we.b: The web of short URLs

Demetris Antoniades
FORTH-ICS
danton@ics.forth.gr

Iasonas Polakis
FORTH-ICS
polakis@ics.forth.gr

Georgios Kontaxis
FORTH-ICS
kondax@ics.forth.gr

Elias Athanasopoulos
FORTH-ICS
elathan@ics.forth.gr

Sotiris Ioannidis
FORTH-ICS
sotiris@ics.forth.gr

Evangelos P. Markatos
FORTH-ICS
markatos@ics.forth.gr

Thomas Karagiannis
Microsoft Research
thomkar@microsoft.com

ABSTRACT

Short URLs have become ubiquitous. Especially popular within social networking services, short URLs have seen a significant increase in their usage over the past years, mostly due to Twitter's restriction of message length to 140 characters. In this paper, we provide a first characterization on the usage of short URLs. Specifically, our goal is to examine the content short URLs point to, how they are published, their popularity and activity over time, as well as their potential impact on the performance of the web.

Our study is based on traces of short URLs as seen from two different perspectives: i) collected through a large-scale crawl of URL shortening services, and ii) collected by crawling Twitter messages. The former provides a general characterization on the usage of short URLs, while the latter provides a more focused view on how certain communities use shortening services. Our analysis highlights that domain and website popularity, as seen from short URLs, significantly differs from the distributions provided by well publicised services such as Alexa. The set of most popular websites pointed to by short URLs appears stable over time, despite the fact that short URLs have a limited high popularity lifetime. Surprisingly short URLs are not ephemeral, as a significant fraction, roughly 50%, appears active for more than three months. Overall, our study emphasizes the fact that short URLs reflect an "alternative" web and, hence, provide an additional view on web usage and content consumption complementing traditional measurement sources. Furthermore, our study reveals the need for alternative shortening architectures that will eliminate the non-negligible performance penalty imposed by today's shortening services.

Categories and Subject Descriptors

C.2.0 [**Computer Communication Networks**]: General; H.3.5 [**Information Storage and Retrieval**]: Online Information Services,Web based services

General Terms

Measurement, Performance

Keywords

Short URLs, Twitter, Online Social Networks

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

1. INTRODUCTION

URL shortening has evolved into one of the main practices for the easy dissemination and sharing of URLs. URL shortening services provide their users with a smaller equivalent of any provided long URL, and redirect subsequent visitors to the intended source. Although the first notable URL shortening service, namely tinyURL [3], dates back to 2002, today, users can choose from a a wide selection of such services. [1] The recent popularity of shortening services is a result of their extensive usage in Online Social Networks (OSNs). Services, like Twitter, impose an upper limit on the length of posted messages, and thus URL shortening is typical for the propagation of content. While short URL accesses represent a small fraction of the *"web hits"* a site receives, they are rapidly increasing by as much as 10% per month according to Alexa [1].

Despite this rapid growth, there is, to the best of our knowledge, no other large-scale study in the literature that sheds light onto the characteristics and usage patterns of short URLs. We feel that understanding their usage has become important for several reasons, including: i) Short URLs are widely used in specialized communities and services such as Twitter, as well as in several Online Social Networks and Instant Messaging (IM) systems. A study of URL shortening services will provide insight into the interests of such communities as well as a better understanding of their characteristics compared to the broader web browsing community. ii) Some URL shortening services, such as bit.ly have grown so much in popularity, that they now account for as much as one percent of the total web population per day [1]. If this trend continues, URL shortening services will become part of the web's critical infrastructure, posing challenging questions regarding its performance, scalability, and reliability. We believe that answering these questions and defining the proper architectures for URL shortening services without understanding their access patterns is not feasible.

To understand the nature and impact of URL shortening services, we perform the first large-scale crawl of URL shortening services and analyze the use of short URLs across different applications. Our study is based on traces of short URLs as seen from two different perspectives: i) collected through a large-scale crawl of URL shortening services, and ii) collected by crawling Twitter messages. The first trace provides insights for a general characterization on the usage of short URLs. The second trace moves our focus onto how certain communities use shortening services. The highlights of our work can be summarized as follows:

[1] http://www.prlog.org/10879994-just-how-many-url-shorteners-are-there-anyway.html

- We study the applications that use short URLs and show that most accesses to short URLs come from IM Systems, email clients and OSN media/applications, suggesting a "word of mouth" URL distribution. This distribution implies that short URLs appear mostly in ephemeral media, with profound effects on their popularity, lifetime, and access patterns.
- We show that the short URL click distribution can be closely approximated by a log-normal curve, verifying the rule that a small number of URLs have a very large number of accesses, while the majority of short URLs has very limited accesses.
- We study the access frequency of short URLs and observe that a large percentage of short URLs are not ephemeral. 50% of short URLs live for more than three months. Further, we observe high burstiness in the access of short URLs over time. short URLs become popular extremely fast suggesting a "twitter effect", which may create significant traffic surges and may pose interesting design challenges for web sites.
- We show that the most popular web sites (as seen by the number of short URLs accesses towards them) changes slowly over time, while having a strong component of web sites which remains stable throughout the examined period. Our experiments also suggest that the web sites which are popular in the short URL community differ profoundly from the sites which are popular among the broader web community.
- We examine the performance implications of the use of short URLs. We find that in more than 90% of the cases, the resulting short URL reduce the amount of bytes needed for the URL by 95%. This result suggests that URL shortening services are extremely effective in space gaining. On the other hand, we observe that the imposed redirection of URL shortening services increases the web page access times by an additional 54% relative overhead. This result should be taken into consideration for the design of future URL shortening services.

2. URL SHORTENING SERVICES

The idea behind URL shortening services is to assist in the easy sharing of URLs by providing a short equivalent. For example, if the user submits `http://www.this.is.a.long.url.com/indeed.html` to bit.ly, the service will return the following short URL to the user: `http://bit.ly/dv82ka`. The user can then publish the short URL on any webpage, blog, forum or OSN, exactly as she would use the original URL. Any future access to `http://bit.ly/dv82ka` will be redirected by bit.ly to the original URL through an "HTTP 301 Moved Permanently" response.

URL shortening services have existed at least as early as 2001 [2]; tinyURL [3] is probably the first such, well-known, service. The rapid adoption of OSNs, and their imposed character limit for status updates, tweets and comments, has led to an increased demand for short URLs. As a consequence, dozens of such services exist today, although only a handful of them, such as bit.ly, ow.ly and tinyURL, capture the lion's share of the market. Aside from the aforementioned services, short URLs are also useful in more traditional systems which either discourage the use of very long words, such as IMs and SMSes, or do not handle long URLs very well, such as some email clients.

