This is the authors' accepted manuscript and may contain minor differences from the published version. Please cite as:

Hale, Scott A., Blank, Grant, and Alexander, Victoria D. (2017). 'Live versus archive: Comparing a web archive and to a population of webpages.' In Niels Brügger and Ralph Schroeder (Eds.), *The Web as History*, London: UCL Press, pp. 45-61. https://www.ucl.ac.uk/ucl-press/browse-books/the-web-as-history.

Live versus Archive: Comparing a Web Archive and to a Population of Webpages

Scott A. Hale,[1,2] Grant Blank,[2] & Victoria D. Alexander[3]

1 Alan Turing Institute, London
2 Oxford Internet Institute, University of Oxford
3 Goldsmiths University of London

Introduction

With its seemingly limitless scope, the World Wide Web promises enormous advantages, along with enormous problems, to researchers who seek to use it as a source of data. Websites change continually and a high level of flux makes it challenging to capture a snapshot of the web, or even a cross-section of a small subset of the web. A web archive, such as those at the Internet Archive, promises to store and deliver repeated cross-sections of the entire web, and it also offers the potential for longitudinal analysis. Whether this potential is realized depends on the extent to which the archive has truly captured the web. Therefore, a crucial question for Internet researchers is: 'How good are the archival data?'

We ask if there are systematic biases in the Internet Archive, using a case study to address this question. Specifically, we are interested in whether biases exist in the British websites stored in the Internet Archive data. We find that the Internet Archive contains a surprisingly small subset, about 24%, of the webpages of the website that we use for our case study (the travel site, TripAdvisor). Furthermore, the subset of data we found in the Internet Archive appears to be biased and is not a random sample of the webpages on the site. The archived data we examine has a bias toward prominent webpages. This bias could create serious problems for research using archived websites, and we discuss this issue at the end of the chapter.

The web has always been an extremely dynamic object. One widely quoted study found that 35–40% of webpages changed content in any given week (Fetterly, Manasse, Najork & Weiner 2004). Another study found that 26% of all webpages visited by users twice within an hour had changed content, and 69% of webpages revisited within a day had changed (Weinreich, Obendorf, Herder, & Mayer, 2008). For researchers interested in the evolution of the web or any part of the web (like the diffusion of certain web technologies), this is a serious challenge. They need historical data, and almost all of this history is lost.

This problem was recognized early in the development of the web, and the Internet Archive was incorporated in 1996 by Bruce Gilliat and Brewster Kahle (Kimpton & Ubios 2006). The goal of the Internet Archive is to collect digital data in danger of disappearing. There has never been any way to completely enumerate all webpages; so, all attempts to archive the web are to some extent incomplete. The general approach is to use a web crawler, a software program that starts with a list of URLs to visit (a seed list) and downloads a copy of the content at each of these URLs. Each downloaded webpage is examined to find all the hyperlinks, which are then added to the list of URLs to be downloaded (subject to certain policies about how much content and what types of content to download). In this way, the software "crawls" from page to page following hyperlinks somewhat like snowball sampling. Despite its best efforts the Internet Archive cannot collect everything. This leads to the question: How much of the web is archived?

In order to answer this question, we looked at two different collections of webpages, one that had been collected and archived by the Internet Archive, and one that we collected ourselves. In this way, we are able to examine the completeness of the data that are held in the Internet Archive, at least with respect to our case study. To achieve this, we needed a case where we could reasonably find and download the full population of historical webpages. It is extremely difficult to find such a population since the Internet is constantly changing, and purposely collected archives are often the only source of historical webpages. We chose TripAdvisor as our case study as the website stores all reviews, including those written years ago, and thus allows us to reconstruct a historical population of webpages.

Our case study compares a full population of webpages from TripAdvisor with the subset stored by the Internet Archive. We defined our population as all tourist attractions in London listed on the TripAdvisor website. We downloaded these attractions from the current TripAdvisor site and found the earliest review of each attraction. We call this data the 'live data', and compare it to Internet Archive data. The specific data we compare to is a copy of all the Internet Archive data for all webpages in the .uk country-code top-level domain from 1996 to 2013 that was copied to the British Library from where we obtained them. We refer to this data as the 'archived data' and note that it is a 'subset' rather than a 'sample' of the web because the Internet Archive does not claim to select a probability sample.

While others have looked at archive coverage in terms of webpages (URLs) generally, notably Ainsworth, AlSum, SalahEldeen, Weigle, and Nelson (2013), this chapter is the first attempt to look at the extent of coverage of an individual website in-depth. The remainder of this chapter is organized as follows. We review the existing literature comparing archived coverage to the web. We describe the Internet Archive and the source of our data, followed by a discussion of TripAdvisor. We report our methodology and results and then discuss the implications of these results for research using web archival data.

Literature

Prior research on the success of web archiving is surprisingly sparse. Two studies, based on small subsets, address this issue. Thelwall & Vaughan (2004) studied differences in website coverage. They used randomly constructed names up to four letters long to find a total of 521 commercial websites related to four countries: the US, Taiwan, China and Singapore and found large differences across the countries. They found that the Internet

Archive had at least one page stored for 92% of the US commercial websites, but had at least one page stored for only 58% of the Chinese commercial websites. Russell and Kane (2008) looked at web citations in history journals. They attempted to retrieve, from the Internet Archive, those citations that were no longer available on live websites. Only 57% of the citations not available online were retrievable from the Internet Archive.

