Telematics and Informatics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

Bit.ly/practice: Uncovering content publishing and sharing through URL shortening services

Daejin Choi[a,1], Jinyoung Han[b,*], Selin Chun[a], Efstratios Rappos[c], Stephan Robert[c], Ted Taekyoung Kwon[a,*]

[a] Department of Computer Science, Seoul National University, Republic of Korea
[b] School of Computing, Hanyang University, Republic of Korea
[c] Institute for Information and Communication Technologies, Haute Ecole d'Ingenierie et de' Geston du Canton de Vaud, Switzerland

ARTICLE INFO

Keywords:
URL shortening service
Short URL
Bit.ly
Content publishing and sharing

ABSTRACT

It becomes the norm for people to share online content such as images, videos, and news over various channels including online social networks, news media, or online communities. One of the popular ways to publish and share online content is using a URL shortening service, which provides a short equivalent URL that is redirected to an original URL of content. This paper comprehensively analyze the practice of using short URLs from their creations to publishing to sharing, using a large scale dataset that contains 4.2 B requests for 80 M URLs created through Bit.ly, one of the most popular URL shortening services. We find that content URLs are mo-sunknown.

1. Introduction

The advances in Internet technologies and the conveniences of online services allow people to share various online content such as images, videos, news, e-books, opinions, write-ups, or product information. Such online content is shared through various online channels including online social networks (OSNs), news media, online communities, instant messages, e-mails, or other web services, by individuals as well as by companies/organizations.

URL shortening services are widely used in publishing and sharing online content by providing a short equivalent URL that is redirected to an original URL of the content (Antoniades et al., 2011). A user, who wishes to share a content page, can submit its original *content URL*, e.g.,www.facebook.com/video/abc.mp4, to a URL shortening service. Then the user can obtain a *short URL* in a concatenation form of the name of the shortening service domain and the hash value, e.g.,bit.ly/2gXUgJI. The main benefit of using a URL shortening service is that a user can publish a short, manageable, and human-unreadable URL to share content (Antoniades et al., 2011; Nikiforakis et al., 2014; Gupta et al., 2014). Hence, for example, microbloggers often use short URLs to share content in their microblogs that have length limits, e.g., 140-characters limit in Twitter (Antoniades et al., 2011). Also, short URLs are often used by spammers, attackers, or users who would like to hide the original content URLs of their content pages (Gupta et al., 2014).

`Bit.ly` is one of the most popular URL shortening services. It has received attention since 2009 when Twitter has started using `Bit.ly` as a default URL shortening service (Gupta et al., 2014). According to *New York Times*, people have created about 600 M URLs through `Bit.ly`, which have been requested over 8 B times (Newman, 2014). The popularity of URL shortening services has

* Corresponding authors: #609, ERICA Center, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggi-do 15588, Republic of Korea.
E-mail addresses: djchoi@mmlab.snu.ac.kr (D. Choi), jinyounghan@hanyang.ac.kr (J. Han), slchun@mmlab.snu.ac.kr (S. Chun), efstratios.rappos@heig-vd.ch (E. Rappos), stephan.robert@heig-vd.ch (S. Robert), tkkwon@snu.ac.kr (T.T. Kwon).
[1] Network Convergence & Security Laboratory, Room #414, Building #138, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea.

https://doi.org/10.1016/j.tele.2018.03.003
Received 6 February 2018; Accepted 4 March 2018
0736-5853/

attracted the research community to investigate usage patterns of short URLs, which provides a valuable insight into understanding its functionalities (Antoniades et al., 2011) and the potential risks of sharing short URLs (Nikiforakis et al., 2014). As short URLs hide their original content URLs, they are often used for sharing malicious content such as spams or phishing (Chhabra et al., 2011; Wang et al., 2013; Maggi et al., 2013; Nikiforakis et al., 2014). However, most of these studies paid little attention to content creation and sharing behaviors using such URL shortening services across various categories such as news, shopping, or adult content.

We comprehensively analyze the practice of using short URLs from their creations to publishing to sharing. Using a large-scale dataset that contains 4.2 B requests for 80 M URLs created through `Bit.ly`, we analyze (i) what types of content pages (e.g., news, adult, or sports) are shortened and shared, (ii) how and where short URLs are published, and (iii) how content pages are shared through different publishing spaces. In particular, we seek answers to the following research questions.

- **Q1 – Who use the URL shortening services?** We investigate who shorten content URLs. A `Bit.ly` user can be either an individual user or a company account. We find that content URLs are primarily shortened through the third party companies, e.g., Twitter (twitterfeed, tweetdeckapi, and twipple), Bit.ly (bitly and zatbitly), Facebook (rssgraffiti), substantially more than by individual user accounts.
- **Q2 – How do users shorten and share URLs for publishing and sharing content?** We find that short URLs are proliferated mostly across OSNs, news/media sites, and computer/electronics sites (e.g., newsfeed service). We also find that short URLs for OSN pages tend more to be requested, and `Facebook` is one of the most popular sources whose content pages are requested through short URLs. However, the URL shortening practice and request patterns show disparate patterns. For example, while there are not so many short URLs for the shopping websites or adult content, they are likely to be requested notably.
- **Q3 – Where short URLs are published?** To shed light on the practice of content publishing through short URLs, we model the relations among content and referrer domains[2], in the form of *content-referrer graph*. The analysis on the content-referrer graph reveals that different domains play different roles in publishing short URLs. For example, search engines, OSNs, and computer/electronics sites are popular spaces for content publishing while news and streaming services are widely used as content sources in general.
- **Q4 – How do users access short URLs in different types of publishing spaces?** We find that users are likely to access different types of content pages through different referrer domains; e.g., adult or malicious content pages tend to be requested from search engines; shopping content is primarily accessed through OSNs; and news are usually clicked through computer/electronics domains. Also, news and shopping pages, published through OSNs, tend to be requested quickly and virally.

The rest of this paper is organized as follows. We present how people use URL shortening services, `Bit.ly`, and review the related work in Section 2. We describe our dataset and the analyzing methodology in Section 3. In Section 4, we report the analysis results and their implication. We first investigate how short URLs are published and shared in terms of request patterns, content types, and temporal/geographical characteristics in 4.1. By modeling the relations among content and referrer domains, we analyze (i) how different domains are associated with others, and (ii) what domains play important roles in publishing short URLs in 4.2. We finally examine how short URLs are published and requested in different types of publishing spaces in 4.3.

2. Background

2.1. Bit.ly: a URL shortening service

URL shortening services assist in publishing and sharing content by providing a short equivalent URL that is redirected to an original URL (Antoniades et al., 2011). A user (who wishes to share content) can submit an original content URL to a URL shortening service, and he/she can obtain a short URL as a concatenation form of the name of the service domain and the hash value, e.g.,bit.ly/2gXUgJI. The user can then publish the obtained short URL to any place in which he/she wishes to publish, such as instant messages, e-mails, OSNs, and newsfeed services. Then, a person who wishes to access the content makes an HTTP request by clicking the corresponding short URL, the URL shortening service redirects the request to the original content URL. Fig. 1 illustrates how a URL shortener shortens an original content URL, and how a URL requester accesses the short URL.

The main benefit of using a URL shortening service is that a user can publish a short, manageable, and human-unreadable URL to share content (Antoniades et al., 2011; Nikiforakis et al., 2014; Gupta et al., 2014). Hence, for example, microbloggers often use short URLs to share content in their microblogs which have length limits, e.g., 140-characters limit in Twitter (Antoniades et al., 2011). Also, users who want to remove semantics from original URLs usually use short URLs for content sharing purposes. As a side effect, the short URLs are also used by spammers, attackers, or users who would like to hide original URLs (Gupta et al., 2014).