Besides providing a short URL for each long one, some of these services provide statistics about the accesses of these URLs. For example, bit.ly provides information about the number of hits each short URL has received (total and daily), the referrer sites the hits came from and the visitors' countries. For each unique long URL that it has shortened, bit.ly provides a unique global hash, along with an information page which provides the overall statistics for the URL. If a registered user creates a short URL for the same long URL, the service will create a different hash that will be given to the user so as to share it as she likes. The information page for this custom hash will contain statistics solely for the hits received by the creator's URL. Nonetheless, overall statistics will still be kept by the global URL's information page. Registered users can create as many custom short URLs as they like for the same long URL.

3. DATA COLLECTION

This section introduces our data collection process and gives a description of the collected data. Overall, we study short URLs from two different perspectives: i) By looking at two shortening services, namely bit.ly and ow.ly, and ii) by examining short URLs and their usage within OSNs, and, in particular Twitter.

3.1 Collection Methodology

We use two approaches to collect short URLs: i) *Crawling*, in which we search Twitter to find tweets which contain URLs and ii) *Brute-Force*, in which we crawl two URL shortening services, that is bit.ly and ow.ly, by creating hashes of different sizes and examining which of them already exist.

As mentioned in the previous section, bit.ly maintains an information page for each created short URL. This page provides detailed analysis regarding the amount of hits a short URL received, its HTTP referrers and the geographical locations of its visitors. The daily amount of hits since the creation of the short URL is also recorded. Information regarding the number of hits from each referrer and country is provided as well. For each bit.ly short URL in our traces we also collect the accompanied information pages. Information pages for short URLs created by registered users also contain a reference to the global short URL for this long URL. For the sake of completeness, our analysis includes the information provided by the global hash. Unfortunately, ow.ly does not provide any such information.

Twitter Crawling: Using the first method, we search for HTTP URLs that were posted on Twitter. Using the Twitter search functionality [6], we collect tweets that contain *HTTP* URLs. Twitter imposes rate limiting in the number of search requests per hour from a given IP address [5]. To respect this policy we limit our crawler to one search request every 5 minutes. Every search request retrieves up to 1500 results (tweets), going no more than 7 days (max) back in time. During our collection period we managed to collect more than 20 million tweets containing HTTP URLs. Only a small fraction of the HTTP URLs (13%) collected was not shortened by any URL shortening service. Among the HTTP URLs collected from Twitter, 50% were bit.ly URLs. The second most popular shortening service was tl.gd with 4%, while tinyURL corresponded to 3.5% and ow.ly amounted to 1.5% of the overall URLs. Hence, part of our analysis focuses on bit.ly URLs.

Brute-Force: Using the second method, we exhaustively search the available keyspace for ow.ly and bit.ly hashes. While the Twitter crawling approach returns links recently "gossiped" in a social network, this approach acts as an alternate source of collection, providing hashes irrespective of their published medium and recency.

In the bit.ly case, we searched the entire keyspace [0-9a-zA-Z] for hashes of up to 3 characters in length. Currently, the shortening service returns 6-character hashes, indicating a significant exhaustion of shorter combinations. In the case of ow.ly, the system does not disseminate random hashes of the user's long URL, but serially iterates over the available short URL space; thus, if the same long URL is submitted multiple times, it will result in multiple different hashes. Considering this deterministic registration mechanism, we collected the full set of short URLs created for a period of 9 days. During that time, we monitored the evolution of the keyspace by

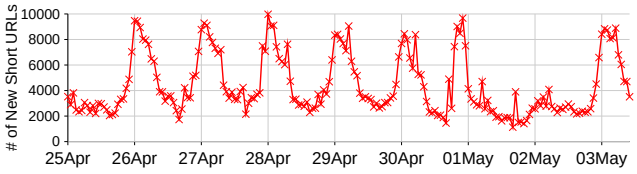

Figure 1: Number of ow.ly short URLs created as a function of time.

creating a new short URL of our own every hour and measuring the distance from the one we had created the previous hour. Using this heuristic, we were able to determine which and how many short URLs were created during that timeframe with a granularity of one hour. Figure 1 shows the number of ow.ly URLs registered as a function of time. As expected, we observe a clear diurnal and weekly cycle, with about 70,000 new short URLs created each day.

Having collected sets of bit.ly short URLs with the aforementioned methods, we proceed with the gathering and analysis of the metadata provided by the shortening service. Initially, we access the corresponding information page and record the resulting long URL, the total number of hits it has received, the name of the user that created it and the global short URL, offering aggregated data. We go on to collect the daily history of hit events for the entirety of the short URL's lifespan. Furthermore we fetch the number of hits per referrer and country. Finally, we follow the global short URL and download the aggregated versions of the metadata as well.

3.2 Collected Data

The previously discussed collection process resulted in four datasets:

- *twitter*: The trace contains 887,395 unique bit.ly short URLs posted on Twitter between the 22nd of April and the 3rd of May 2010. For each short URL, all the accompanied metadata are also collected.
- *twitter2*: The trace contains over 7M unique bit.ly short URLs posted on Twitter between the 6th of May and the 2nd of August 2010. In this trace we limit our metadata gathering to only the total and daily accesses for each short URL.
- *owly*: This trace contains 674,239 ow.ly short URLs created between the 26th of April and the 3rd of May 2010. As described in the brute-force methodology, this constitutes the entire population of ow.ly short URLs created in that period.
- *bitly*: Contains 171,044 unique bit.ly short URLs collected by exhaustively searching the available key space for hash sizes of 1 to 3 characters. All the accompanied metadata for each short URL are also collected.

Table 1 summarizes the data collected.

3.3 Representativeness

Before proceeding with the analysis of the collected data, we first examine the representativeness of these traces. To provide an estimation on the ratio of tweets that contain bit.ly URLs, we retrieved the total number of tweets, for a specific time window, using the public timeline feature of the Twitter API. For the same time window, we also collected the total number of tweets containing bit.ly, through the live search feature of Twitter. We examined both quantities for 144 10-minute windows, for the total period of 1 day. On average, we observed that 4.9% of all posted tweets contained bit.ly short URLs. With our relaxed crawling methodology we managed to retrieve about 7% of all new tweets containing one or more bit.ly short URLs. To estimate the benefit of a more aggressive crawling methodology, we used a second crawler, deployed only for the limited time period of a single day, issuing a search request every thirty seconds. The aggressive approach was able to harvest almost four times more tweets than the moderated one.

As discussed in Section 4, our findings remain the same when comparing statistics across the two crawling rates. The only observable difference is that, as expected, a more aggressive rate results in the collection of a larger number of less popular short URLs, i.e., short URLs that received one or two hits. Taking into consideration the ethical aspects of web crawling and considering that our tweet sampling ratio was large enough to allow the extraction of valid characteristics and behaviors, we followed the relaxed collection rate for the results presented throughout the paper.

4. THE WEB OF SHORT URLS

We begin our analysis with a general characterization of short URLs. Over the following sections, we identify where short URLs originate from, the type of content they point to, and analyze their popularity patterns.

