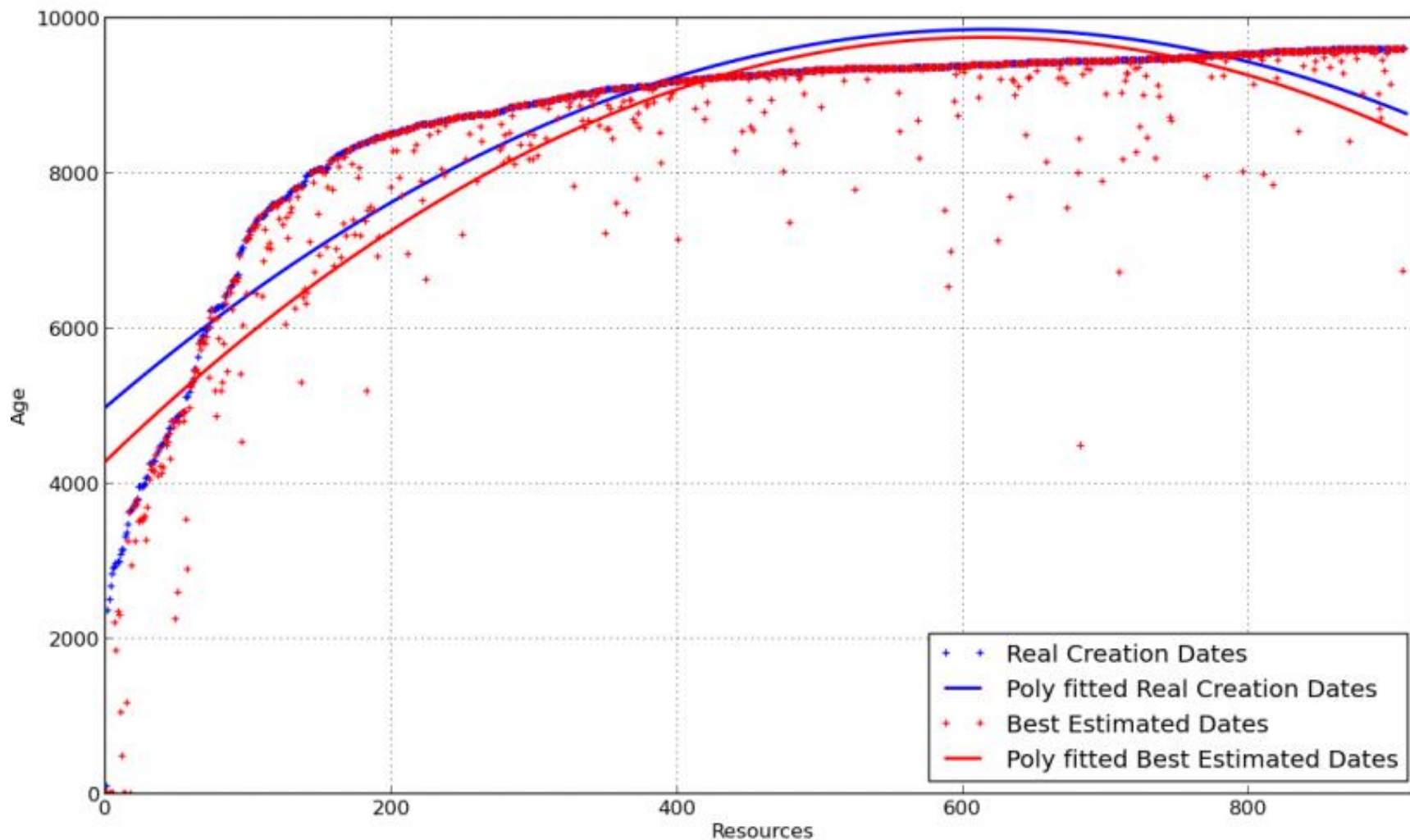
	Number of Resources	Percentage
An estimation found	910	76%
Exact matching estimation	393	33%
No estimation found	290	24%
Total Resources	1200	100%



Evaluation

Age Estimation Method	Resources Found By Using The Method As Best Estimate	Percentage Of Resources Found	Contribution	
			Resources Contributed	Percentage Contributed
Google	370	40.66%	709	59.13%
Topsy	236	25.93%	632	52.71%
Archives	152	16.70%	578	48.21%
Bitly	96	10.55%	554	46.21%
Last Modified	53	5.82%	134	11.18%
Backlinks	3	0.33%	180	15.01%
Total Estimate	910	75.90%	1199	100%

Actual Vs. Estimated Dates



So what happens if one of these 6 methods failed?