Both of these studies examined only small number of websites, and Russell and Kane's selection was not a random sample. The most complete study on the extent to which the web is archived is Ainsworth, AlSum, SalahEldeen, Weigle, and Nelson (2013).[1] They sampled 1,000 Uniform Resource Locators (URLs) each from the Open Directory Project (DMOZ), the recent URLs bookmarked on the social bookmarking site Delicious, randomly created hash values from Bitly, and the Google search engine index. They used the Memento API (Van de Sompel et al. 2009; Van de Sompel et al. 2010) to search 12 archives (including the Internet Archive) for each of the samples of 1,000 URLs and found that between 35% and 90% of the web was archived.

This is not a very satisfactory answer because it is such a wide range, but it broadly confirms the results from the smaller projects of Thelwall & Vaughan (2004) and Russell & Kane (2008). Large parts of the web are not included in any archive. A major weakness of these studies is a lack of detail about how much of each website has been archived. Thelwall & Vaughan (2004) counted a website as present in the archive as long as at least one page was archived. Ainsworth et al. (2013) and Russell & Kane (2008) looked at webpages (URLs) from many websites but also did not examine how much of each site was in the archive. We address this gap by analysing how much of a website has been archived and whether the archived pages in the website differ in a systematic way from the population of all pages on the website.

There is a large literature on the use of Internet Archive data; however, this literature is less helpful to scholars than it could be, because it largely discusses what authors think should be possible without reference to the reality of what actually is possible (e.g. Arms, Huttenlocher, Kleinberg, Macy and Strang 2006; Weber 2014). Our study uses a computational approach to assess what actually is possible to learn from Internet Archive data.

Case selection

We study London attractions found on the travel website TripAdvisor (TripAdvisor.co.uk). TripAdvisor, according to its own strapline, is the 'world's largest travel website'. TripAdvisor (2014) cites Google Analytics as showing that it receives 315 million unique visitors each month. This figure shows the extraordinary importance of TripAdvisor in the travel business. It is therefore not surprising that most academic research on TripAdvisor is found in the tourism literature and focuses on hotel reviews. Previous studies tend to focus on practical issues such as how users decide how to trust reviews, the response of hotels to reviews, or the content of negative reviews and complaints (Ayeh, Au and Law, 2013; Cunningham, et al., 2010; O'Connor, 2008; Sparks and Browning, 2010; Stringam and Gerdes, 2010). In contrast, our substantive interest, discussed elsewhere, is in how TripAdvisor works to convey cultural meanings. By studying reviews of cultural organisations, we examine the blurring of

[1] This paper is an updated version of Ainsworth, AlSum, SalahEldeen, Weigle, and Nelson (2011).

distinctions between high and popular culture and between commercial and non-profit venues (Alexander, Blank and Hale, in preparation).

TripAdvisor displays user-generated reviews across categories such as hotels, restaurants, and attractions. (Attractions encompass all elements of a city that are not restaurants or hotels.) Each review is comprised of a star rating, a title and a textual description. When starting a review, users enter a name of their hotel, restaurant or attraction, and if the target has been reviewed already, TripAdvisor suggests matches. Users can choose to review an item that already exists in TripAdvisor, or they can create an entry for a new, previously unreviewed one. For each review, users must choose a star rating, ranging from one star (negative) to five stars (positive). It is not possible for users to post reviews without choosing a star rating. Users then enter a short title or description in a free-form text box, and this serves as the title of their review, and then they write the review itself, which can be as short or as long as they wish. TripAdvisor ranks hotels and attractions within categories based on their reviews using a proprietary method and these rankings may have a profound effect on the livelihood of hoteliers (Scott and Orlikowski, 2012). From our perspective, however, a crucial benefit of the reviews is that they provide a simple star rating combined with a more nuanced textual description. The star ratings allow an explicit comparison across different types of data, in this case, the archived data and our own live data.

We limited our live data to TripAdvisor's user-generated reviews of London attractions on TripAdvisor's UK site (tripadvisor.co.uk). This offers us major advantages. London is a world-class metropolis with an enormous variety of attractions, providing us with a large range of reviews. Despite its size, however, London is still a bounded space so that our dataset can include the entire population of attractions and the entire population of reviews. Using TripAdvisor's UK site for London attractions makes it an appropriate vehicle for comparison to the archived data.[2]

The British Museum is currently the top attraction in London, and is described as '#1 of 1,277 things to do in London' (TripAdvisor, 2015). We have compiled a data set of these attractions, as detailed below. This allows us to compare across data sets (live data versus archived data) on easily measured variables, such as number of attractions and reviews, the average star rating for each attraction, and the dates of reviews. Table 1 lists example attractions in each of TripAdvisor's top-level categories.

[2] TripAdvisor operates a number of domain names (e.g., tripadvisor.com, tripadvisor.es, etc.) in over 30 countries; however, most of the content about specific attractions on these sites is the same.