4.1 Where do short URLs come from?

Despite the fact that short URLs are typically seen within OSN services, URL shortening services have already existed for a number of years. Thus, a natural question to ask is whether there are particular communities of users or applications where the usage of short URLs is dominant.

To this end, we study the "referrers" of each short URL, information that is provided by bit.ly for each short URL. Table 2 lists the top-5 most popular referrers for the URLs in traces *twitter* and *bitly*. We see that in both cases the vast majority of users (that is, 60% and 72% respectively) arrive at bit.ly from non-web applications; these include Instant Messaging and email clients, mobile applications like Twitterific and BlackBerry mail, Twitter desktop applications and directly (by pasting/typing the URL in a browser). For those users that do access short URLs through web applications, we observe that they mostly come from Twitter, and various other social-networking-related sites. This suggests that bit.ly (and possibly other URL shortening services) are most popular in social networking applications/communities.

The distribution of referrers in Table 2 reveals an entirely new browsing model for short URLs users. According to our findings, short URLs do not frequently appear in traditional web pages but are distributed via Instant Messaging (email,IM,phone) and social network channels (twitter.com, facebook.com), suggesting a "word of mouth" type of propagation. This has significant impact on the browsing habits and patterns of short URL users as we show in the following sections.

4.2 Where do short URLs point to?

Having observed that short URLs mostly originate in non-browser type of applications, we now aim at understanding the type of web pages that are popular through bit.ly links. To achieve this, we manually classified the content of the 100 most accessed domains in the *twitter* trace. Similarly, we classified the links of the *owly* trace, which was obtained via the Brute-Force method and presents a perhaps more general view of the content served through short URLs. In the case of ow.ly, the number of accesses per short URL is not available so we selected the most popular domains based on the number of shortened URLs under each domain.

Table 3 presents the top categories for each case. One may notice that news and informative content come first. This observation corroborates the finding of Kwak et al. [18], which suggested that Twitter acts more as a information-relaying network rather than as a social networking site. However, while this study suggests that

trace name	service	number of URLs	accesses	first URL access	last URL access
twitter	bit.ly	887,395	101,739,341	2008-07-08	2010-04-29
twitter2	bit.ly	7,401,026	2,202,442,600	2008-06-27	2010-09-25
owly	ow.ly	674,239	not available	2010-04-26	2010-05-03
bitly	bit.ly	171,044	15,096,722	2008-07-07	2010-05-06

Table 1: Summary of data collected

Rank	***twitter***		***bitly***	
	Site	% of Accesses	Site	% of Accesses
1	eMail,IM,apps,phone,direct	59.32	email,IM,apps,phone,direct	72.72
2	twitter.com	23.49	twitter.com	11.77
3	partners.bit.ly	3.02	www.cholotube.com	2.16
4	www.facebook.com	2.17	www.facebook.com	1.72
5	healthinsuranceexchange.info	1.57	partners.bit.ly	1.63

Table 2: The 5 most prolific Referrers of short URLs.

twitter		***owly***	
Category	% Sites	Category	% Sites
news (inc. portals)	25	news (inc. portals)	51
info / edu	18	various	17
various	13	info / edu	10
entertainment	10	social networking	5
personal	9	media sharing	5
twitter-related	9	shorten urls	4
commercial	6	commercial	4
media sharing	4	twitter-related	2
social networking	4	sharing articles	1

Table 3: Most popular types of content.

Rank	***twitter***		***bitly***	
	Site	% of Accesses	Site	% of Accesses
1	US	42.12	US	54.15
2	JP	12.20	GB	5.59
3	None	8.95	None	4.83
4	GB	5.96	CA	4.14
5	CA	4.58	PE	3.48

Table 4: The 5 Countries with the largest number of clicks.

trending topics are related to news by as much as 85%, the fraction of news related short URLs is significantly lower in our case (25% and 51% for the two traces). A surprising finding is that 4 of the most accessed URLs in the *owly* trace were shortening services. Such cases reflect short URLs packed inside other short URLs to avoid exposure of the long URLs from tools that unwrap the first level of redirection. Spammers use such techniques to avoid detection, as mentioned by Grier et al. in [16]. Manualy examining a number of these URLs confirmed this suspission with a large number of short URLs pointing to spam content. We plan further inverstication of this phenomenon as future work.

4.3 Location

We now examine the geographic coverage of short URL usage, i.e., whether short URL users follow the distribution of Internet/web users or whether short URLs are a niche application of some particular countries. Table 4 shows the distribution of the country of origin of short URL accesses in the *twitter* and *bitly* traces. Most of these accesses come from the United States, Japan, and Great Britain. Interestingly enough we do not see any accesses from China and India, which are ranked in the top-5 countries with the largest number of Internet users [8]. Our conjecture is that applications which use short URLs are probably not popular or widespread in the above countries, suggesting that the penetration of short URL use is significantly different from the Internet/web one.

4.4 Popularity

As discussed in Section 4.2, short URLs primarily refer to news and other information related content. In this section, we examine the particular domains visited through short URLs, and their popularity over time. First, however, we examine the popularity distribution of individual URLs. Popularity is measured by examining the number of hits a URL received.

URL Popularity: Large systems that provide content to users typically exhibit a power-law behavior [9, 23] with respect to the offered content (e.g., [11]). That is, a small fraction of the content is very popular, while most of it is considered uninteresting, characterized by moderated access rates. Figure 2 (top) depicts the popularity distribution of the short URLs in the *twitter* and *twitter2* trace, and the corresponding Cumulative Distribution Function (CDF) –bottom. As is the case with other content provider services, the distribution has a heavy tail.

Figure 2 also plots the popularity distribution and corresponding CDF for the short URLs collected through the aggressive harvesting, pressented in Section 3.3. As we observe the sampling rate we employ on the Twitter crawling method does not bias our findings. The only observable difference is that, as expected, more aggressive sampling result's in the collection of a large number of less popular short URLs, i.e., short URLs that received one or two hits.

Since our trace might be populated with recently created URLs, the distribution may be biased. To examine this hypothesis, we eliminate all short URLs whose creation was during the last week of our trace collection period. Further, we split short URLs into *active* and *inactive*. As "inactive", we consider short URLs for which no hit was observed during the last week of our trace. To define the inactivity threshold for our study we experimented with several different values. Figure 3 shows the popularity distribution for threshold values from 7 to 56 days. Using threshold values larger than 7 days does not affect the popularity distribution. [2]

Figure 4 separately examines the distribution of the active and

[2]Similar results were observed when examining the lifetime curve for different inactivity thresholds.