`Bit.ly` is one of the most popular URL shortening services. It has received attention since 2009 when Twitter has used it as a default URL shortening service (Gupta et al., 2014). According to *New York Times*, people have created about 600 M URLs through `Bit.ly`, which have been requested over 8 B times (Newman, 2014). `Bit.ly` also offers supporting functions for companies such as custom domain supports (e.g., nyti.ms for New York Times, pep.si for Pepsi) and analytics tools, which increases the popularity of the services for companies as well as for individuals.

[2] A referrer domain indicates a domain where a short URL is published, while a content domain represents a domain whose content is created.

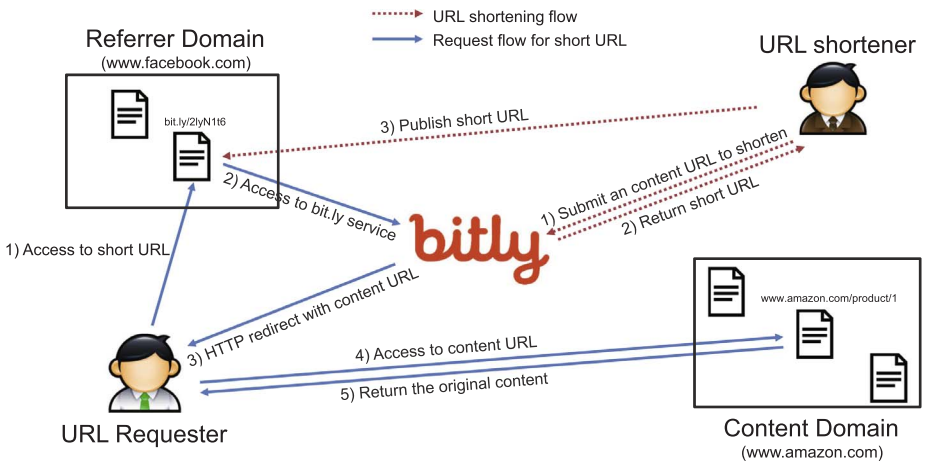


Fig. 1. We illustrate how a URL shortener shortens an original content URL, and how a URL requester accesses the short URL.

2.2. Related work

Online Content Publishing and Sharing: People share various online content such as images, videos, or news through different Internet systems, e.g., online communities, OSNs, e-commerces, or social curating services. Online communities or news services are one of the popular places where users share news or new information by uploading external content or URLs (Choi et al., 2015). Other users can write comments and exchange their thoughts to the uploaded content. Many researchers have investigated content sharing patterns in such online communities or news services (Kumar et al., 2010; Marcoccia, 2004; Gómez et al., 2008). Choi et al. analyzed posts and comments in `Reddit`, a popular online community, and characterized commenting patterns in terms of volume, responsiveness, and virality (Choi et al., 2015). They explored how characteristics of content and user participation behavior are associated with commenting patterns in `Reddit`. Using `Yahoo!`, `USENET`, and `Twitter` datasets, Kumar et al. observed content propagation in terms of volume and depth, and proposed a propagation growth model based on the observations (Kumar et al., 2010). Gomez et al. explored content propagation patterns in `Slashdot` (Gómez et al., 2008), a technology-related news website, and found that the degrees of comments follow a log-normal distribution. Wang et al. proposed a model to predict the volume of comments in `Digg.com`, a popular social news service, and applied the proposed model to different platforms such as `Twitter` and `Reddit` (Wang et al., 2012).

As OSNs have become one of the most popular places where various content types are shared, there have been many efforts in understanding and predicting content sharing patterns. Rodrigues et al. analyzed the word-of-mouth exchange of URLs among `Twitter` users and showed that URLs are likely to be shared among users who are geographically close together (Rodrigues et al., 2011). Bakshy et al. examined the patterns of information sharing in `Facebook`, and found that weak ties play a more important role in dissemination of content in `Facebook` (Bakshy et al., 2012). Cheng et al. showed that temporal and structural features are key factors to predict the size of a photo cascade generated by resharing in `Facebook` (Cheng et al., 2014). Cha et al. analyzed propagation patterns of photo content in `Flickr` and showed that photos do not spread widely and quickly (Cha et al., 2009). Goel et al. investigated the propagation patterns of URLs in `Yahoo!` and `Twitter`, and found that the majority of the diffusions occur within one hop from a seed node (Goel et al., 2012).

Recently, social curation services such as `Pinterest` have been reported as the vibrant places that encourage users to collect, organize, and share content by their tastes or interests (Han et al., 2014; Gelley and John, 2015), which reveal distinct consumption patterns compared to other online services. Han et al. investigated content propagation patterns in `Pinterest` using the collected large-scale data, and showed that sharing pins in Pinterest is mostly driven by pin's properties like its topic, not by users' characteristics such as the number of followers (Han et al., 2014). This was confirmed by Gelley and John (Gelley and John, 2015), who showed that 'following' is not significantly utilized in content sharing in `Pinterest`. Chang et al. investigated which categories are popular to male and female users in `Pinterest`, and showed that male and female users differ in collecting content across different topics. Han et al. showed that content creation and diffusion patterns are associated with users' different motivations and genders in `Pinterest` (Han et al., 2015).

While these studies provide valuable insights into understanding content publishing and sharing patterns in online systems, we focus on how content is published and shared in a form of a *short URL*. In particular, we explore how the content pages with different categories (e.g., news, shop, adult content) are shortened, published, and shared across different online systems.

URL Shortening Services: The characteristics of short URLs motivate people to use URL shortening services and its variants with different goals, which in turn has spurred active research into usage patterns of URL shortening services (Antoniades et al., 2011; Nikiforakis et al., 2014). Demetris et al. investigated how short URLs are shared based on the information pages of short URLs, providing daily statistics of short URLs, and tweet/retweets including the URLs (Antoniades et al., 2011). They provided *a macro-level view* of the short URL usage shared in Twitter, such as the daily click counts of tweets. On the other hand, we perform *a micro-level analysis* of the short URL usage including how content pages are created and published through short URLs, and how short URLs are

shared through various types of domains (e.g., search engine, computer & electronics, not to mention OSNs), based on the detailed request logs for Bit.ly short URLs.

As short URLs themselves can hide their original content URLs, they are often used for sharing malicious content such as spams or phishing. Hence, many studies have focused on the potential risks of sharing short URLs. Using the dataset of qc.rx, a well-known URL shortening service, Klien and Strohmaier studied how short URLs are used for spamming from a geographical perspective (Klien and Strohmaier, 2012). Chhabra et al. investigated how phishing URLs are shared and propagated in online services based on the Bit.ly and PhishTank datasets, and showed that short URLs in Twitter tend to be more requested from more countries for a longer time than other services (Chhabra et al., 2011). Wang *el al.* observed the spam short URLs published in Twitter, and developed a model for detecting spams (Wang et al., 2013). Using the two-years large-scale dataset from several URL shorting services, Maggi et al. analyzed how many users are exposed to malicious short URLs, and found that the threats of using short URLs are not as serious as those of using long URLs (Maggi et al., 2013). Nikiforakis (Nikiforakis et al., 2014) reported a high portion of short URLs created from ad-based URL shortening services are likely to be used for infecting users with malware and exfiltrating private data. On the contrary, we comprehensively analyze (i) what content types (e.g., news, adult, or sports) are shortened and shared, (ii) how and where short URLs are published, and (iii) how content pages are shared through different publishing spaces.

3. Materials and method

3.1. Dataset description

To investigate the practice of using URLs shortened by Bit.ly, we perform a measurement study using a large-scale dataset from Bit.ly. Our dataset consists of two parts – (i) short URL data and (ii) request data for the short URL. The short URL data includes the content (or original) URL a user shortens, the global hash of the target URL, a user id, and its creation time. Each request log consists of a global hash of short URL, its original URL, its referrer URL where the short URL is published, and the temporal, geographical request information such as request time, country, city, and timezone. Note that only anonymized user data is used for this research, and no personally identifiable information is used.