Table 1: Categories of attractions on TripAdvisor in 2015

Category	Number of Attractions in Category[a]	Example attractions
Amusement Parks	3	The London Dungeon; Shrek's Adventure!
Boat Tours & Watersports	45	Canal and River Cruises Day Tours; Capital Pleasure Boats
Casinos & Gambling	17	Hippodrome Casino; Kempton Park Racecourse
Classes & Workshops	90	Hairy Goat Photography Tours; Bread Angels; East London Wine School
Food & Drink[b]	120	Eating London Food Tours; Spice Monkey Cookery School
Fun & Games	232	ClueQuest - The Live Escape Game; HintHunt; Secret Studio
Museums	280	Victoria and Albert Museum; National Gallery
Nature & Parks	129	St James's Park; Thames River;
Nightlife	1231	City of London Distillery; Comedy Store London; The Cavern Freehouse
Outdoor Activities	139	London Duck Tours; Moo Canoes Ltd.; Fishing London Coaching and Guide Service
Shopping	571	Covent Garden; Harrods
Sites & Landmarks	519	Houses of Parliament; Big Ben
Spas & Wellness	210	Pure Massage; The Body Retreat
Theatre & Concerts	292	Les Miserables; Brick Lane Music Hall
Tours & Activities	521	Alternative London Tours; BrakeAway Bike Tours; Shoreditch Street Art Tours
Transportation	67	London Tube; King's Cross Station
Traveller Resources	30	Barbican Centre; City of London Information Centre
Zoos & Aquariums	6	London Zoo

Source: Data on categories and number of subtopics is from the live data on TripAdvisor. The number of attractions per category and examples are drawn from TripAdvisor (2015).

[a] Attractions often appear in more than one category; so, the total adds to more than the number of attractions in the dataset.

[b] The Food and Drink category does not include restaurants, but does include food and drink available in other attractions, such as a museum café, cookery school, or food-related tour.

Data and methods

There are many technical issues to resolve in order to study webpages. We found all the London attraction pages on TripAdvisor had the form of "Attraction_Review-.*-London_England.html" where .* indicates any (or no) characters. We used the sitemap files published by tripadvisor.co.uk that list all webpages on the site to create a complete list of all the attractions in London available on TripAdvisor for the current, live site

and wrote a custom web crawler in Python3 to fetch the HTML of all the pages. Each attraction page had up to 10 user reviews on it. For attractions with more than 10 reviews, we downloaded all additional pages of reviews.

We crafted regular expressions to extract the elements of the attractions and user reviews in which we were interested. For attractions, we extracted the following elements:

- The name of the attraction
- The number of reviews for the attraction
- The average star rating of the attraction
- The category of the attraction as determined by TripAdvisor / its users
- The rank of the attraction among other attractions in London
- The total number of 5-star, 4-star, 3-star, 2-star, and 1-star reviews

We also extracted the date that each review was added to each attraction. We performed all data collection in July 2015. Our final live dataset therefore contains all London attractions listed on TripAdvisor at that time and all available reviews to these attractions.

TripAdvisor, like many websites, does not include all content in the HTML of each webpage, but loads some content separately using JavaScript. For TripAdvisor, the text of all user reviews is truncated in the HTML page and foreign-language reviews are not included at all. As the website still exists, we were able to emulate the JavaScript requests needed to collect the full text of reviews as well as foreign-language reviews for the live site but not for the archived data. Even so, within the live dataset, we were unable to collect 123 foreign-language reviews and hence our dataset contains 516,641 (99.98%) of the 516,764 reviews available in July 2015.

The Internet Archive is the oldest and biggest web archive, founded in 1996. A non-profit organization headquartered in San Francisco, it was created to preserve a historical copy of the World Wide Web. The UK Joint Information Systems Committee (JISC, now 'Jisc', a third-sector, charitable body) commissioned the Internet Archive to extract all stored webpages within the .uk domain from its archives. This data was stored in a new data centre at the British Library and forms the JISC UK Domain Dataset (UK Web Archive Open Data, n.d.). This Internet Archive data is the data we use within this chapter, and note that this data is the broadest data set of UK domains available for the time period we study (1996-2013). [3] In partnership with the British Library, we extracted all TripAdvisor webpages stored in the archive with URLs matching "Attraction_Review-.*-London_England.html". The data includes the HTML of the webpages as well as information about when the pages were added to the archive. We refer to this data simply as the archived data.

[3] In addition to the JISC UK Domain Dataset comprised entirely of Internet Archive data, the British Library has also independently collected web content related to the UK. Prior to 2014, the British Library manually selected important UK websites and crawled the websites whose owners could be contacted and gave permission to be included in the BL Web Archive. In 2014, the British Library started running its own crawls of the .uk domain, completely separate from the Internet Archive. We do not use any data that the BL crawled itself as the selective crawls did not include TripAdvisor and the 2014 crawl was not available at the time we extracted our data.

Results

Data overview

The earliest review in the live dataset was written on 26 August 2001, and the number of reviews on the site has been growing exponentially since that time (Figure 1, note that the vertical axis is a logarithmic scale).