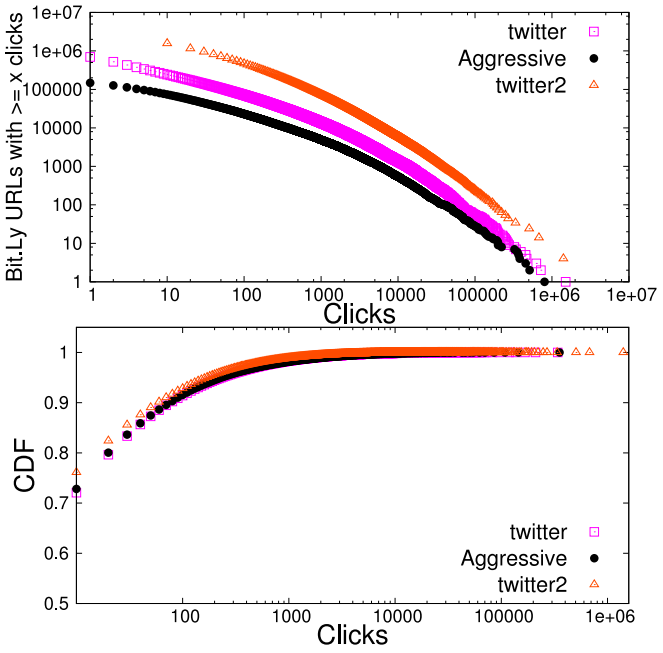

Figure 2: Popularity of bit.ly URLs.

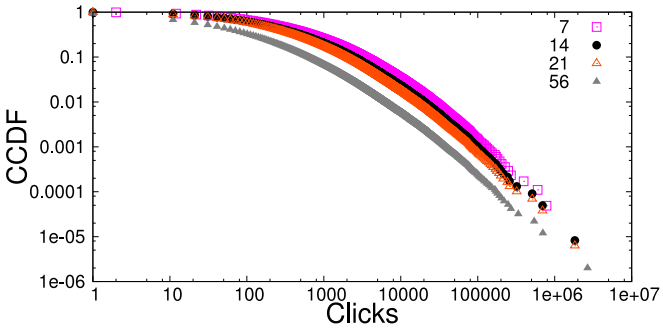

Figure 3: Popularity of bit.ly URLs using different activity thresholds.

inactive short URLs for the *twitter2* trace. Both curves appear similar to the original distribution. Further, a 90-10 rule seems to apply to the distribution. That is, we see that 10% of the short URLs are responsible for about 90% of the total hits seen in our trace.

Content Popularity: So far we have analyzed the overall popularity of individual short URLs, and examined its distribution. We now proceed to study *which web sites people access using short URLs*. Using the daily access information from the *twitter* and *bitly* traces, we try to answer questions such as: i) Which are the most popular web sites accessed through short URLs? ii) Are these sites similar to the ones found in the "traditional" web? iii) Does the set of these popular web sites change over time, and if so, how?

Table 5 lists the 10 most popular web sites: that is, the sites which received the highest numbers of hits through the short URLs in the two traces. Surprisingly, besides familiar sites, such as Youtube and Facebook, we observe others that are less known or popular according to well known ranking services such as Alexa and Netcraft; for example, `pollpigeon.com` (a service for very short opinion

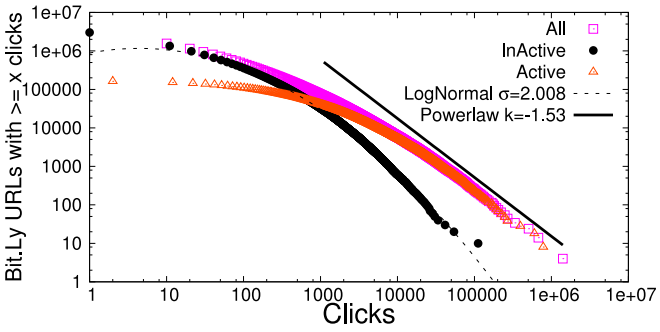

Figure 4: Popularity distributions for Active and In-Active bit.ly URLs from *twitter2* trace.

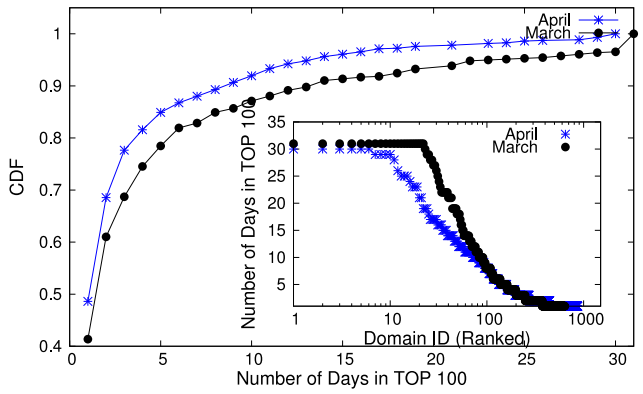

Figure 5: Number of days a domain name is in TOP-100 during March and April 2010.

polls), `mashable.com` (a social media news site), `twibbon.com` (a Twitter campaign support site), etc. Note that the list does not significantly change when using the data collected through aggressive crawling (Section 3.3), nor the larger Twitter trace (*twitter2*). This further supports that our selected sampling gives a good representation of the overall statistics collected through Twitter.

As we have observed previously, short URLs are mostly found in social networking or interaction environments and, thus, their popularity reflects the interests of the particular communities. For example, taking short polls is very common in social networking sites. Thus, such URLs rank very high in accesses through short URLs, even though they may not rank high in a more general web browsing environment. Overall, our findings indicate that while the community which browses the web through short URLs shares some interests with the broader web browsing community, it also presents a distinctive focus on web sites of special interest.

In addition to identifying the popular web sites, we are also interested in understanding whether these web sites significantly change over time. To this end, we calculated the 100 most popular web sites per day for the entire months of March and April 2010. 868 and 636 different sites where present in the daily top-100 respectively. Figure 5 displays the number of days a site appears in the top-100 each month. The Figure shows that there are about 6 sites which appear every single day of April 2010 in the top-100 (22 sites for March 2010). These compose a kernel of popular sites which does not seem to change over time, and has captured the attention and interest of bit.ly users. Additionally, we see that there are about

Rank	*twitter*				*bitly*			
	Site	% of Accesses	Alexa Rank	NetCraft Rank	Site	% of Accesses	Alexa Rank	NetCraft Rank
1	www.youtube.com	10.42	3	3	winebizradio.com	15.2	2693058	N/A
2	mashable.com	2.14	315	1175	www.youtube.com	10.51	3	3
3	www.facebook.com	1.91	2	2	livesexplus.com	3.98	15250029	N/A
4	www.47news.jp	1.51	3376	14605	mashable.com	2.28	315	1175
5	pollpigeon.com	1.24	57842	153550	inws.wrh.noaa.gov	2.27	1169	N/A
6	www.omg-facts.com	1.1	N/A	150669	www.alideas.com	2.26	7536010	N/A
7	twibbon.com	0.76	21271	55376	about:blank	1.87	N/A	N/A
8	itunes.apple.com	0.75	52	673	googleblog.blogspot.com	1.63	2251	2223
9	www.newtoyinc.com	0.72	167768	988477	addons.mozilla.org	1.56	247	197099
10	www.guardian.co.uk	0.65	273	231	www.google.com	1.53	1	1

Table 5: The 10 most popular web sites as seen through the real user accesses of the bit.ly URLs in traces *twitter* and *bitly*.