To characterize the URL properties, we additionally investigated the category of each of content and referrer domains. To this end, we first extracted the domain name of a content (or referrer) URL by removing all characters after the first delimiter '/'. For example, the domain name of a content URL ' www.facebook.com/video/abc.mp4' is ' www.facebook.com'. We then submitted the domain name of the content to a commercial URL scanner, VirusTotal[3], which scans a submitted URL over a corpus of five website scanning engines, and returns the category for the given URL. Note that the returned category name is usually different across the five engines, and the categorization is often not consistent even within a single engine. Also, some engines even require users to manually label categories. To address this problem, we perform a semi-manual categorization. That is, we made the standardized set of categories, each of which is provided by SimilarWeb[4]. Note that we mostly used the second-level categories in SimilarWeb. If there are only 1st level categories, we used them as they are. In addition, we added 'Violence & Illegal', 'Blogs', and 'Streaming' categories to the standardized category set, which are provided by VirusTotal engines but not by SimilarWeb. Finally, we have total 64 categories in our standardized category set.

Our dataset contains more than 80 M short URLs and their 4.2 B requests generated from more than 2.1 B devices and more than 220 countries during one month, June 2012. The top 3 countries by the number of requests are USA, China, and Japan. Considering the report that the portion of Internet users in these countries are 10.2%, 22.4% and 4.2%, respectively (DGTraffic, 2012), the result indicates that the short URLs are more heavily requested from the USA. Note that the numbers of content and referrer domains are 3.1 M and 2.2 M, respectively.

3.2. Content-referrer graph model

To explore how short URLs are published through domains, we model the relations among content and referrer domains as a *Content-Referrer graph*, a directed graph $G = (V,E,W)$, where V is the set of all domains, including content and referrer domains, and E is the set of edges. Each edge connects from a content domain to a referrer domain, where a short URL of a content URL is published. Note that any domain can be a content domain, a referrer domain, or both. The weight of an edge is the number of content URLs published in the referrer domain. Here, we consider only the content URLs requested at least once. Fig. 2 illustrates a relation between a referrer domain (Twitter) and a content domain (Facebook), which is modeled as the directed edge between the two nodes in the content-referrer graph. Note that the content-referrer graph is a forest that consists of multiple distinct components across which there is no reachable path.

We finally build a content-referrer graph based on more than 3 M content domains and 2 M referrer domains. The number of nodes, edges, and components are about 4.3 M, 12 M, and 48 K, respectively. Note that requests from non-websites (e.g., Instant Message and Apps) are labeled as 'direct' in Bit.ly, and are removed in constructing the content-referrer graph.

[3] https://www.virustotal.com/.
[4] https://www.similarweb.com/category.

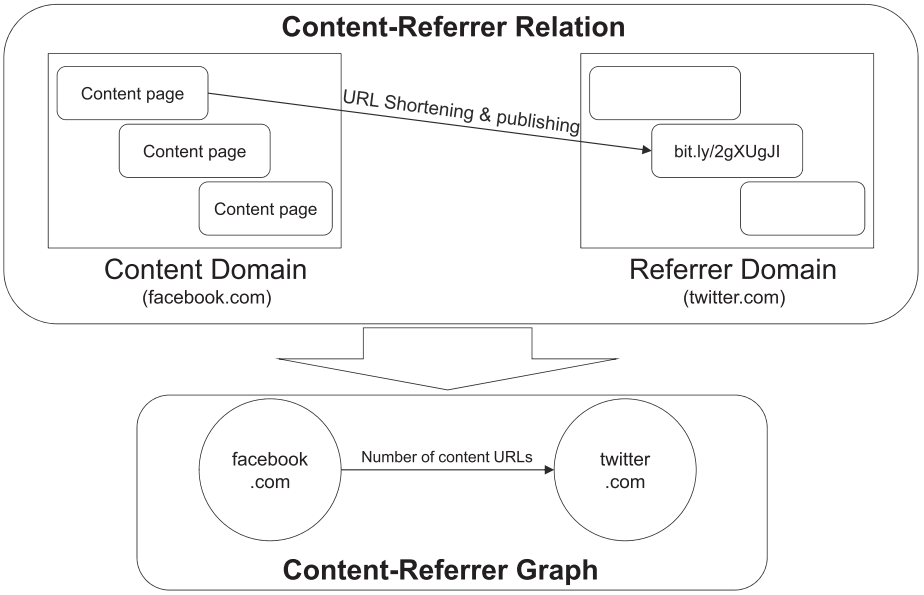


Fig. 2. We model the relations among content and referrer domains in the form of a *content-referrer* graph. If a tweet has a short URL for a content page in Facebook, there is a directed edge from Facebook (content domain) to Twitter (referrer domain). The weight (of an edge) is the number of content URLs published in a referrer domain.

4. Results

4.1. Content sharing patterns thorough Bit.ly

We first investigate how people create short URLs through `Bit.ly`. Figs. 3(a) and (b) show the distributions of the number of short URLs created by and that of domains shortened by each user, respectively. As shown in Fig. 3(a), 35% of users shorten only 1 URL while 3.66% of users create more than 100 short URLs. Fig. 3(b) shows that around half of users create short URLs of only a single domain, but 0.32% of users shorten content pages in more than 100 domains, meaning that URL shortening users are likely to shorten content for a small number of domains. Interestingly, the CCDF of empirical data is under the fitting function in [100,10000] ranges, but is over the fitting function when the number of domains is greater than 10000, meaning that URL shorteners tend to create short URLs for either only a small number of domains or a large number of domains. Note that the two distributions both follow the power-law (Barabasi and Albert, 1999) with $(1.7095, -0.84918, x \geqslant 10)$ and $(0.75502, -1.0172)$ as parameters.

We next investigate relatively 'active' `Bit.ly` users who create more short URLs than others. Here, a `Bit.ly` user can be either an individual user or a company account. Table 1 shows the top 10 `Bit.ly` users in terms of number of created short URLs. As shown

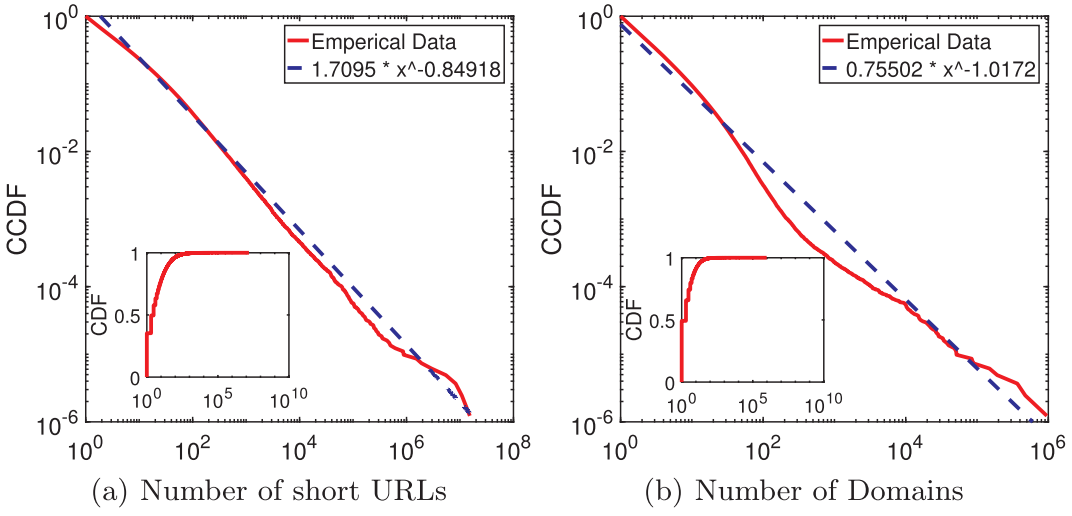


Fig. 3. Numbers of short URLs and domains for each user are plotted.