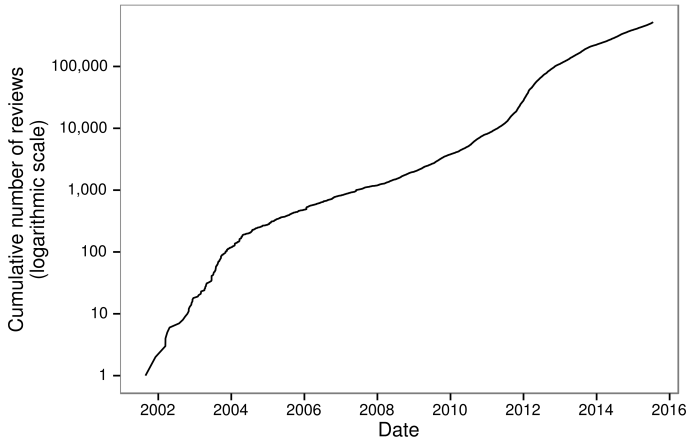

Figure 1: Cumulative number of reviews in the live dataset

TripAdvisor does not indicate when an attraction was first added to the website; so, we take the date of the earliest review as a proxy for this measure. Measuring growth in this way, we found that the number of attractions on the website has also been growing each year (Figure 2, again note the logarithmic scale on the vertical axis).

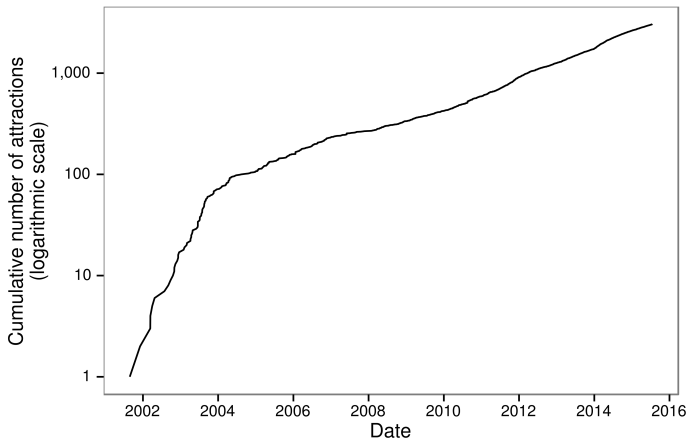

Figure 2: Cumulative number of attractions in the live dataset by first appearance. The date of earliest review is used as the date the attraction first appeared on the site.

The archived data contains 1,169 TripAdvisor webpages containing 340 unique attractions. The webpages of most attractions (57%) were only archived once, but some attractions were archived multiple times. The median number of copies was 1, the mean 3.4, and the maximum 31 (the most-archived attraction was "Alternative London Tours").

The most recent data in the archived dataset is from 1 May 2013. Using the live dataset and the date of the first review for each attraction as a proxy for when that attraction was added to TripAdvisor, we estimate there were at least 1,406 attractions listed on

the TripAdvisor website at that time. Thus, the 340 attractions covered in the archived dataset represent at most 24% of all the attractions available on the site at that time. This is the first indication of what proportion of the website is contained within the archived dataset. The top panel of Figure 3 shows the number new attractions added to the archived dataset each month based on the date that the webpage was crawled. The bottom panel of Figure 3 shows the number of new attractions added to the live website each month based on the date of the earliest review. Figure 4 shows the estimated proportion of attractions in the archived data compared the live dataset.

The actual percentage of attractions stored in the archived dataset is probably lower as the live dataset does not include attractions that were on TripAdvisor but later removed. This appears to apply to 37 attractions in the archived dataset that do not appear in the live dataset. This means that there are actually 303 attractions in both the archived data and the live data. In addition, our numbers do not include the 734 attractions in the live data (8 of these are in the archived data) with no reviews and hence no proxy for when they were added.

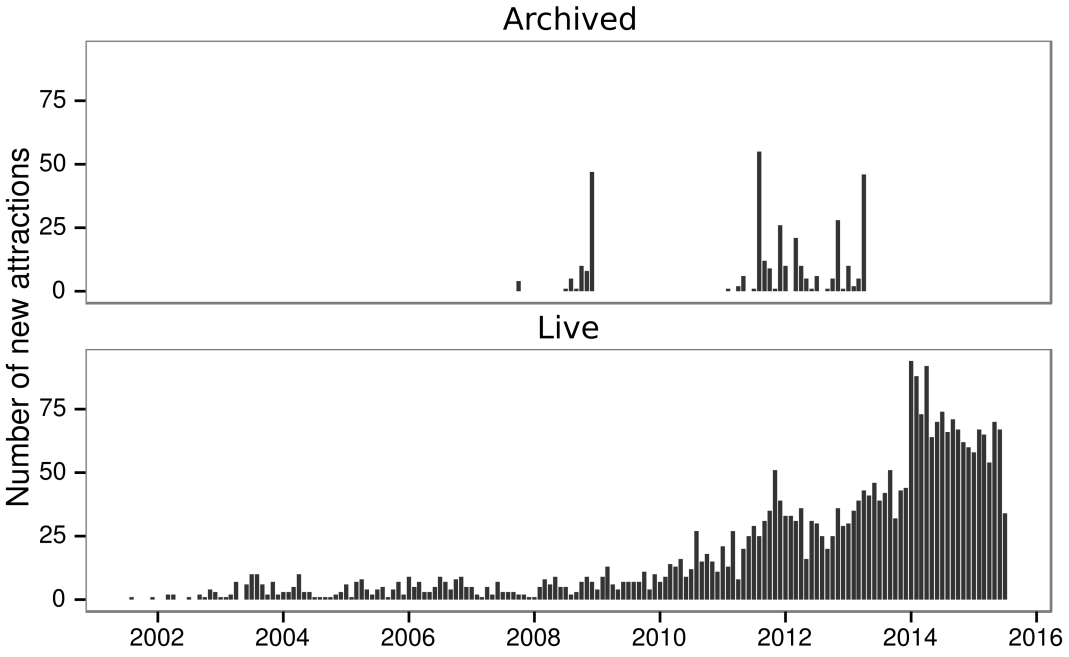

Figure 3: The number of new London attractions added each month to the TripAdvisor website based on archived data and live data. For the archived data, the date of a new attraction is the date that the webpage of the attraction was first crawled while for the live data the date of a new attraction is the date of the oldest review for that attraction.