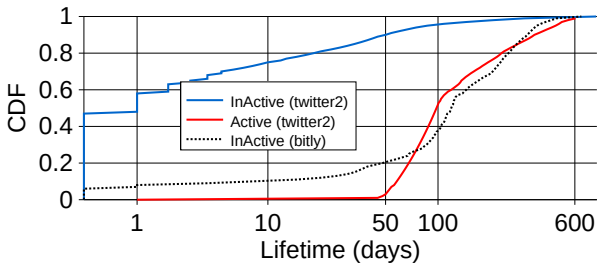

Figure 6: Lifetime analysis of short URL in traces *twitter2* and *bitly*.

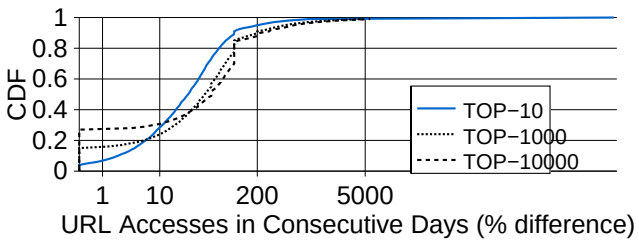

Figure 7: Cumulative Distribution Function for the daily click differences for the TOP-10/1000/10000 short URLs.

400 sites which appear once or twice in the top-100, enjoying short bursts of popularity. The results for the top 10 most popular web sites per day show similar behavior in a smaller scale. We further examine this burstiness effect in detail in Section 5.

5. EVOLUTION AND LIFETIME

The analysis throughout the previous section highlights the fact that short URLs differ from traditional URLs in many ways. Being published through social networking applications (Section 4.1), they have inherent idiosyncrasies that affect their observed activity over time. Indeed, the liveness of a short URL depends on factors such as the visitor's activity and her screen real estate. Since news feeds in social network environments typically display recent activity and are frequently updated, once a short URL disappears from the visitor's screen, it has almost no chances of getting clicked. Furthermore, short URLs are not directly "searchable" and are far from easy to remember, therefore users rarely access them explicitly.

In this section, we analyze how active a short URL is, by examining its hit rate over time. Specifically, we ask the following questions: i) Are short URLs ephemeral or do they survive for long periods of time? ii) How is the hit rate of a short URL spread across its lifetime? We consider such queries pertinent to the cacheability of short URLs that provide implications for the design of shortening services (e.g., URL recycling).

5.1 Life Span of short URLs

To examine the life span of short URLs we focus our attention on the *twitter2* and *bitly* traces. Both traces refer to the same shortening service which provides the daily hit rate per short URL. We define the life span, or *lifetime*, of a URL as the number of days between its last and first observed hit.

Figure 6 displays the lifetime CDF of the two traces. The figure further splits URLs into *active* and *inactive*, as these are defined in section 4.4. Recall that, as "inactive", we consider all short URLs for which no hit was observed during the last week of our trace. This split provides a feel of how the lifetime distribution depends on the activity of the URL, and will also be clarified in the following section when we examine the temporal characteristics of the URL hit rate.

One out of two short URLs are not ephemeral! While one might expect that short URLs are mostly ephemeral URLs, i.e. lasting for a few days, the aforementioned figure shows that 50% of the active short URLs for the *twitter2* and *bitly* traces have a lifespan of 98 and 124 days respectively. On the other hand, inactive URLs have a shorter lifespan as expected, with 51% only lasting for a day for the *twitter2* trace. Still a significant fraction of short URLs (more than 15%) last at least one month.

5.2 Temporal evolution

Having observed that a significant fraction of URLs survives for numerous days, we will now turn our focus on how hits are spread throughout a URL's lifetime. For the remainder of this section, we will use the *twitter2* trace, unless otherwise specified.

Looking at the evolution of the number of hits per day per URL as a function of time for several high volume URLs we observe several distinct patterns. Some show sudden increases or spikes while others have a significant decrease in hit rate. However, in all cases the bursty nature of access patterns was evident.

We attempt to characterize this burstiness in a more generic fashion across several URLs, by measuring the daily change in the number of hits for each short URL for the top-10, top-1000 and top-10000 short URLs (see Figure 7). We observe that the median value is around 24% for the top-10 URLs and around 40% and 50% for the top-1000 and top-10000 URLs. In other words, the number of accesses for a typical short URL varies by as much as 40% from

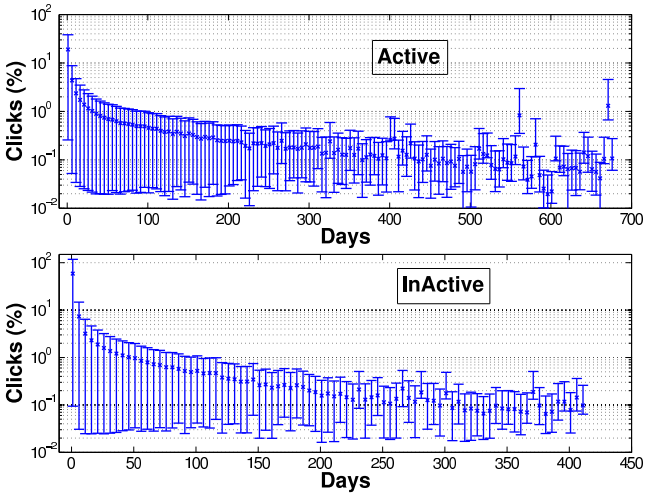

Figure 8: Mean and confidence intervals for the fraction of daily hits over the total clicks versus the lifetime of the URL.

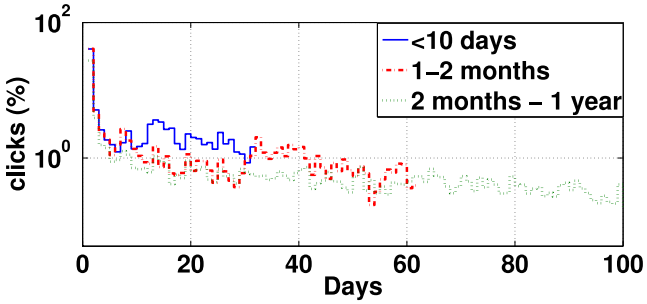

Figure 9: Fraction of hits per day conditioned on different lifetimes.

one day to the next. Moreover, for 10% of the days, this change is at least 100% for the top-10 and around 200% for the top-1000 and top-10000 URLs. Overall, we notice that as less popular URLs are included, that is as we move from the top-10 to the top-1000 and top-10000, we observe increasingly larger daily changes. This reflects the existence of URLs that only enjoy a few days of high popularity, and are then "forgotten".

1 day of fame. We further examine the evolution of hit rate across the lifetime of the short URLs in Figure 8, where we examine the mean, and confidence intervals of the fraction of a short URL's total hit rate over its lifetime, across all short URLs (with 0 denoting the creation day of the short URL). The figure depicts both active (top) and inactive (bottom) short URLs which show two distinctive patterns. For the inactive URLs, we observe that on the average 60% of hits are observed during their first day. As a short URL ages, its hit rate drops sharply and then stays roughly constant as the hit ratio converges to 0. In fact, this observation holds irrespective of the lifetime of the short URL (see Figure 9). In contrast, while this first-day effect is also evident for active short URLs albeit with at a smaller fraction (at roughly 18%), we also observe a significant hit rate for recent days. This reflects popular short URLs that still enjoy a significant hit rate.

As previously mentioned, Figure 9 shows no obvious dependence of the daily hit rate with a short URL's lifetime for inactive

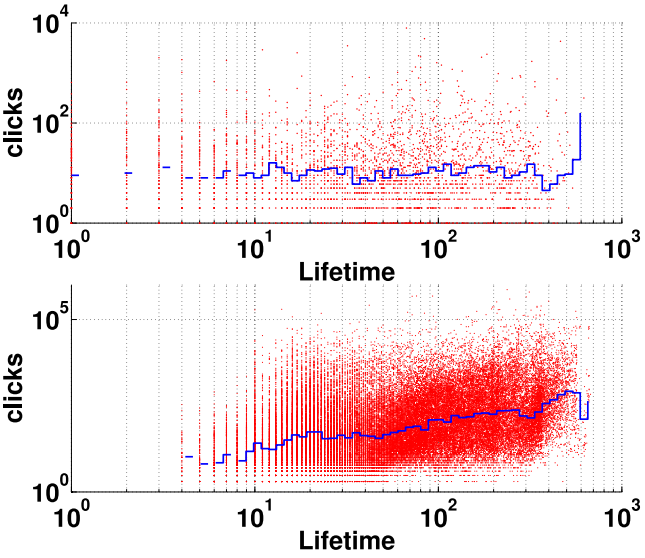

Figure 10: Lifetime of a short URL vs. number of hits.

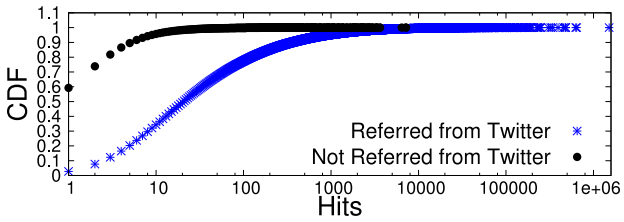

Figure 11: *The Twitter effect.* Difference in popularity for Twitter referred short URLs vs. non-Twitter referred ones.

short URLs. We examine this relationship in more detail by looking at the total number of hits as a function of the short URL's lifetime (Figure 10, median hit rate). The figure is in accordance with our previous observation for the inactive short URLs (top), namely that no obvious relationship exists. On the contrary, active short URLs (bottom) appear to exhibit a linear relationship in log-log scale with the lifetime of the URL.

Summarizing our discussion in this section, contrary to our expectations, we observe one out of two short URLs are not ephemeral. More than 50% of the active short URLs tend to live for more than three months. Moreover, a large number of short URLs enjoy occasional hits that may skew their lifetime. This implies that design mechanisms for shortening services should not expect a short lifespan of short URLs that is in the order of days. In addition, most short URLs enjoy a high hit rate relative to their total hits during their first day of creation, with the fraction of hits significantly dropping after.

6. PUBLISHERS

In this section, we focus our interest on the publishers of short URLs, i.e., users who include short URLs in Twitter messages. Twitter provides a unique opportunity for users to easily increase the popularity of their published content in a social network, which may not be possible with some of the other short URL sources. Figure 11 confirms this hypothesis by plotting the popularity of short URLs that received at least one hit from a Twitter user versus

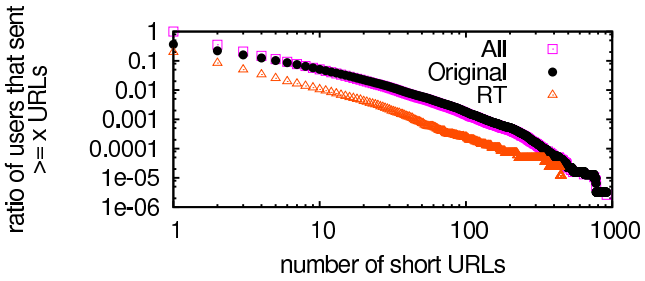

Figure 12: CCDF of posted short URLs per Twitter user. The distribution is heavy-tailed with a small percentage of users posting a large number of short URLs

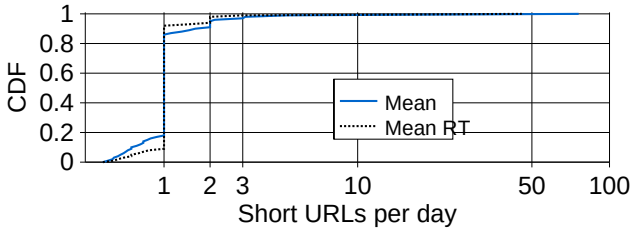

Figure 13: Number of posted short URLs per day per user.

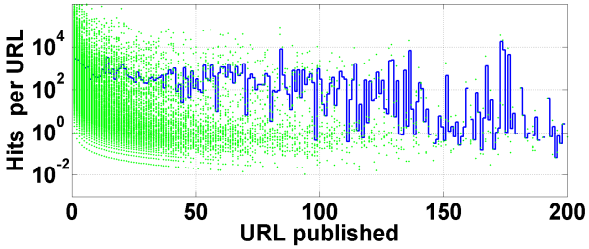

Figure 14: Expected hits as a function of the URLs published per user.

the popularity of all other short URLs. The *Twitter effect* is obvious: short URLs referred from Twitter enjoy significantly higher popularity compared to short URLs not experiencing this type of "word-of-mouth" propagation. Thus, examination of the publish rate and the popularity of published tweets relates to the propagation of *User Generated Content (UGC)* within social networks (e.g., [11, 13]), although the content reflected by the short URL in this case might not have been generated by its publisher. Note that Twitter messages may reflect original messages or "retweets", i.e., messages that are re-postings of an original message.

Our driving questions are: i) What does the distribution of published URLs per user look like? Are there any automated users which publish disproportionately large numbers of short URLs? ii) What is the activity of a typical user? This question relates to the publish rate of new URLs over time. Furthermore, do most users publish original URLs or retweet existing ones? iii) Does a higher publish rate per user imply a higher hit rate for the URLs published? This is pertinent to the propagation of a user's published URL and the population this URL may reach.

Figure 12 plots the Complementary CDF (CCDF) of posted short URLs per Twitter user. Most users published a handful of tweets with short URLs (the median is equal to 1 short URL). Overall, 90% of the users generated 5 or less such tweets each, and 65% of the users generated only one tweet containing a short URL. On the other hand, we see that some users generated hundreds of such tweets. For example, the most prolific user generated just under one thousand such tweets. Interestingly, the majority of tweets with short URLs are original Twitter messages and not retweets (RT).

Publishing about a thousand tweets in a week is an impressive number of published messages. For this reason, we now focus on the most prolific publishers in order to understand their behavior. We subsequently inspected the profiles of the top 12 publishers. Each tweet carries a label indicating the way it was posted, i.e., via the web site, the official API or a third-party application. From these top publishers, 10 uploaded their messages via twitterfeed [7] and the other two via TweetDeck [4] and the API respectively. Twitterfeed is an application designed specifically for automatically relaying the contents of an RSS feed via tweets. Furthermore, we visually identified bursty message patterns in all profiles with tweets coming in batches of two or three, every few minutes. All the above clearly indicate a semi-automated behavior.

To examine the users' daily publish rate of short URLs, Figure 13 displays the corresponding CDF. We observe that the median rate is 1 short URL per day, while 98% of the users publish no more than 5 short URLs per day. For prolific publishers we also observe a high number of short URL in a daily basis, also explained by the several automated applications used by Twitter users.

Intuitively, a users' publish rate should correlate with the total number of hits observed for his published URLs. However, the nature of this relationship is not evident, and depends on whether a users' followers indeed click on the posted short URL. For example, spammers or advertisers may not observe as many hits for subsequent published URLs. We examine this relationship in Figure 14, which displays the expected hits per URL as a function of the published URLs across users. We see that as the number of URLs published by a poster increases, the expected hit rate drops. This may imply either spamming-type behavior for heavy publishers, or that only a few short URLs from each publisher enjoy high hit rates compared to the rest of the user's published short URLs.

7. SHORT URLS AND WEB PERFORMANCE

Having studied the access patterns of short URLs, we now turn our attention to understanding potential performance implications of their use. We consider two such cases, namely: i) To what extent do short URLs offer space reduction compared to long ones? ii) short URLs introduce an extra step of indirection in the process of accessing web content. Hence, we attempt to quantify the performance penalty of this extra step. For example, could it turn out to be a major performance bottleneck?

7.1 Space Reduction

In this section we explore the amount of space saved through URL shortening services. As gain, we define the relative ratio of the URLs' length before and after the shortening service. Figure 15 displays this gain for the short URLs in traces *twitter* and *owly*. For roughly 50% of the URLs, we observe a 91% reduction in size, or about a factor of 10. Furthermore, for 90% of the URLs, the short version takes up to 95% less space than the long one - a factor of 20 improvement. Therefore, we see that URL shortening services are *quite* effective at reducing URL size and can provide significant

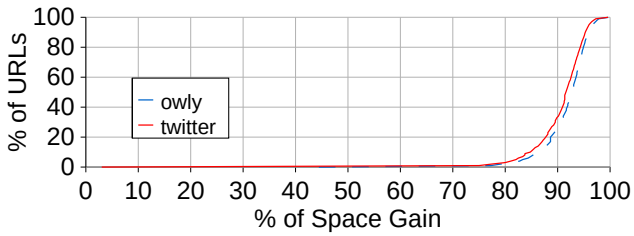

Figure 15: Reduction in URL size achieved by URL shortener services.

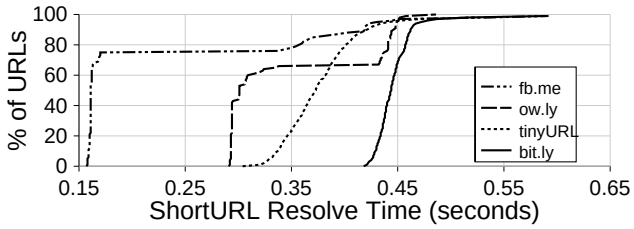

Figure 16: Latency in seconds imposed by 4 different URL shortening services.

benefit in environments where space is at a premium. A real-world approximation of the space saved by short URLs is the case of Twitter, where users place short URLs in their messages. While each tweet is limited to 140 characters, we assume that users, who would not be able to fit a long URL in their message, would either create a second tweet or not tweet at all. In our *twitter* trace, we replaced the bit.ly URLs in all tweets with their equivalent long versions and found that only 31% remained under the character limit.

7.2 Latency

Although URL shortening services offer a substantial space benefit over long URLs, they nonetheless impose an additional indirection in the user's web request. This may result in an increased web page access time, user-perceived latency and an overall degradation of performance. In this section, we quantify the latency such URL shortening services add to the overall web experience by exploring whether this imposes a significant overhead in web access times.

To estimate the overhead added by URL shortening services, we periodically accessed the 10 most popular short URLs in each of four such services, namely bit.ly, ow.ly, tinyURL.com and fb.me, as seen in the *twitter2* trace. Each short URL was accessed every 5 minutes for a time frame of 30 days. For each access we logged the total time of the web page transfer and the time needed for the redirection imposed by the URL shortening service. Figure 16 shows the extra cost incurred due to the redirection. Three of the services are closer together, exhibiting a median value of this overhead in the order of 0.37 seconds, while, in any case, none of them lies lower than 0.29 seconds. The fourth service, fb.me, a Facebook.com shortening service, appears to have a much smaller median value, in the order of 0.16 seconds and a lower bound very close to that. However it exhibits a bimodal behavior in terms of latency with 75% of redirections imposing no more than 0.17 seconds delay and 25% slowing down the user's requests by more than 0.33 seconds. Furthermore, the distance between the fastest and slowest 5% of accesses is 0.272 seconds. ow.ly shows a similar bimodal behavior with 66% of redirections imposing less than 0.33 seconds delay and the rest 34% adding a delay around 0.44 seconds. On the other hand, bit.ly appears to be the slowest but shows more consistent behavior with a distance of 0.046 seconds. We speculate that this bimodal behavior of fb.me and ow.ly to be due to caching policies followed by the two services. Though, we do not observe any correlation with the time of day for either service.

Figure 17 puts the redirection overhead of bit.ly in perspective and displays it as a percentage of the total web page access time. Using the top 200 short URLs from *twitter* we measure the additional overhead imposed for accessing a web page through a short URL. We observe that in more than 50% of the accesses, the URL shortening redirection imposes a relative overhead of 54%, while in 10% of the accesses this overhead is about 100% - a factor of two.

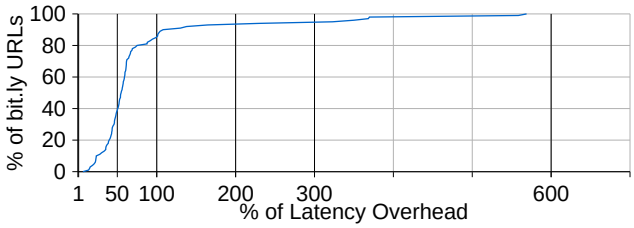

Figure 17: Latency imposed by URL shortening services for the 200 most popular URLs in *twitter* trace. The latency is plotted as an overhead percentage relative to the web page access time.

We see then, that even though the additional delay seems to be less than half a second and may be considered small by some people, it turns out to be comparable to the final web page access time in a significant fraction of the examined cases. Therefore, should URL shortening services become even more widespread, their latency may prove even more evident, with a non-negligible penalty on performance; this implies that alternative shortening architectures for eliminating such overheads may be required in the future.

8. RELATED WORK

Interest in online social networks and services has been significant over the past years. Several measurement studies have examined basic graph properties such as degree distributions or clustering coefficients [14,21] or their particular structure [17]. While part of our traces originates from Twitter, our work significantly differs from these studies as we focus on the use of short URLs and their presence within a social network, rather than network itself.

Part of our analysis relates to the evolution of content popularity [12, 13], information propagation through social links [13, 20], as well as popularity of objects and applications in social networks [11, 22]. For example, in [12, 13] the authors study how Flickr images evolve and how information propagates through the Flickr social graph. Lerman and Ghosh in [19] examined the information spread in Twitter and Digg and showed that although Twitter is a less dence network and spreads information slower than Digg, information continues to spread for longer and penetrates further the social graph. In a spirit similar to these studies, we examine how content becomes popular over time. However, in this work, we focus on how this popularity is reflected by the hit rate of short URLs. Cha et al [11] also deal with content popularity by performing a study of user generated content via crawling the YouTube and Daum sites. The authors observed the presence of the Pareto principle. Our analysis confirms that this is also the case in the popularity of short URLs. Our observations on the dispersion of

the hit rates of short URLs are consistent with the well-documented findings on the existence of Zipf's Law and heavy-tailed distributions in WWW (e.g., [10, 15]). However, our work further highlights that a web site's popularity does not necessarily translate in an equivalent popularity in the "web of short URLs".

Information propagation in Twitter has been studied in [18]. The authors have crawled the Twitter network and analyzed the temporal behavioral of trending topics. The authors suggested that Twitter is mostly a news propagation network, with more than 85% of trending topics reflecting headline news. Indeed, this observation is also confirmed by our study. A large fraction of short URLs points to news-related domains; however, the percentage of news related URLs appears lower in our study, 7 out of the top-100 URLs.

9. CONCLUSIONS

We have presented a large-scale study of URL shortening services by exploring traces both from the services themselves and from one of the largest pools of short URLs, namely the Twitter social network. To our knowledge, this paper presents the first extensive characterization study of such services.

Specifically, we provided a general characterization on the web of short URLs, presenting their main distribution channels, their user community and its interests, as well as their popularity. Furthermore, we explored their lifetime and access patterns showing an activity period of more than a month with an increased popularity over the first days of their life. We explored the publishers of short URLs, and show a possibility of increased popularity when short URLs are accessed through Twitter. Additionally, a publisher of such URLs is more likely to be considered a spammer and enjoy decreased popularity when operating at an aggressive rate. Finally, we quantified the performance of URL shortening services, showing a high space gain in terms of bytes used, but also increased overhead in the web page transfer times when accessed through short URLs. This overhead increases web page access time by more than 54% in 50% of the cases, implying that alternative shortening architectures may be required in the future.

Acknowledgment

This work is supported in part by Herakeitos II PhD Scholarship in the area of "Internet traffic classification". Elias Athanasopoulos is funded by the Microsoft Research PhD Scholarship project, which is provided by Microsoft Research Cambridge. This work is supported in part by the Marie Curie Actions - Reintegration Grants project PASS and by the project SysSec funded in part by the European Commission, under Grant Agreement Number 257007. We thank the anonymous reviewers for their valuable comments and Christos Papachristos for his valuable help. Demetris Antoniades, Iasonas Polakis, Georgios Kontaxis, Elias Athanasopoulos and Evangelos P. Markatos are also with the University of Crete.

10. REFERENCES

[1] Alexa Traffic Stats. `http://www.alexa.com/siteinfo/bit.ly#trafficstats`.

[2] Announcement of URL shortening service available at makeashorterlink.com. `http://www.metafilter.com/8916/`.

[3] TinyURL.com. `http://tinyurl.com/`.

[4] TweetDeck. `http://www.tweetdeck.com/`.

[5] Twitter Rate Limit. `http://apiwiki.twitter.com/Rate-limiting`.

[6] Twitter Search. `http://search.twitter.com/`.

[7] TwitterFeed. `http://twitterfeed.com/`.

[8] Wikipedia - List of countries by number of Internet users. `http://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users`.

[9] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. In *IN INFOCOM*, pages 126–134, 1998.

[10] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *IN INFOCOM*, pages 126–134, 1998.

[11] M. Cha, H. Kwak, P. R. P., Y.-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM IMC '07, San Diego, CA, USA*, pages 1–14, 2007.

[12] M. Cha, A. Mislove, B. Adams, and K. Gummadi. Characterizing Social Cascades in Flickr. In *ACM SIGCOMM Workshop on OSNs*, 2008.

[13] M. Cha, A. Mislove, and K. P. Gummadi. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proc. of the 18 Intl. World Wide Web Conference (WWW)*, 2009.

[14] H. Chun, H. Kwak, Y. Eom, Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *IMC '08: Proc. of the ACM SIGCOMM conference on Internet measurement*.

[15] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[16] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *CCS '10: Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM, 2010.

[17] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, 2006.

[18] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[19] K. Lerman and R. Ghosh. Information contagion: n empirical study of the spread of news on digg and twitter social networks. In *Proceedings of the 3th AAAI Conference on Weblogs and Social Media (ICWSM'10)*, pages 90–97, 2010.

[20] J. Leskovec, L. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *Proceedings of the 7th ACM conference on Electronic commerce (EC)*, 2006.

[21] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc of the 5th ACM/USENIX Internet Measurement Conference (IMC'07)*, 2007.

[22] A. Nazir, S. Raza, and C. Chuah. Unveiling facebook: a measurement study of social network based applications. In *IMC '08: Proc. of the ACM SIGCOMM conference on Internet measurement*.

[23] M. Ripeanu, I. Foster, and A. Iamnitchi. Mapping the Gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 2002.