Table 1
Top 10 Bit.ly users in terms of number of created short URLs are shown.

Rank	User Name	Number of Created Short URLs
1	twitterfeed	15,276,864
2	dens	10,820,680
3	bitly	8,289,480
4	rssgraffiti	5,448,646
5	tweetdeckapi	2,952,763
6	addthis	1,908,226
7	ameba	1,539,152
8	ifttt	884,396
9	twipple	863,334
10	zatbitly	615,117

in Table 1, the top shorteners are likely to be the third party companies rather than individual users. For example, the third party services of Twitter (twitterfeed, tweetdeckapi, and twipple), Bit.ly (bitly and zatbitly), and Facebook (rssgraffiti) are ranked in the top 10 list. Moreover, 'dens' and 'ameba', made by service providers (foursquare.com and ' ameblo.jp', respectively) to encourage their users to publish content by short URLs for their services, are also heavily used for URL shortening. Interestingly, management tools for web services such as 'addthis', 'ifttt' are widely used in shortening URLs.

We then examine the top categories and top domains in terms of number of URLs in Tables 2 and 3, respectively. Table 2 shows that content in 'Social Network', 'News & Media', and 'Computer & Electronics' (e.g., newsfeed service) is likely to be published through short URLs. The top domain (in terms of number of short URLs) in Table 3 is ' foursquare.com'. We find that most of the URLs associated with ' foursquare.com' are the check-in information, which are published by users who wish to share their current location information with others. Interestingly, ' ameblo.jp' (a Japanese microblog service) ranks higher than other global companies such as Google, Facebook, and Twitter, which implies a heavy usage of short URLs for content in 'ameblo.jp'. This may be partially because the global companies provide their own URL shortening services: goo.gl, fb.me, and t.co for Google, Facebook, and Twitter, respectively. The portion of content URLs for the 'Shopping' category is over 8%, and they are mostly originated from ' www.amazon.com', implying that short URLs are widely used in publishing shopping content.

To investigate how uniformly each user shortens content pages across the categories and domains, we count the number of content URLs a user has shortened, and calculate the *category entropy* and *domain entropy* for each user u as follows:

$$Entropy_u = -\sum_{m=1}^{N_u} p_m^u \log p_m^u \tag{1}$$

where N_u is the number of categories/domains associated with the URLs of user u, and p_m^u is the URL portion of the m^{th} category/domain of user u.

Fig. 4 shows the median, average, and uniform entropy values as the number of categories/domains (of a single user) increases. Note that the uniform values are calculated when the numbers of content URLs are equal across the categories/domains. The gap of entropy values between uniform and median (and average) increases as the number of categories (or domains) increases. This signifies the skewness of users' interests in shortening URLs – although there are a small number of users who shorten URLs in many categories or domains, most users are likely to focus on a few categories (and domains) in shortening URLs.

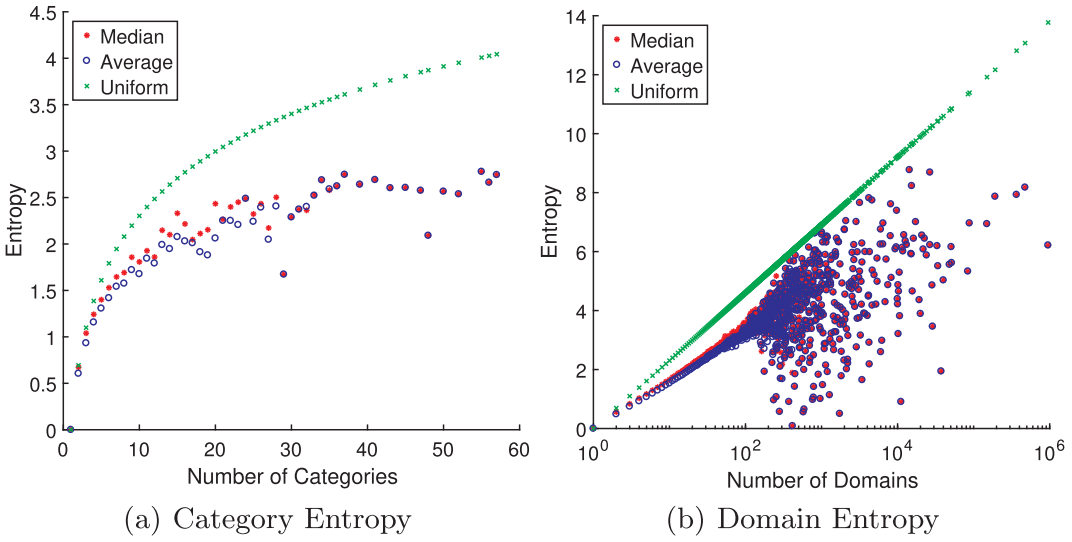
We next investigate what types (or categories) of content are requested through short URLs. Tables 4 and 5 show the top 10 categories and domains in terms of number of short URL requests (i.e., through URL clicks), respectively. Overall, short URLs for 'Social Network' are heavily requested; content pages in ' www.facebook.com' are most requested through short URLs. In addition, ' www.youtube.com' and ' www.amazon.co.jp' are also in top 10 by the number of requests. These results seem to be related to the global popularity of websites reported in Campbell (2012). That is, popular websites (e.g., Google, Facebook, Amazon, and Youtube), whose content pages are heavily accessed in general, are also more likely to be requested through short URLs. However, interestingly,

Table 2
Top 10 categories in terms of number of short URLs are shown.

Rank	Category	Portion of URLs (%)
1	Social Network	25.59
2	News & Media	15.39
3	Computer & Electronics	9.51
4	Shopping	8.04
5	Business & Industry	4.49
6	Blogs	4.10
7	Search Engine	3.24
8	Sports	3.11
9	Arts & Entertainment	3.03
10	File Sharing	2.59

Table 3
Top 10 domains (in terms of number of short URLs) and their associated categories are shown.

Rank	Domain	Category	Portion of URLs (%)
1	foursquare.com	Social Network	13.39
2	ameblo.jp	Social Network	2.30
3	feedproxy.google.com	Computer & Electronics	2.28
4	www.amazon.com	Shopping	1.76
5	www.google.com	Search Engine	1.32
6	www.facebook.com	Social Network	1.27
7	www.youtube.com	Streaming	1.25
8	twitter.com	Social Network	1.11
9	news.google.com	News & Media	1.06
10	apps.facebook.com	Social Network	1.05



(a) Category Entropy (b) Domain Entropy

Fig. 4. The median, average, and uniform values of category and domain entropy are plotted as the number of categories/domains increases.

Table 4
Top 10 categories in terms of number of requests are listed.

Rank	Category	Portion of URL Requests (%)
1	Social Network	14.95
2	Shopping	14.55
3	News & Media	9.86
4	Computer & Electronics	7.68
5	Adult	7.08
6	File Sharing	5.82
7	Arts & Entertainment	5.20
8	Streaming	5.09
9	Business & Industry	3.67
10	Sports	3.27

the portion of requests of the content pages in relatively less popular domains such aswww.pornhub.com,mlks.co, andwww.lapatilla.com are also high, meaning that content pages in these domains tend to be accessed through short URLs rather than direct access. The gap of popularity may come from the functionalities of short URLs; For example, since the informative text (e.g., domain name or content title) represented in URL can be hidden through URL shortening, adult content pages are likely to be published and shared through short URLs. Note that the URL publishing practice and access patterns are disparate when we compare Tables 2 and 4; while there are not many short URLs for content in the 'Shopping' and 'Adult' categories, they are likely to be requested many times.