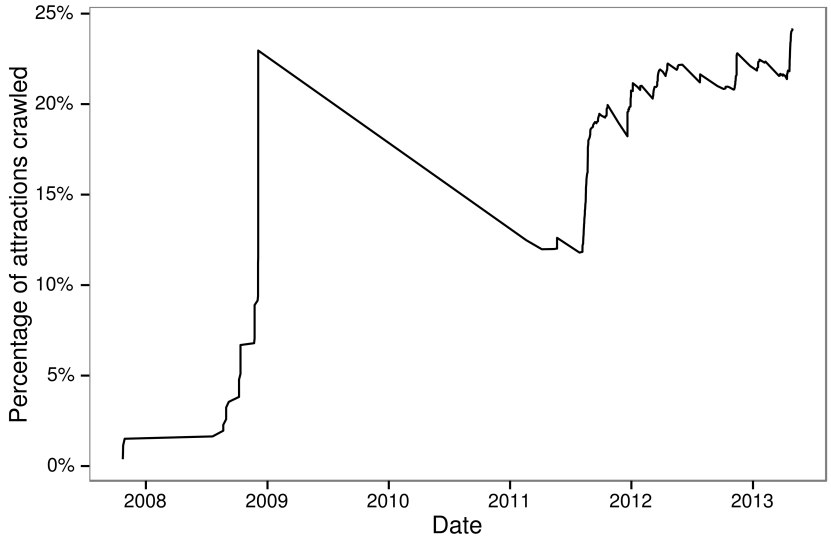

Figure 4: The proportion of attractions stored in the archived dataset increased irregularly to around 24% of all attractions on the TripAdvisor website from 2007 to 2013 even as the overall number of attractions on TripAdvisor continued to grow.

Comparing the two datasets

We proceed by comparing the 303 attractions in both the archived dataset and the live site with the 1,409 attractions known to be on the live site at the last date of a new page in the archived data. We find that the attractions in the archived dataset differ significantly and are not representative of those on the live site.

Attractions within the archived dataset have a considerably different distribution of reviews per attraction than attractions in the live dataset. We demonstrate these differences using two statistical techniques.[4] Figure 5 shows the distribution of the number of reviews per attraction using a kernel density (note that the horizontal axis uses a logarithmic scale). Since the live data represents the actual population, we use a 1-sample t-test, which shows that the mean number of reviews per attraction in the archived data differs significantly from the population mean ($t=5.7$, $p<0.001$, $N=303$). The distribution of the archive data is skewed to the right; it contains attractions with 928 more reviews on average, probably an indication that the archived data have a bias toward more visible and prominent webpages. Figure 6 (also a kernel density, but with linear scales) shows that attractions in the archived dataset have higher average star ratings compared to attractions in the live dataset: an indication that the archived data tend to be biased toward more popular attraction. This difference is confirmed by a one-sample t-test ($t=3.2$, $p=0.002$, $N=303$). Finally, Figure 7 (also a kernel density with linear scales) shows that attractions in the archived dataset tend to have a similar distribution of ranks. A 1-sample t-test shows that the mean rank of attractions in the archived data does not differ significantly from the mean of the population, the live data

[4] We have used a technique called kernel density estimation with a Gaussian kernel to estimate the distributions of the two datasets. We also use a standard hypothesis-testing technique, a one-sample t-test, to compare the mean of a sample to a known population mean in order to assess the probability that the sample (the archived data) was drawn from the population (the live data).

(t=-1.2, p=0.22, N=303). The fact that one of the three measures of bias did not show a statistically significant difference is noteworthy; however, rankings are probably the least useful indicator because TripAdvisor reports attraction rankings within a number of different subcategories and the particular ranking criteria are not public.

Finally, in Table 2 we examine the percentage of attractions in each dataset in each of the 18 top-level categories on the current TripAdvisor website. Museums are most overrepresented in the archived dataset, nine percentage points higher than in the live data. The archived data also include an excessive number of Tours and Activities (6.6%). Nightlife is the most underrepresented, 6.9% less in the archived data compared to the live data. If a researcher were interested in using the archived data as a proxy for attractions, these deviations could certainly cause biased results.

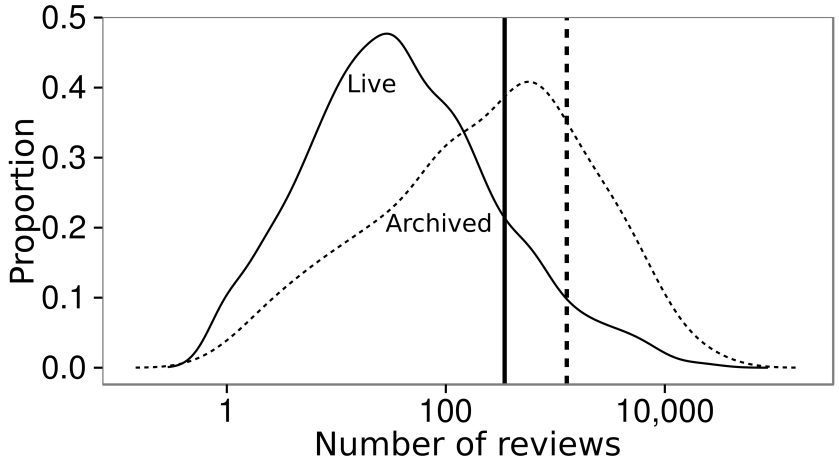

Figure 5: Distribution of reviews per attraction in the live dataset and the archived data. Vertical lines are means. Note that horizontal access uses a logarithmic scale.