Isolation and Elimination

Age Estimation Method	Area Under Curve	
	AUC	Percentage lost in AUC
Topsy	720.61	5.51%
Last Modified	725.59	4.86%
Archives	741.23	2.81%
Google	742.52	2.64%
Bitly	758.73	0.51%
Backlinks	762.64	0%
Total Estimate	762.64	0%

Carbon Date API

← → ↻ cd.cs.odu.edu/cd/ ☆

Carbon Dating The Web

Predict the Birth day of a webpage!

```
{
  "URI": "http://www.cs.odu.edu/",
  "Estimated Creation Date": "1996-02-09T21:47:46",
  "Last Modified": "",
  "Bitly.com": "2011-08-31T14:05:22",
  "Topsy.com": "2012-06-19T19:26:50",
  "Backlinks": "2003-02-18T20:54:12",
  "Google.com": "",
  "Archives": {
    "Earliest": "1996-02-09T21:47:46",
    "By_Archive": {
      "api.wayback.archive.org": "1996-02-09T21:47:46",
      "webcitation.org": "2012-01-22T21:01:29"
    }
  }
}
```

Web Science and Digital Libraries - Department of Computer Science, Old Dominion University, Norfolk VA - 23529



<http://cd.cs.odu.edu/cd/<Your URL Here>>

```
curl -i http://cd.cs.odu.edu/cd/http://www.mementoweb.org
```

```
HTTP/1.0 200 OK
```

```
Date: Fri, 01 Mar 2013 04:44:47 GMT
```

```
Server: WSGIServer/0.1 Python/2.6.5
```

```
Content-Length: 550
```

```
Content-Type: application/json; charset=UTF-8
```

```
{
  "URI": "http://www.mementoweb.org",
  "Estimated Creation Date": "2009-09-30T11:58:25",
  "Last Modified": "2012-04-20T21:52:07",
  "Bitly": "2011-03-24T10:44:12",
  "Topsy.com": "2009-11-09T20:53:20",
  "Backlinks": "2011-01-16T21:42:12",
  "Google.com": "2009-11-16",
  "Archives": {
    "Earliest": "2009-09-30T11:58:25",
    "By Archive": {
      "wayback.archive-it.org": "2009-09-30T11:58:25",
      "api.wayback.archive.org": "2009-09-30T11:58:25",
      "webarchive.nationalarchives.gov.uk": "2010-04-02T00:00:00"
    }
  }
}
```



Carbon Date API on GitHub

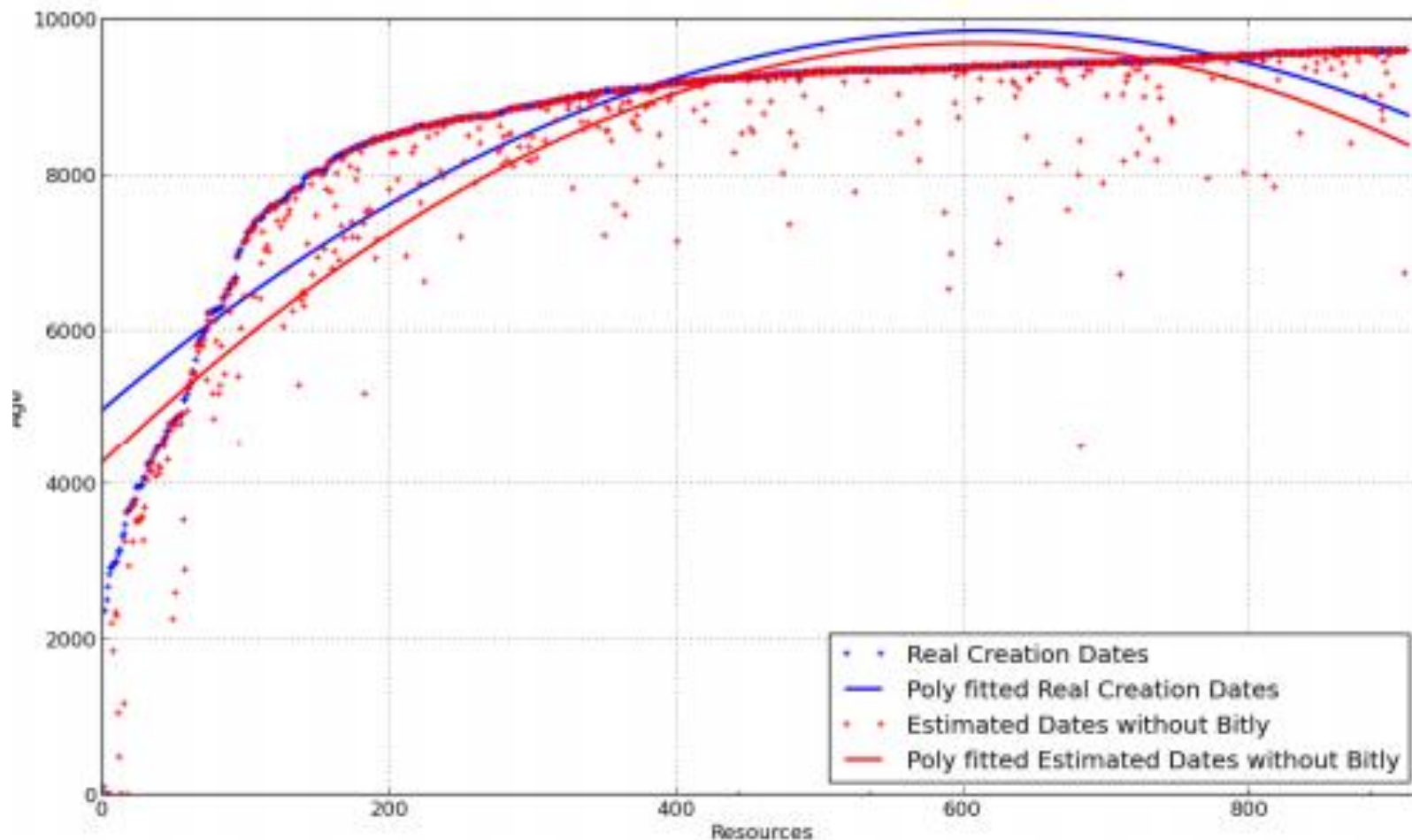
- Due to the slow response we advise that you download the module and install it on your machine.
- <https://github.com/HanySalahEldeen/CarbonDate>



Extra Slides

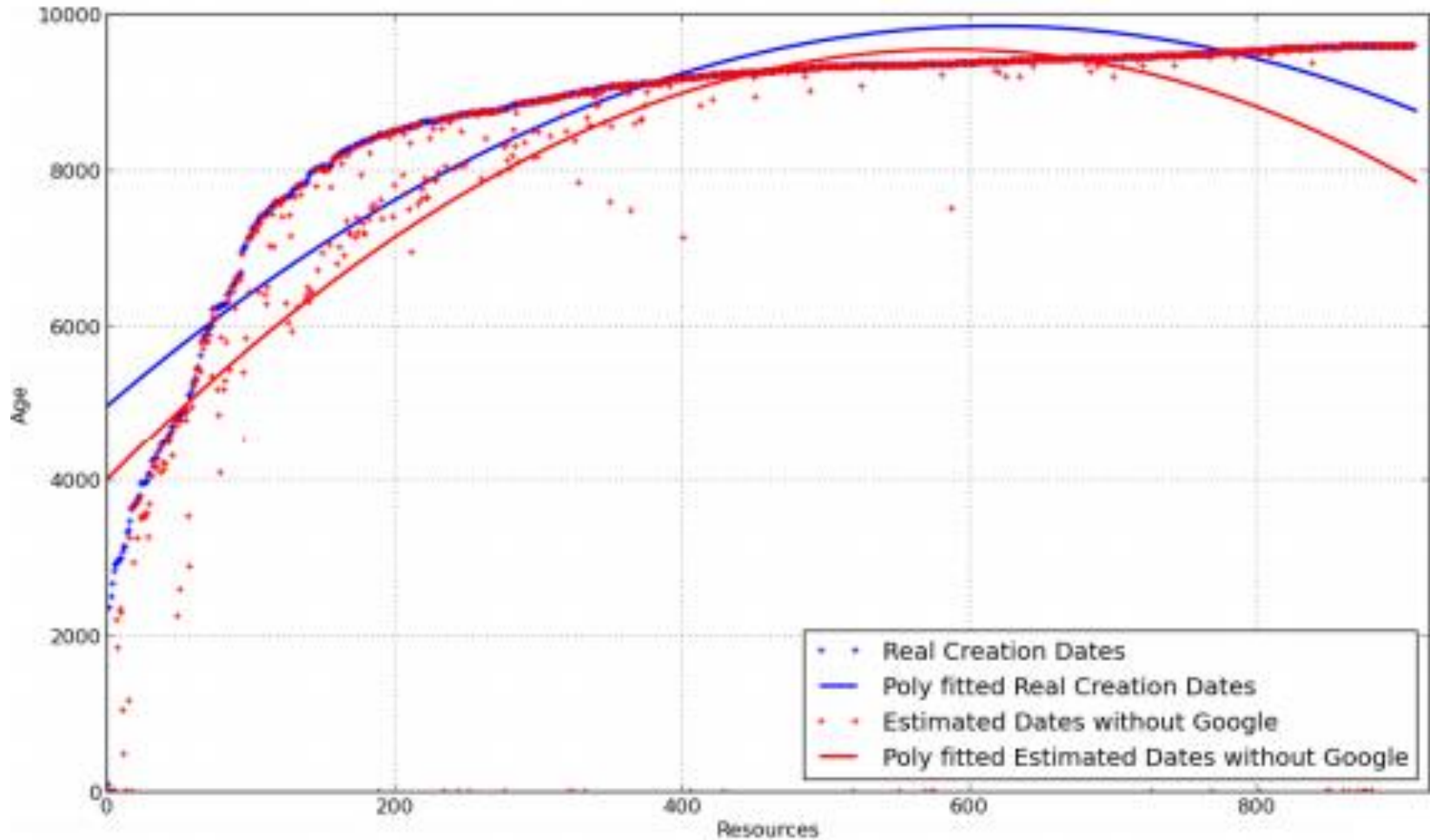


Without Bitly



(a) Without Bitly, AUC=758.73

Without Google



(b) Without Google, AUC=742.52