We next investigate how many domains the short URLs are requested from, which are called *referrers*. For example, if a user clicks a short URL in Facebook,www.facebook.com is a referrer domain. Fig. 5(a) shows the distribution of the numbers of referrer domains for a given short URL. About 60% of short URLs are requested from only one referrer domain while 0.01% of short URLs are requested from more than 100 referrer domains. Note that the CCDF of empirical data is over the fitting line, meaning that there are a few short

Table 5
Top 10 domains (in terms of number of URL requests) and their associated categories are listed.

Rank	Domain	Category	Portion of URL Requests (%)
1	www.facebook.com	Social Network	8.38
2	www.pornhub.com	Adult	4.96
3	apps.facebook.com	Social Network	1.99
4	rtm.ebaystatic.com	Shopping	1.74
5	www.youtube.com	Streaming	1.56
6	itunes.apple.com	File Sharing	1.42
7	mlks.co	Uncategorized	1.13
8	www.amazon.co.jp	Shopping	1.06
9	www.lapatilla.com	News & Media	1.01
10	feedproxy.google.com	Computer & Electronics	0.69

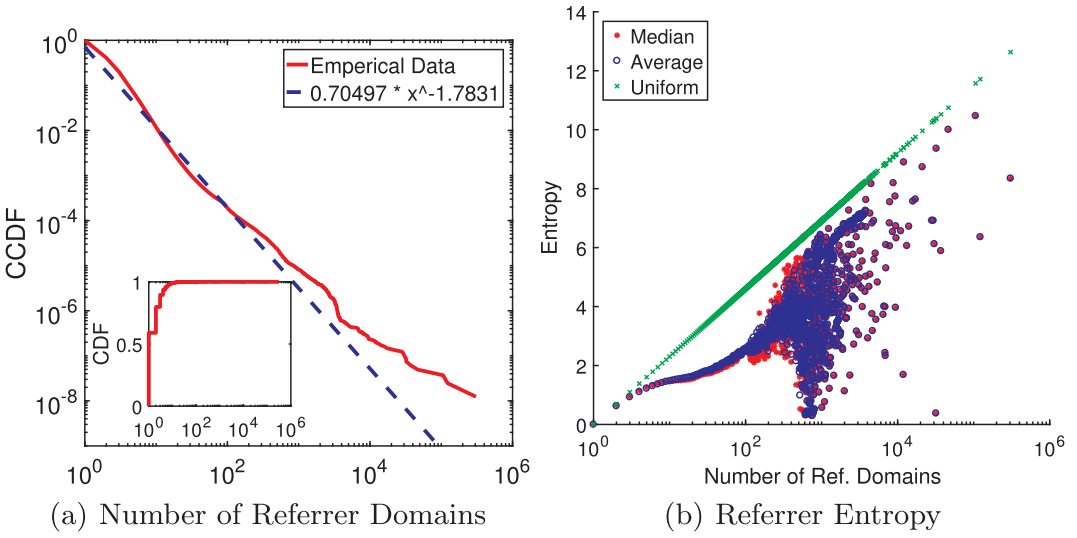


Fig. 5. The distributions of the numbers of referrer domains and referrer entropies are plotted as the number of referrer domains increases.

URLs accessed from a large number of referrer domains. We also plot the referrer entropy in Fig. 5(b), which quantifies how uniformly requests are distributed across the referrer domains, whose calculation is similar to Eq. (1). As shown in Fig. 5(b), the median and average values of the request entropies across referrer domains do not increase as much as those of the uniform case, meaning that most URL requests are generated in a few referrer domains.

We next examine the temporal characteristics of the requests to short URLs. To this end, we group the short URLs based on their creation dates and count the numbers of short URLs requested in our measurement period. Note that we describe the number of short URLs created in (i) whole period (Fig. 6(a)), and (ii) recent 1-month period (Fig. 6(b)).

As shown in Fig. 6(a), the number of short URLs increases as the short URLs are continuously created. Note that the numbers of short URLs created in June 2012 is 10 times larger than the ones for short URLs created in the previous month (i.e., May 2012), which implies that short URLs which are relatively recently created are likely to be requested. We also observe that thousands of URLs generated from 2009 to 2011 are still requested in June 2012. When we zoom in the creation time on June 2012, there are distinct temporal patterns between weekdays and weekend; the number of short URLs becomes higher for weekdays and lower for weekends. This is in line with the finding that more URLs are shortened in weekdays than weekends (Antoniades et al., 2011).

We also investigate how short URLs are requested from a geographical perspective. We observe the top 5 domains and categories in terms of number of requests for five representative countries (i.e., USA, Japan, China, Brazil, and GBR) where short URLs are mostly used. As shown in Table 6, the URL access patterns are different across the countries. The 'Shopping' URLs are mostly accessed in USA and China; URLs for the 'Social Network' category are actively requested in Japan and Brazil. Note that ' www.lepirata.com', a shopping site that sells football jerseys, ranks high in Brazil. When we look at the top 5 domains in USA and China, the CDNs for **Ebay** and **Taobao**, respectively, are the dominant contributors in URL requests. Interestingly, URL requests by Japanese users are likely to go toward localized social services such as ' amablo.jp' orblog.livedoor.jp. Note that people in GBR are likely to request news content, mostly created in ' bbc.co.uk', implying that BBC is a major news platform for the short URLs in Britain.

We note that there exist several 'malicious' or 'black' domains whose content pages are highly requested. For example, ' www.wmybuy.com', a gambling portal web pretending to be a shopping portal, is the most highly-requested domain in China. Also, URLs for 'Violence & Illegal' domains are highly requested in Brazil, and ' www.tufos.com.br', a site for 'Adult' content, is highly accessed.

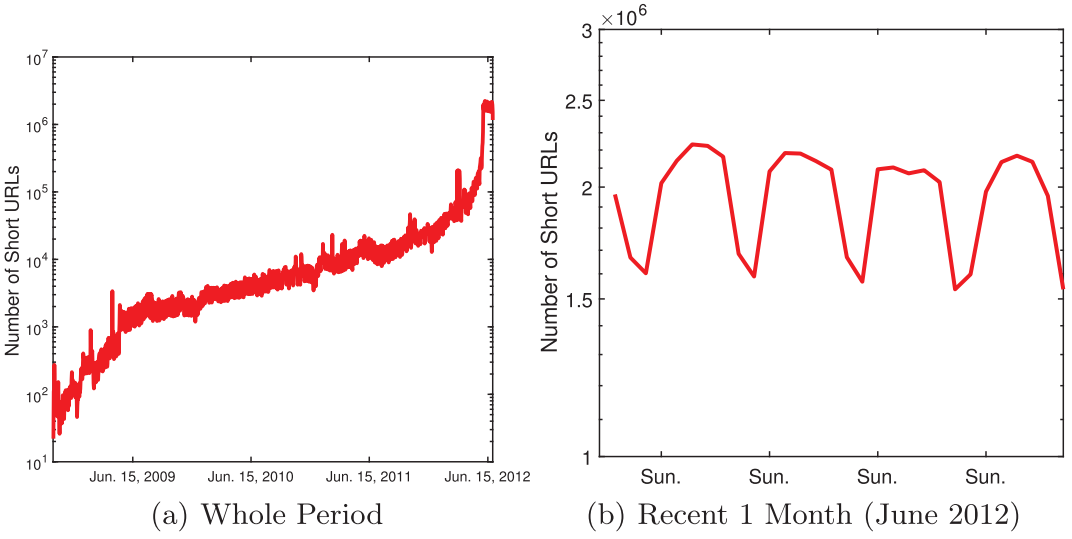


Fig. 6. The numbers of requested short URLs created in (a) whole period and (b) June 2012 are plotted.