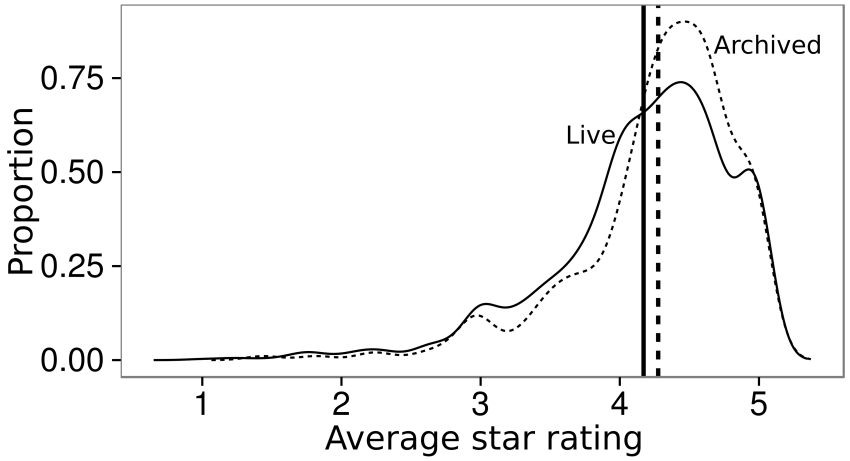

Figure 6. Distribution of star ratings in live dataset and the archived data. Vertical lines are means.

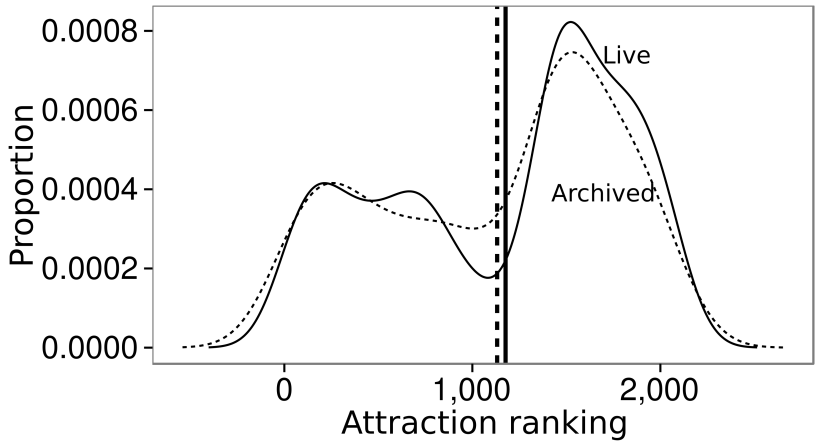

Figure 7. Distribution of attraction rankings in the live dataset and the archived data. Vertical lines are means.

Table 2.2: Percentages in each attraction category in the live data and archived data

Category	Live data	Archived data	Difference
Amusement Parks	0.1	0.4	0.3
Boat Tours & Water Sports	1.5	2.3	0.8
Casinos & Gambling	0.5	0.8	0.3
Classes & Workshops	1.9	1.9	0.0
Food & Drink	1.4	1.2	-0.3
Fun & Games	5.8	5.0	-0.8
Museums	11.8	20.8	9.0
Nature & Parks	5.6	5.8	0.2
Nightlife	18.1	11.2	-6.9
Outdoor Activities	3.6	5.8	2.1
Shopping	15.3	12.3	-3.0
Sights & Landmarks	22.0	24.2	2.2
Spas & Wellness	4.0	0.8	-3.2
Theatre & Concerts	11.2	12.7	1.5
Tours & Activities	15.7	22.3	6.6
Transportation	0.7	1.9	1.2
Traveller Resources	1.3	1.2	-0.1
Zoos & Aquariums	0.3	1.2	0.9

Note: The percentages in the live data and the archived data add to more than 100% because some attractions are categorized in more than one category

Discussion

Much has been promised for the use of web archives, and there have been a number of studies. For example Chu, Leung, Van Hui and Cheung (2007) tracked the longitudinal development of site content on e-commerce websites. Mike Thelwall with various colleagues (Thelwall and Wilkinson, 2003; Vaughn and Thelwall, 2003; Payne and Thelwall, 2007) used web data to demonstrate the interdependence of academic institutions on the web. Hackett and Parmanto (2005) used the Internet Archive's Wayback Machine to analyse how technological advances were manifest in changes in website design over time. Hale, Yasseri, Cowls, Meyer, Schroeder, and Margetts (2014) studied the evolution of the presence of British universities on the web using the same .uk webarchive dataset that we used here.

The work with web archives has not been as extensive as the original founders anticipated, because, at least in part, there remain major challenges to using web archives. Scholars using the biggest archive, the Internet Archive, are mining data from an 18-petabyte dataset as of August 2014 (Internet Archive, 2015). Confronted with this enormous amount of data, few tools exist to help scholars find information. Furthermore, webpages are not well-structured or consistently structured, and they can be extremely difficult to transform into a format that can be used for large-scale quantitative research. In addition, changes in webpage format and changes in content often occur simultaneously. This complicates longitudinal research because just getting the data into a consistent format may be difficult and slow. It may not be something that many scholars will want to invest in, given their need to publish.

Once the data have been put in a consistent format what, exactly, do researchers have? This is the question we have addressed. First, researchers using web archive data have a subset of the full web. Using Ainsworth et al.'s (2013) estimates of webpages they might have between 35% and 90% of the web. By constructing their sample of URLs from DMOZ, Delicious, Bitly, and Google, Ainsworth et al. (2013) almost certainly examined the inclusion of more popular and prominent URLs (i.e., the URLs included in DMOZ or added to Delicious are by definition more popular and prominent than the URLs that no one adds to these platforms). We have avoided this bias by comparing to the entire population of London attraction webpages on TripAdvisor. Although TripAdvisor is a prominent website, we still found only 24% of the webpages about London attractions were archived.

This suggests that previous results are dramatic overestimates of the amount of the web that has been stored in archives. Our findings also complement the results from previous studies that have examined the percentage of web content included in web archives (e.g., Thelwall & Vaughan 2004). Whereas these studies looked at the inclusion of at least one page of a website in the archive, we looked deeper into the site itself at whether webpages within the site are stored. Even though the TripAdvisor site itself is included in our archived data, only at most 24% of the pages about London attractions have been stored. It may also suggest that there is enormous variation in the archival coverage, and the simple presence of one webpage from a website in the archive does not provide an indication of how much of that website is actually within the archive.

We also found that the archived pages do not resemble a random probability sample. There is a clear bias toward prominent, well-known, and highly-rated webpages. Smaller, less well-known and lower-rated webpages are less likely to be archived. It is worth noting that all the archived data we used came from the Internet Archive; so, the archived data are probably the best, most complete source possible for this time period but it is clearly not complete, and it contains significant biases. In 2014, the BL began conducting its own crawls of UK websites, but the representativeness and completeness of this data is yet to be determined.

What are the implications of these results for research using web archives? Much of the appeal of the Internet is that it seems to provide broader data than conventional sources. Advocates talk about it being unrestricted in scale or geographic scope. One reason web archives were seen as valuable was because they promised to provide full historical data on things like diffusion of innovations, community formation, emergence of issues, or the formation and dynamics of networks (Arms et al. 2006). The Internet is certainly broader than most conventional data sources, but the web archive we examined is broader in a certain way. It focuses on the big and the prominent. Due to the limits on

the number of pages found and crawled from any one website, web archives are necessarily incomplete even when they start with a seed list of all domain names (as is now the case for the British Library crawls of the .uk country-code top-level domain). In some instances the limits on the pages for each website are relatively high—as in the case of national web archive in Denmark (see Brügger in press)—but it remains difficult to assess what content is not archived (as archiving strategies change overtime and technical issues in capturing dynamic/JavaScript content arise). So, a web archive-based study of diffusion of innovation on the Internet would actually be a study of diffusion among prominent, highly-rated webpages, not among all webpages. A study of network formation or network dynamics would be a study of networks of well-known, highly-rated webpages. It would not be a study of diffusion among all webpages. Hale et al.'s (2014) study of British university websites, for instance, is a study biased toward hyperlinks on more prominent webpages.

The incomplete nature of web archives limits the type of analyses available to researchers. We were only able to conduct our analysis, for instance, at the level of attractions in London and not about the content of reviews because the archived data is so incomplete with reference to review text that it did not make sense to even attempt such a comparison. These problems are only getting worse as content moves off the Web to other channels (e.g., mobile apps), personalization means there is no definitive version, and dynamic sites use JavaScript or other technologies to fetch content separately from the HTML page.

The promise raised by Arms et al. (2006) was that web archives would eliminate the need to proactively collect data for longitudinal studies of networks, innovations, community formation, etc., and instead allow for fine-grained, retrospective analyses over longer periods of time. Web archive data can certainly provide insights that would otherwise be unavailable (e.g., we were able to find attractions that had been deleted from TripAdvisor in the archive that were unavailable on the live site) and with suitable modelling, networks of hyperlinks from web archive data may be compared to null model controls. However, our study highlights that web archive data does not replace the need to collect specific data proactively over set periods of time for many types of longitudinal analysis, and the level of incompleteness of web archive data raises questions about the extent to which archived web data can be used to conduct longitudinal research at all. An approach that would yield much higher quality data is the same as we might have used for pre-Internet longitudinal data. That is, collect repeated cross-sectional datasets proactively in real time and then do retrospective, time-series analyses of the data only at the end of the study period. The irony is striking, but the point is that web archives do not provide any free lunch to good research.

These are serious problems. Web archives are an extensive and permanent record, but they are also an incomplete and biased record. While it is certainly possible to analyse larger numbers of many things, are large, biased numbers a good idea? The answer is that a biased set of data remains biased no matter how many cases it contains and biased datasets provide biased answers regardless of their sizes. So researchers have to confront the bias problem. Web archives do not contain a complete population, except perhaps in certain limited areas, and what is missing from the archives is often unknown.

References

Ainsworth, S., AlSum, A., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2013). How much of the Web is archived? Technical Report arXiv:1212.6177v2.

Ainsworth, S., AlSum, A., SalahEldeen, H., Weigle, M.C., & Nelson, M.L., (2011). How much of the web is archived? *JCDL 2011*, ACM Press, Ottawa, Canada, 133–136

Alexander, V. D., Blank, G. & Hale, S. A. (2015). Hive Mind meets Distinction: Using Digital Trace Data to Examine Cultural Capital in TripAdvisor Reviews of London Cultural Attractions. Oxford Internet Institute, Working Paper.

Arms, W. Huttenlocher, D., Kleinberg, J., Macy, M., & Strang, D. (2006). From Wayback Machine to Yesternet: New opportunities for social science. In *Proceedings of the 2nd International Conference on e-Social Science*.

Ayeh, J. K., Au, N. & Law, R. (2013). Do we believe in TripAdvisor? Examining credibility perceptions and online travelers' attitude toward using user-generated content, *Journal of Travel Research*, 52(4) 437–452.

Brügger, N. (in press) Probing a nation's web domain A new approach to web history and a new kind of historical source. In *Routledge Handbook of Internet Histories*, edited by G. Goggin and M. McLelland.

Chu, S.C., Leung, L.C., Hui, Y .V ., & Cheung, W., 2007. Evolution of e-commerce Web sites: A conceptual framework and a longitudinal study. *Information and Management*, 44(2), 154–164.

Cunningham, P., Smyth, B., Wu, G., & Greene, D. (2010). Does TripAdvisor makes hotels better? Technical Report, UCD-CSI-2010-06, School of Computer Science & Informatics, University College Dublin.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2004). A large-scale study of the evolution of web pages. *Software-Practice and Experience*, 34, 213–237.

Hackett, S. & Parmanto, B., (2005). A longitudinal evaluation of accessibility: Higher education web sites. *Internet Research,* 15(3), 281–294.

Hale, S. A., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK webspace: Fifteen years of British universities on the Web. In *Proceedings of the 2014 ACM conference on Web science (WebSci '14)*. ACM, New York, USA, 62–70. http://dx.doi.org/10.1145/2615569.2615691

Internet Archive. (2015). Internet Archive: Petabox. Retrieved from https://archive.org/web/petabox.php.

O'Connor, P. (2008). User-generated content and travel: A case study on Tripadvisor.com. *Information and Communication Technologies in Tourism*, 47–58.

Payne, N. & Thelwall, M., (2007). A longitudinal study of academic webs: Growth and stabilization. *Scientometrics*, 71(3), 523–539.

Russell, E. & Kane, J. (2008). The missing link: Assessing the reliability of Internet citations in history journals. *Technology and Culture*, 49(2), 420–429.

Scott, S. V. & Orlikowski, W. J. (2012). Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, Organizations and Society*, 37, 26–40.

Sparks, B. A. & Browning, V.(2010). Complaining in cyberspace: The motives and forms of hotel guests' complaints online. *Journal of Hospitality Marketing & Management*, 19(7), 797–818.

Stringam, B. B. & Gerdes, J. Jr (2010). An analysis of word-of-mouse ratings and guest comments of online hotel distribution sites. *Journal of Hospitality Marketing & Management*, 19(7), 773–796.

Thelwall, M. & Vaughan, L. (2004). A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2), 162–176.

Thelwall, M. & Wilkinson, D. (2003). Three target document range metrics for university websites. *Journal of the American Society for Information Science and Technology*, 54(1), 29–38.

TripAdvisor (2015). Top Things to Do in London. Retrieved from http://www.tripadvisor.co.uk/Attractions-g186338-Activities-London_England.html.

TripAdvisor (2014). About TripAdvisor. Retrieved from http://www.tripadvisor.co.uk/PressCenter-c6-About_Us.html.

UK Web Archive Open Data (n.d). JISC UK Web Domain Dataset (1996–2013). Retrieved from http://data.webarchive.org.uk/opendata/ukwa.ds.2/.

Weinreich, H., Obendorf, H., Herder, E., & Mayer, M. (2008). Not quite the average: An empirical study of Web use. *ACM Transactions on the Web*, 2(1), 1–31. http://doi.acm.org/10.1145/1326561.1326566

Van de Sompel, H., Nelson, M. L., Sanderson, R., & Balakireva, L. L., Ainsworth, S., & Shankar, H. (2009). Memento: Time travel for the Web. Technical Report, arXiv:0911.1112.

Van de Sompel, H., Sanderson, R., Nelson, M., Balakireva, L., Shankar, H., & Ainsworth, S.. (2010). An HTTP-based versioning mechanism for linked data. In *Proceedings of Linked Data on the Web Workshop (LDOW2010)*. Retrieved from http://events.linkeddata.org/ldow2010/papers/ldow2010_paper13.pdf.

Vaughn, L. & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology*, 54(1) 29–38.

Weber, M. (2014). Observing the Web by understanding the past: Archival Internet research. In *Proceedings of the 14th International World Wide Web Conference (WWW'14 Companion)*. Seoul, Korea, 1031–1036. http://dx.doi.org/10.1145/2567948.2579213