4.2. Content-referrer graph

Based on the proposed graph model described in Section 4.2, we investigate (i) how different domains are associated with others, and (ii) which domains play important roles in publishing short URLs. Fig. 7 shows the distributions of weighted in-degrees, weighted out-degrees, and weights of edges, respectively. The portions of nodes that only have in-degrees (i.e., the 'referrer-only' domains) and out-degrees (i.e., 'content-only' domains) are 48.19% and 42.21%, respectively, indicating that a high portion of domains tend to be used as only either a referrer or a content source. Furthermore, around 15% and 23% of nodes have only 1 weighted out-degree and in-degree, respectively, while 5.8% and 2.7% of domains have more than 100 weighted out-degrees and in-degrees, respectively. This implies that a small number of domains play significant roles in publishing and sharing short URLs.

Fig. 7 also shows that in-degrees, out-degrees, and weights follow power-law with $(0.67949, -0.83296, x \geqslant 10)$, $(0.19429, -0.85657, x \geqslant 5)$ and $(0.62659, -1.1343)$ as parameters, respectively. Interestingly, the CCDF of in-degrees is below the fitting function and the gap becomes larger as the in-degree increases, implying that short URLs relatively less tend to be published in popular publishing spaces (i.e., referrer domains). Note that weighted in-degrees are larger than out-degrees in general, which indicates that content is generally shortened in fewer content domains and published in more referrer domains.

We next investigate how domains are linked amongst themselves in terms of the content-referrer relation. To this end, we first visualize the content-referrer graph, as shown in Fig. 8, to reveal relations among domains in a global view. Note that we plot only the top 0.1% edges in terms of the weight for the visualization purposes, and the sum of weights (i.e., total number of content URLs) in this graph is around 40 M, which accounts for 43.8% of the total weights. As shown in Fig. 8, the content-referrer graph mainly consists of two giant groups – `Facebook.com` and `t.co` (i.e., Twitter), meaning that both representative domains are heavily used to publish short URLs.

To reveal the heavy relations between domains in the content-referrer graph, we analyze the top 5 relations in the content-referrer graph in terms of number of short URLs (i.e., weight). As shown in Table 7, two representative OSNs, Facebook and Twitter, are the dominant referrer domains where short URLs are largely published. However, interestingly, we find that different content domains are likely to use the two OSNs as referral domains. The content URLs in foursquare.com and ameblo.jp tend to be published through Twitter (t.co), while app.facebook.com is likely to be published in Facebook. Note that feedproxy.google.com, an online newsfeed service, tends to use both Twitter and Facebook as primary referrers.

We next investigate how domain categories (or types) play roles in publishing short URLs. To this end, we first classify domains into twelve categories, as described in Section 4.1. Note that 'Adult or Malicious' is a set of the following five categories, which are linked to the malicious or adult content: (i) 'Parked', (ii) 'Spam', (iii) 'Phishing', (iv) 'Violence & Illegal', and (v) 'Adult', whose content URLs for these categories are mostly shortened for hiding the original URLs (e.g., adult content, spamming, etc.) (Nikiforakis et al., 2014; Gupta et al., 2014).

Fig. 9(a) shows the average weighted in-degrees and out-degrees for each (domain) category. Obviously, different categories play substantially different roles in the content-referrer graph. The average in- and out-degrees of both **Search Engine** and **Social Network** domains are higher than others mostly, implying that these domains play roles as both content sources and publishing spaces. Also, the average in-degree of **Computer** & **Electronics** domains are higher than its average out-degree, while the tendency is reversed in the case of **News** & **Media** and **Streaming**. This indicates that content URLs in **News** & **Media** and **Streaming** domains tend to be published in many referrer domains while content URLs from multiple content domains tend to be published in **Computer** & **Electronics** domains such as newsfeed services.

We next investigate how many domains (in different categories) play crucial roles in the content-referrer graph, by extracting the

Table 6
Top 5 domains and categories in terms of number of requests for five representative countries (i.e., USA, Japan, China, Brazil, and GBR) are summarized.

(a) Top 5 domains

Country	Domain	Portion of Requests for the Domain (%)
USA	rtm.ebaystatic.com	10.79
	mobile.ebay.com	3.96
	api.ning.com	2.34
	trib.al	1.75
	www.facebook.com	1.50
Japan	ameblo.jp	11.57
	www.amazon.co.jp	11.54
	cdn1.ustream.tv	10.98
	blog.livedoor.jp	2.51
	p.twipple.jp	2.12
China	www.wmybuy.com	15.38
	img01.taobaocdn.com	6.71
	img03.taobaocdn.com	6.56
	img02.taobaocdn.com	6.52
	img04.taobaocdn.com	5.88
Brazil	www.facebook.com	8.39
	www.lepirata.com	7.32
	www.faston.com.br	6.12
	www.tufos.com.br	2.45
	feedproxy.google.com	2.15
GBR	www.bbc.co.uk	4.05
	www.facebook.com	1.82
	dist1.terasoft.lt	1.78
	viper.w12.org	1.65
	api.ning.com	1.26

(b) Top 5 categories

Country	Category	Portion of Requests for the Category (%)
USA	Shopping	20.47
	News & Media	13.30
	Computer & Electronics	10.82
	Arts & Entertainment	10.67
	Business & Industry	6.17
Japan	Social Network	22.04
	Shopping	15.96
	Streaming	13.89
	Computer & Electronics	7.87
	Search Engine	4.39
China	Shopping	35.95
	Computer & Electronics	21.70
	File Sharing	8.64
	Business Services	4.21
	Games	3.93
Brazil	Social Network	12.83
	Violence & Illegal	10.68
	Shopping	10.27
	Arts & Entertainment	8.27
	File Sharing	7.73
GBR	News & Media	17.74
	Arts & Entertainment	12.57
	Computer & Electronics	11.65
	Sports	7.21
	Business & Industry	5.94

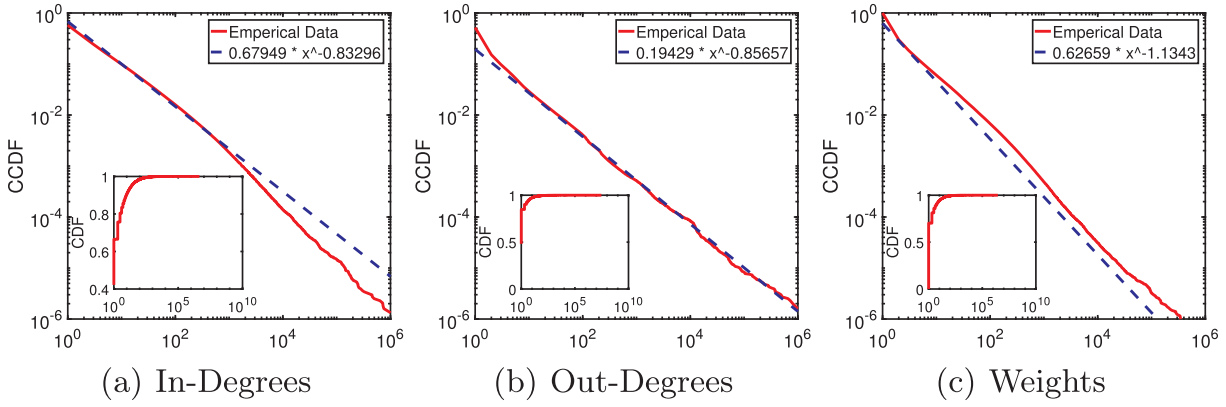


Fig. 7. The distributions of weighted in-degrees, out-degrees, and weights of the content-referrer graph are plotted.

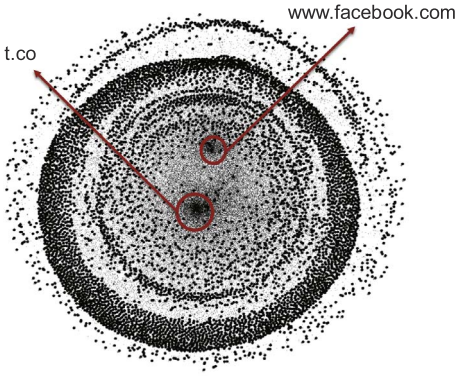


Fig. 8. The content-referrer graph is plotted. Only the top 0.1% relations of the content-referrer graph in terms of weight are shown for visualization purposes.

Table 7
Top 5 relations in terms of weight are listed.

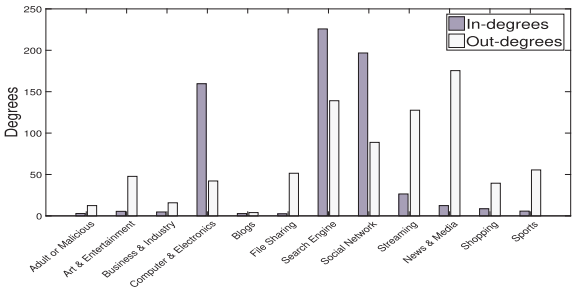
Content Domain	Referrer Domain	Portion of URLs (%)
foursquare.com	t.co	3.16%
ameblo.jp	t.co	1.02%
apps.facebook.com	www.facebook.com	0.83%
feedproxy.google.com	t.co	0.80%
feedproxy.google.com	www.facebook.com	0.64%

top 0.1% of domains in terms of weighted in- and out-degrees. We calculate the relative ratio of the top 0.1% domains in each category. That is, if there are a thousand of domains whose category is A and three of them are in the top 0.1% of all the domains in terms of weighted in-degrees, then the relative ratio is $3/(1K*0.001) = 3$, which indicates that there are three times more domains whose category is A in the top 0.1% list compared to the other categories.

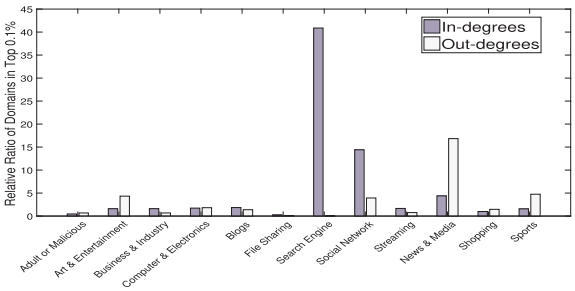
As shown in Fig. 9(b), the relative ratio of in-degrees of **Computer** & **Electronics** is almost zero while the relative ratios for **Search Engine** and **Social Network** are significantly high (42 and 14, respectively). Considering that average weighted in-degrees of the two categories are higher than others (as shown in Fig. 9(a)), this implies that substantial numbers of the domains of the two categories play crucial roles as publishing spaces. Note that only a few **Computer** & **Electronics** domains are used as heavy publishing sources while most of the domains in the **Computer** & **Electronics** are used only as referrers in general. Similarly, the relative ratios of out-degrees of **File Sharing**, **Search Engine**, and **Streaming** are low while their average out-degrees are high (as shown in Fig. 9(a)), meaning that only a small number of domains in these categories play important roles as publishing sources. Note that the relative ratio of out-degrees of **News** & **Media** are substantially high (16 times more than the average), implying that these domains typically play the role of content providers.

4.3. Referrer analysis

We examine how content pages are published and requested from different referral domains. Here, we focus on the three referrer



(a) Average Weighted In- and Out-Degrees across Category

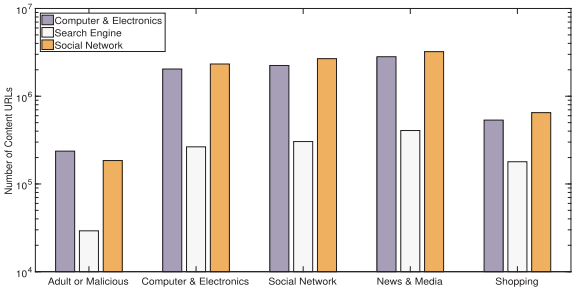


(b) Normalized Ratio of Domains in Top 0.1% by In- and Out-degrees

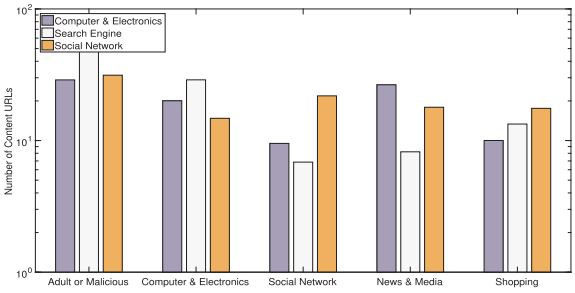
Fig. 9. The average weighted in- and out-degrees across category and the relative ratio of domains in top 0.1% by weighted in- and out-degrees are plotted.

domains, **Computer** & **Electronics**, **Search Engine**, and **Social Network** domains, each with high in-degrees (see Fig. 9(a)), which signifies significant roles in sharing content. In particular, we investigate how content URLs created in the domains of five representative categories (in terms of number of requests as shown in Table 4) are published in the above three referrer categories: (i) **Adult or Malicious**, (ii) **Computer** & **Electronics**, (iii) **Social Network**, (iv) **News** & **Media**, and (v) **Shopping**.

Fig. 10 shows the number of published content URLs (in the form of short URLs) and average number of requests for each content URL shared through **Computer** & **Electronics**, **Search Engine**, and **Social Network** categories. As shown in Fig. 10(a), content URLs



(a) Number of Content URLs



(b) Average Number of Requests

Fig. 10. The number of short URLs and average number of requests per short URL across the categories are shown.

created from the five categories are likely to be published through **Computer** & **Electronics** and **Social Network** referrer domains rather than **Search Engine**. This indicates that the content URLs created from the five categories tend to be published mostly through the **Computer** & **Electronics** and **Social Network** referrer domains, while the content URLs created from other categories (e.g., **Streaming**, **File Sharing**) are likely to be published through the **Search Engine** referrer domains. Note that we showed that **Search Engine** domains are heavily used as the publishing spaces of short URLs (see Section ??).

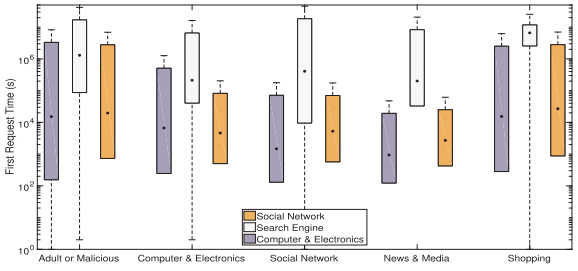
Interestingly, when we look at Fig. 10(b), content access patterns are disparate from content publishing ones (see Fig. 10(a)). For example, the average number of requests of the content URLs for **Adult or Malicious** and **Computer** & **Electronics** published through **Search Engine** referrer domains is higher than the ones through other referrer domains. That is, users tend to access **Adult or Malicious** and **Computer** & **Electronics** content through the **Search Engine** referrer domains rather than the **Computer** & **Electronics** or **Social Network** referrer domains. The **Social Network** and **Shopping** content pages tend to be requested more through the **Social Network** referrer domains, which indicates that people interested in **Social Network** and **Shopping** are likely to request such content though in **Social Network** domains. The **News** & **Media** content pages are largely requested through the **Computer** & **Electronics** referrer domains, meaning that the **Computer** & **Electronics** domains are major channels for **News** & **Media** content.

In summary, three popular referrer domains (as publishing spaces) play different roles in sharing content from different categories. In other words, there exist effective spaces (i.e., referrer domains) that can attract users' requests for different content types, which sheds important insights for content publishers who wish to elicit more user responses or attentions.

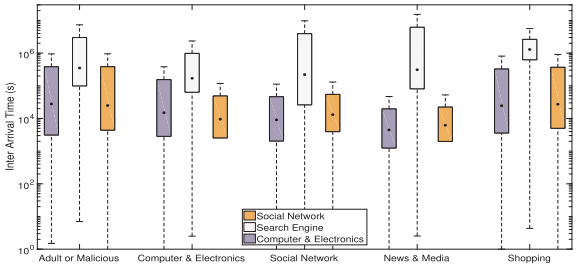
We next investigate the access patterns of content URLs through the three referrer domains from a temporal perspective. To this end, we measure two metrics which reflect user responsiveness to content: (i) *first request time* of a URL as the time difference between the URL creation time and its first requested time, and (ii) *inter arrival time* of a URL, which is defined as the average time between two consecutive requests for the URL from the first request to the last request. Note that we take into account only URLs that are requested at least twice.

As shown in Fig. 11, the first request time and inter arrival time of content URLs (in the five categories) are different across different referrer domains. That is, users' responses are different temporally across different publishing spaces. Overall, both the first request time and inter arrival time of content URLs published in the **Search Engine** referrer domains are higher than those published through the **Computer** & **Electronics** and **Social Network** referrer domains. Note that the gaps between the **Search Engine** and other referrer domains become relatively larger for the **News** & **Media** and **Shopping** content than other content, which indicates that, if news or shopping content pages are published through **Computer** & **Electronics** and **Social Network**, users tend to access the content quickly and virally.

Fig. 11 also reveals that user access patterns for different content categories are various even though they are published in the same referrer domains. For example, as shown in Fig. 11(a), the first request times of **News** & **Media** and **Social Network** content published through the **Computer** & **Electronics** and **Social Network** referrer domains are lower than those of other content categories. This implies that news or SNS-related content tend to be requested quickly. Note that the lengths of boxes (i.e., range from 25% to 75%) of the **Adult or Malicious** and **Shopping** content published in **Computer** & **Electronics** and **Social Network** referrer domains are longer than others, implying that the users' first responses spread more in these cases.



(a) First Request Time (s)



(b) Inter Arrival Time of Requests

Fig. 11. First request time and inter arrival time for five categories are plotted.

5. Conclusions

We comprehensively studied the practice of using short URLs from their creations to publishing to sharing. Using the dataset that contains 4.2 B requests for 80 M URLs created through Bit.ly, we analyzed (i) how content is shortened and shared depending on its types, (ii) how and where short URLs are published, and (iii) how content is shared through different publishing spaces. We found that content URLs are primarily shortened through the third party companies, such as Twitter, rather than individual users. We also found that the URL shortening practice and access patterns are disparate. For example, while there are not so many short URLs for the shopping and adult content, they are likely to be requested more. Our analysis on the content-referrer graph reveals that different domains play different roles in publishing short URLs. For example, search engines, OSNs, and computer & electronics are popular domains for content publishing, while news and streaming domains are widely used as content sources. We further revealed that users are likely to access different content types through different web sites; e.g., adult or malicious content tend to be requested from search engine domains; shopping content is primarily accessed through OSNs; and news are usually clicked through computer & electronics domains. We plan to take a machine-learning approach to build a prediction model for content-referrer relations.

Declaration of interests

None

Acknowledgement

This work was supported in part by the National Research Foundation of Korea through PF Class Heterogeneous High Performance Computer Development (NRF-2016M3C4A7952587) and the research fund of Hanyang University (HY-2017-N). The ICT at Seoul National University provides research facilities for this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttp://dx.doi.org/10.1016/j.tele.2018.03.003.

References

Antoniades, D., Polakis, I., Athanasopoulos, E., Ioannidis, S., Markatos, E.P., Karagiannis, T., 2011. we.b: The web of short urls. In: Proceedings of the 20th International World Wide Web Conference (WWW 2011).

Bakshy, E., Rosenn, I., Marlow, C., Adamic, L., 2012. The role of social networks in information diffusion. In: Proceedings of the 21st International World Wide Web Conference (WWW 2012).

Barabasi, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286.

Campbell, J., 2012. Top 12 websites in the world – then and now: 2012, 2007, 2001.https://www.socialtalent.com/blog/recruitment/top-12-websites-in-the-world-then-and-now-2012-2007-2001. Online; accessed 20-Oct-2016.

Cha, M., Mislove, A., Gummadi, K.P., 2009. A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th International World Wide Web Conference (WWW 2009).

Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J., 2014. Can cascades be predicted? In: Proceedings of the 23rd International Conference on World Wide Web Conference (WWW 2014).

Chhabra, S., Aggarwal, A., Benevenuto, F., Kumaraguru, P., 2011. Phi.sh/ocial: The phishing landscape through short urls. In: Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference.

Choi, D., Han, J., Chung, T., Ahn, Y.Y., Chun, B.G., Kwon, T.T., 2015. Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In: Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN 2015).

DGTraffic, 2012. Indonesia internet users.http://www.dgtraffic.com/indonesia-internet-users/. Online; accessed 20-Oct-2016.

Gelley, B., John, A., 2015. Do i need to follow you?: Examining the utility of the pinterest follow mechanism. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work (CSCW 2014).

Goel, S., Watts, D.J., Goldstein, D.G., 2012. The structure of online diffusion networks. In: ACM Conference on Electronic Commerce (EC 2012).

Gómez, V., Kaltenbrunner, A., López, V., 2008. Statistical analysis of the social network and discussion threads in slashdot. In: Proceedings of the 17th International Conference on World Wide Web Conference (WWW 2008).

Gupta, N., Aggarwal, A., Kumaraguru, P., 2014. bit.ly/malicious: Deep dive into short URL based e-crime detection. CoRR abs/1406.3687.

Han, J., Choi, D., Choi, A.Y., Choi, J., Chung, T., Kwon, T.T., Rha, J.Y., Chuah, C.N., 2015. Sharing topics in pinterest: Understanding content creation and diffusion behaviors. In: Proceedings of the 2015 ACM on Conference on Online Social Networks (COSN 2015).

Han, J., Choi, D., Chun, B.G., Kwon, T., Kim, H.C., Choi, Y., 2014. Collecting, organizing, and sharing pins in pinterest: Interest-driven or social-driven? In: Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS 2014).

Klien, F., Strohmaier, M., 2012. Short links under attack: geographical analysis of spam in a url shortener network. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT 2012).

Kumar, R., Mahdian, M., McGlohon, M., 2010. Dynamics of conversations. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010).

Maggi, F., Frossi, A., Zanero, S., Stringhini, G., Stone-Gross, B., Kruegel, C., Vigna, G., 2013. Two years of short urls internet measurement: Security threats and countermeasures. In: Proceedings of the 22nd International Conference on World Wide Web (WWW 2013).

Marcoccia, M., 2004. On-line polylogues: conversation structure and participation framework in internet newsgroups. J. Pragmatics 36, 115–145.

Newman, A., 2014. Bitly helps the red cross get to hope.ly – 2014.https://www.nytimes.com/2014/12/02/business/media/bitly-helps-the-red-cross-get-to-hopely.html?r = 0.

Nikiforakis, N., Maggi, F., Stringhini, G., Rafique, M.Z., Joosen, W., Kruegel, C., Piessens, F., Vigna, G., Zanero, S., 2014. Stranger danger: exploring the ecosystem of ad-based url shortening services. In: Proceedings of the 23rd International Conference on World Wide Web Conference (WWW 2014).

Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K., Almeida, V., 2011. On word-of-mouth based discovery of the web. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC 2011).

Wang, C., Ye, M., Huberman, B.A., 2012. From user comments to on-line conversations. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2012).

Wang, D., Navathe, S.B., Liu, L., Irani, D., Tamersoy, A., Pu, C., 2013. Click traffic analysis of short url spam on twitter. In: Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing.