Short Links Under Attack: Geographical Analysis of Spam in a URL Shortener Network

Florian Klien
Graz University of Technology
klien@student.tugraz.at
f@qr.cx

Markus Strohmaier
Knowledge Management Institute
Graz University of Technology and Know-Center
markus.strohmaier@tugraz.at

ABSTRACT

URL shortener services today have come to play an important role in our social media landscape. They direct user attention and disseminate information in online social media such as Twitter or Facebook. Shortener services typically provide short URLs in exchange for long URLs. These short URLs can then be shared and diffused by users via online social media, e-mail or other forms of electronic communication. When another user clicks on the shortened URL, she will be redirected to the underlying long URL. Shortened URLs can serve many legitimate purposes, such as click tracking, but can also serve illicit behavior such as fraud, deceit and spam. Although usage of URL shortener services today is ubiquituous, our research community knows little about how exactly these services are used and what purposes they serve. In this paper, we study usage logs of a URL shortener service that has been operated by our group for more than a year. We expose the extent of spamming taking place in our logs, and provide first insights into the planetary-scale of this problem. Our results are relevant for researchers and engineers interested in understanding the emerging phenomenon and dangers of spamming via URL shortener services.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*System issues*

General Terms

Measurement, Experimentation

Keywords

URL Shortener, link analysis, spam

1. INTRODUCTION

URL Shortener services have been available for at least 10 years [1]. Such services i) take a long URL as an input, ii) offer a short URL in return, and iii) permanently redirect traffic from the short URL to the long URL. One of the first URL shortener services was TinyURL, founded in 2002 [2]. The idea of shortening URLs originated from problems observed with links in emails, where links were often re-wrapped by clients and thus rendered unclickable. However, it has not been until recently that URL shortener services have gained popularity through online social media such as Twitter where space is limited. Although URL shortener services can be used for a multitude of different purposes including link tracking, click analysis or spam, our research community knows little about how exactly these services are used and what purposes they serve. In this paper, we study usage logs of a URL shortener service that has been operated by our group for more than 20 months. During this time, the service has *shortened* more than one million URLs and has *resolved* more than nine million links. At the same time, it has attracted significant attention of spammers. In this work, we expose the extent of spamming taking place on our URL shortener service, and provide first insights into the national and international scale of this problem.

Based on usage log data provided by our URL shortener service, this short paper addresses the following research questions: (i) What is the extent of spam in URL shortener services? (ii) Does usage of URL shortener services differ across countries, and if yes, in what way? (iii) Do shortened URLs “travel” across countries, and if yes, what is the nature of interaction between countries? and (iv) What are promising features for identifying spam in URL shortener services?

While the analysis presented in this paper is based on one dataset only, we believe that our results are interesting to the Hypertext community for at least two reasons: First, they provide unique insights into the use and misuse of a URL shortener service that was operated over a period of more than 20 months. Second, other URL shortener services such as bit.ly do not share usage logs that can be studied openly. Our work therefore provides a rare view into the operation of such services over a significant period of time. We leave the task of applying our analysis to other datasets to future research.

2. RELATED WORK

In [10], Inoue et al. present a study of a URL Shortener that was build right after the Great East Japan Earthquake in March 2011. The authors of this paper use a combination of a CDN (Content Delivery Network) and a URL Shortener to prevent server overload on heavily visited websites. By using CoralCDN, they distribute requests on servers around the world [5]. CoralCDN works on basis of DNS and HTTP mirrors. One can divert the traffic from a website to the mirrors of CoralCDN by inserting “.nyud.net” after the original domain name. The resulting address serves the website via the CoralCDN mirrors. Inoue et al. use the URL Shortener to

HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA.
Copyright 2012 ACM 978-1-4503-1335-3/12/06 ...$10.00.

rewrite long URLs to be *coralized* and thus redirect users to mirrors of the CoralCDN. They analyzed the content users shortened with their own shortener. Their data set is available on-line [3]. Antoniades et al. analyzed a data set of short URLs they obtained through crawling [6]. They crawled Twitter for short URLs of *bit.ly* and *ow.ly*. Further they guessed short URLs via brute forcing hashes. Their results show that the lifetime of 50% of short links exceeds 100 days. Further they investigated the latency of short URL services and show that most request are delayed by about 0.35 seconds. Grier et al. point out that many spammers use shortener services to obfuscate their links in tweets. They find that blacklisting URLs is no optimal solution for fighting spam on Twitter since blacklists often lag Twitter and a spammer's link often reaches the public before it can be blacklisted [9]. Our work is related to this work in a sense that we will present promising features for spam classification that might be helpful to fight spam. Leskovec and Horovitz examined a very large graph of a personal messaging tool. It consisted of millions of links between millions of individuals, where every edge represented one conversation or chat. Users were located all around the world and the resulting connections that could be derived between countries and people led to the insights that most connections seemed to be between countries that had ethnic or historical connections: Germany - Turkey, Portugal - Brazil and China - Korea. We too will focus on the connections that evolve between countries via short URLs [11]. Chhabra et al. did a study concerning the distribution of phishing links across Twitter. Their research analyzes the connection between users and their vulnerability to click malicious links. Further they look into the geographical distribution and the lifetime of phishing URLs [8]. In our own previous work, we have studied the behavior of bots in social networks [12]. Our work builds on and relates to the state-of-the-art by conducting geographical analysis of spam in a usage log created by a URL shortener service that we have operated.

3. EXPERIMENTAL SETUP

We use data from the URL Shortener *qr.cx* that started operating in June 2009. It is a regular HTTP service with an API in place. A *creator* (a user who wants to shorten a URL) can send GET requests with a long URL as an argument and receive a short URL (e.g. `http://qr.cx/1r8`) as a response. Our URL shortener generates one random string as an identifier, which is appended to the base URL and becomes the short link or short URL. This short URL redirects the visitor via a 301 HTTP response to the previously provided long URL. In the following, we will use the term *resolve* whenever we refer to the function of returning a long URL in exchange for a short URL that has been requested by a *resolver*. Other terms that are sometimes used to denote this function include *redirect*, *click* or *expand*. Each provided long URL is only registered once with our service, it cannot be re-registered and the long URL cannot be changed. Unlike other popular URL shortener services like *bit.ly*, our service does not offer user accounts (for privacy concerns) and provides no possibility to get multiple short URLs for an already registered long URL.

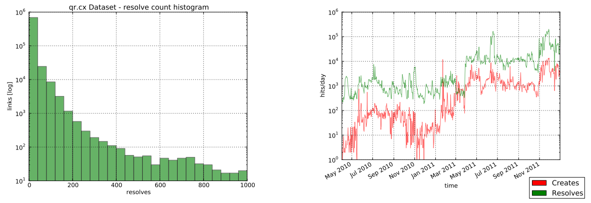
	Resolves	Creates	Sum
complete data set	7,919,892	731,624	8,651,516
with location	7,918,221	731,228	8,649,449

Table 1: Data set characteristics: The observation period ranges from 1st of April 2010 to 31st of December 2011. For our evaluation we generated a subset of this data that contains location information.

Our data set ranges from 1st of April 2010 to 31st of December 2011 and contains 7,919,892 *resolves* of short URLs and 731,624 *created* short URLs (see Table 1). A subset of this data contains geographical longitude and latitude information for users which we obtained from geo-locating their IP. There are some limitations and biases in our dataset.

Given that the URL shortener service was operated from within Austria, there might be a local social influence. Furthermore, popular online social media are often hosted in US-based territories, which represents another source of bias. In addition, these services sometimes use bots to *resolve* short URLs and explore the mentioned content. Other biases are possible (e.g. with regard to users' preferences with regard to certain shortener services). In general however, we believe that - based on our comprehensive logs - our URL shortener service reflects certain characteristics of URL shortener services in general.

Figure 1a shows a histogram of *resolves* for our data set and Figure 1b plots the number of *creates* and *resolves* over our observation period. *Resolves* and *creates* correlate strongly in the second half of 2011. The traffic for URL *creates* has increased by two orders of magnitude between December 2010 and December 2011. In the same time period, *resolves* have increased by a factor of about 25. There exist 51,675 links that have only been *resolved* once and one link that has been *resolved* around 30,000 times. A spam wave hit our service in April 2011. This is depicted in Figure 1b, and in a video visualization of our data that we have made available [4].



(a) *Resolve* Histogram

(b) *Resolves* and *creates* over time.

Figure 1: 1a shows the *Resolve* histogram for the data set including *resolves* up to a limit of 1000. The y-axis shows a logarithmic scale count of links, the x-axis shows the number of *resolves*. 1b: *Resolves* per day are depicted in green (upper line), *creates* per day are depicted in red (lower line). An increase of both *creates* and *resolves* can be observed for April 2011. From a deeper look at our logs, we find that this increase was caused by a spam wave that hit the URL Shortener service at that time.

3.1 The URL Shortener Network

For our following analysis, we define a URL shortener network as a weighted directed network of users, where *resolves* correspond to edges between *creators* and *resolvers*. The number of *resolves* observed corresponds to the weight of the edge. This network gives us the possibility to study network-theoretic characteristics of URL shortener networks. To reduce the number of nodes in the network, one can group users by IP address, IP subnet, country, region or provider. By grouping IPs in different ways, analysis can focus on different levels of granularity and different aspects depending on the data set. In the following, our analysis groups users by country in order to enable us to do geographical analysis of URL shortener service usage on a planetary scale.

	Country	Link Creates	**Resolves**	IRR	RC Ratio	Outdegree	Indegree	Resolver Pct.	Creator Pct.
1	**US**	81,341	6,250,743	**79.82%**	**98.72%**	148,264	6,209,235	**97.67%**	2.33%
2	**GB**	2,409	699,804	**7.19%**	**99.66%**	33,869	696,590	**95.36%**	4.64%
3	DE	6,544	357,036	22.16%	98.20%	69,520	352,607	83.53%	16.47%
4	RU	7,599	108,996	3.04%	93.48%	119,712	103,572	46.39%	53.61%
5	JP	8,015	102,979	0.75%	92.78%	111,082	101,433	47.73%	52.27%
6	KR	8,967	50,679	1.63%	84.97%	126,600	50,629	28.57%	71.43%
7	FR	26,272	43,886	1.84%	62.55%	179,489	43,613	19.55%	80.45%
8	CA	1,155	34,263	12.4%	96.74%	16,817	33,687	66.70%	33.30%
9	NL	1,686	32,454	2.88%	95.06%	21,185	32,244	60.35%	39.65%
10	CN	907	30,626	17.67%	97.12%	7,705	29,155	79.10%	20.90%
11	GR	1,196	16,293	0.05%	93.16%	15,295	16,243	51.50%	48.50%
12	**IN**	31,798	15,620	**0.54%**	**32.94%**	436,535	14,761	**3.27%**	96.73%
13	IE	283	14,930	0.02%	98.14%	3,390	14,877	81.44%	18.56%
14	AU	278	11,877	2.64%	97.71%	4,408	11,532	72.35%	27.65%
15	AT	1,120	9,580	34.99%	89.53%	75,869	9,428	11.05%	88.95%

Table 2: Top 15 Countries by *resolves*. Note: The Indegree would be the same as the *resolves* column, if the data set would not be limited to the observation period between 04/2010-12/2011. The column Indegree has smaller values because URL creations before 04/2010 have no known creation country and are ignored. India, ranked at 12th place, shows an interesting pattern of link *creates* to *resolves*, where *creates* outnumber *resolves* twice. The U.S. and Great Britain show far more *resolves* than *creates*.

3.2 Metrics

To characterize different countries and their URL Shortener usage, we present and use the following measures: The ratio between *resolves* and *creates*, the *Internal Resolve Rate*, the *Resolver Percentage* and the *Creator Percentage* as the ratio between Indegree and Outdegree [13].

The *RC Ratio* between *resolves* and *creates* tells us if a particular group (in this paper: a country) visited more links than it *created.*

$$RC\ Ratio = \frac{\#\ of\ Resolves}{(\#\ of\ Creates + \#\ of\ Resolves)} \tag{1}$$

This ratio models *resolves* and *creates* as percentage. 100% *RC Ratio* means the group has no link *creates*. If it is lower than 50% it means the group *resolved* fewer links than it *created*. This shows the group as a whole "sends out" more links than it uses (*resolves*). A variation of this metric would be to only count *resolves* of self-*created* short URLs.

The *Internal Resolve Rate* (*IRR*) is the percentage of link *resolves* by a group that were created by themselves:

$$Internal\ Resolve\ Rate\ (IRR) = \frac{\#\ of\ Resolves\ by\ Group}{\#\ of\ all\ Resolves} \tag{2}$$

The ratios between Indegree and Outdegree show the "consumption" or "broadcasting" activity of a group. In this work, the Indegree is defined as the cumulative *resolve* count of a group. The Outdegree is defined as the cumulative *resolve* count of links *created* by that group. Whenever someone *resolves* a link, the creator's Outdegree is increased by one. The

$$Resolver\ Percentage = \frac{Indegree}{(Indegree + Outdegree)} \tag{3}$$

shows how much a group consumes and the

$$Creator\ Percentage = \frac{Outdegree}{(Indegree + Outdegree)} \tag{4}$$

shows how much a group "broadcasts". The Resolver Percentage differs from the Internal Resolve Rate as it does count link *resolves* that were not *created* locally.

4. RESULTS

In this section, we organize the presentation of our results in correspondence to the research questions introduced before:

4.1 What is the extent of spam?

We annotated a random sample of 5,957 shortened URLs by visiting the long URL and evaluating its content.

Our annotated set contained 4,780 spam and 1,177 non-spam links. This results in a spam rate of 80.24%. We also evaluated whether *creators* had *resolved* their link themselves. The resulting sub set can be seen in Table 3.

	self resolved	not self resolved	Sum
spam	17	4,763	4,780
non-spam	18	1,159	1,177
Sum	35	5,922	**5,957**

Table 3: The annotated sub set of our data set. 80.24% of URLs are labelled as spam.

To visualize temporal and geographical aspects of this data, we plotted all redirects and link creations on a world map for a given time frame, and then assembled the different frames in sequential order into a video that covers the 20 months observation period. Binning our data in frames with a window of 6 hours resulted in a 1 min 40 seconds video, which we have made available for viewing online: `http://youtu.be/06Mhn0L23Tk` [4]. In the video, yellow dots represent *resolves* and pink dots *creates*. Especially noteworthy are "waves" of URL creations around minute 0:57. This nicely illustrates how spammers use bot nets to *create* a lot of URLs in a very short time. Links are *created* all over the world but are *resolved* mostly in the U.S. and Europe, which indicates that the primary targets of spammers are US and Europe based populations or services. The video also suggests that the number of *resolves* in spam attacks using URL shortener services can likely be used as a proxy measure for gauging the success of a spam wave.

4.2 Does usage of URL shorteners differ across countries, and if yes, in what way?

Answering this question would give us insights into the local usage of URL shorteners. Details on usage can potentially be used for tackling a number of problems including marketing, infrastructural planning or abuse detection.

We found that the usage differs significantly between different countries. First, absolute *resolve* counts differ a lot. The U.S. has almost 10 times more *resolves* than the second biggest "consumer", Great Britain, and 200 times more than the 10th place China. In

Table 2 the top 15 Countries by *resolves* are listed in more detail. Second, relative usage numbers differ too. The U.S. has a total *resolve* count of 6,250,743 and a total link *create* count of 81,341. India has a total *create* count of 31,798 short links and a *resolve* count of only 15,620, as seen in Table 2. We use the measures from Section 3.2 to look into these numbers more deeply.

As shown in Figure 4, one can see that a lot of countries *resolve* more links than they *create* (green) but even more *create* more links than they *resolve*, as seen in red. Generally put, a high Outdegree seems to be indicative of creating nations (which might be linked to spamming) and a high Indegree seems to be indicative of spam-receiving countries (the targets of spam campaigns). Looking at the *IRR* of countries we see very different usage types of countries. We picked two countries that differ significantly with regard to usage. For example, our dataset produces a 100% "export" ratio or 0% *IRR* for India. This high external resolve ratio is significant as there are more than 30,000 links that were *created* in India. At the same time, the U.S. has the highest *IRR* distribution. The numbers show that more than half of the links *created* in the U.S. are *resolved* in the U.S.

In addition, we compared countries based on their In- and Outdegree (cf. equations 3 and 4). Taking the absolute values and plotting them in a scatter plot (Figure 2) reveals drastic differences in different countries' usage. Kazakhstan has an extremely high Outdegree compared to it's Indegree. Mexico and Thailand have the highest Outdegree but do not match Germany, Great Britain or the U.S. based on their Indegree.

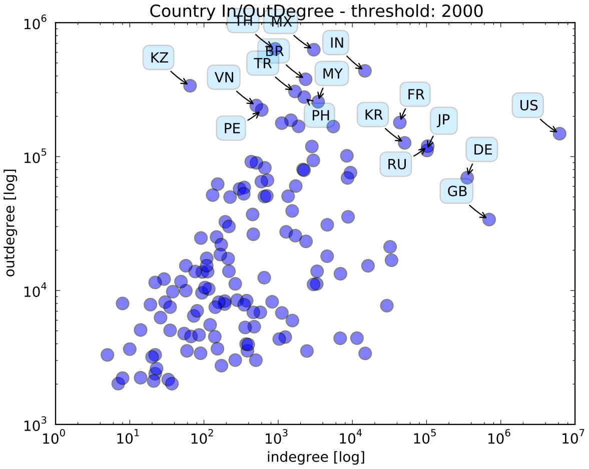


Figure 2: Scatter plot of countries by Indegree and Outdegree on a logarithmic scale. The U.S. has the highest Indegree and Mexico and Thailand have high Outdegrees.

4.3 Do shortened URLs "travel" across countries, and if yes, what is the nature of interaction between different countries?

In order to "travel" across countries, we require a URL to be *created* in country *A* and be *resolved* in country *B*. We consider this analysis to be of interest as it might be useful to distinguish locally-relevant from internationally relevant content or identify international flows of information and diffusion.

In Figure 3, we illustrate directionality of *resolves* in a graph of the top 15 countries from Table 2. *Resolves* are edges that start from a *creating* country and point to a *resolving* country. Edges are weighted according to the number of short URLs that have been *resolved* between two countries. The graph only shows the top 15 countries ranked by *resolves* in our data set and starts displaying edges with a weight higher than 5,000. From this graph, we can observe that the U.S. dominates the consumption of short links in our data set. We also noticed that there are only four countries that appear to be *resolving* their own links (given our thresholds). India accounts for the strongest edge to the U.S. with 340,000 *resolves*.

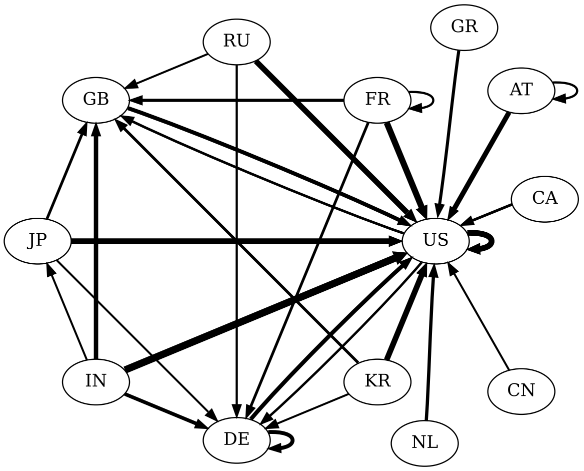


Figure 3: The international flow of shortened URLs: Edges go from countries where links have been *created* to countries that *resolve* them. Edges' widths correspond to logarithms of their *resolve* count. Smallest edge weight is 5,000, biggest 340,000.

4.4 What are promising features for spam identification?

While presenting and evaluating a spam classification method is beyond the scope of this paper, we want to explore the usefulness of initial features for spam identification. While potential features include but are not limited to content-based or behavior-based features, in the following we are interested in behavioral features as they can be assumed to work across languages and work independant of the type of content (e.g. textual vs. images). We evaluate a simple hypothesis that states that spammers would be less likely to verify the shortened URLs or use them themselves. Evaluating a statistical independence between hypothetical non-spammers that *resolve* their links and spammers that do not *resolve* their links we find a significant difference. Using a random subset from our dataset, a little under 1% drawn from each month, we manually labled links to be spam or not (see Table 3). A Pearson's chi-squared test showed that the two variables (self *resolved*, spam) are not independent [7]: $\chi^2_{1,5957} = 20.3091$, $p < 10^{-5}$, this corroborates our intuition.

While these initial investigations are promising and relevant for future work on spam classification algorithms in the context of URL shortener services, more work is warranted.

5. DISCUSSION AND CONCLUSIONS

URL shortener services such as bit.ly and others play a critical role on the web today. While usage of these services is ubiquitous, we know little about how exactly these services are used, and what purposes they serve. The work in this paper exposes that spam represents a pressing problem both for operators and for users

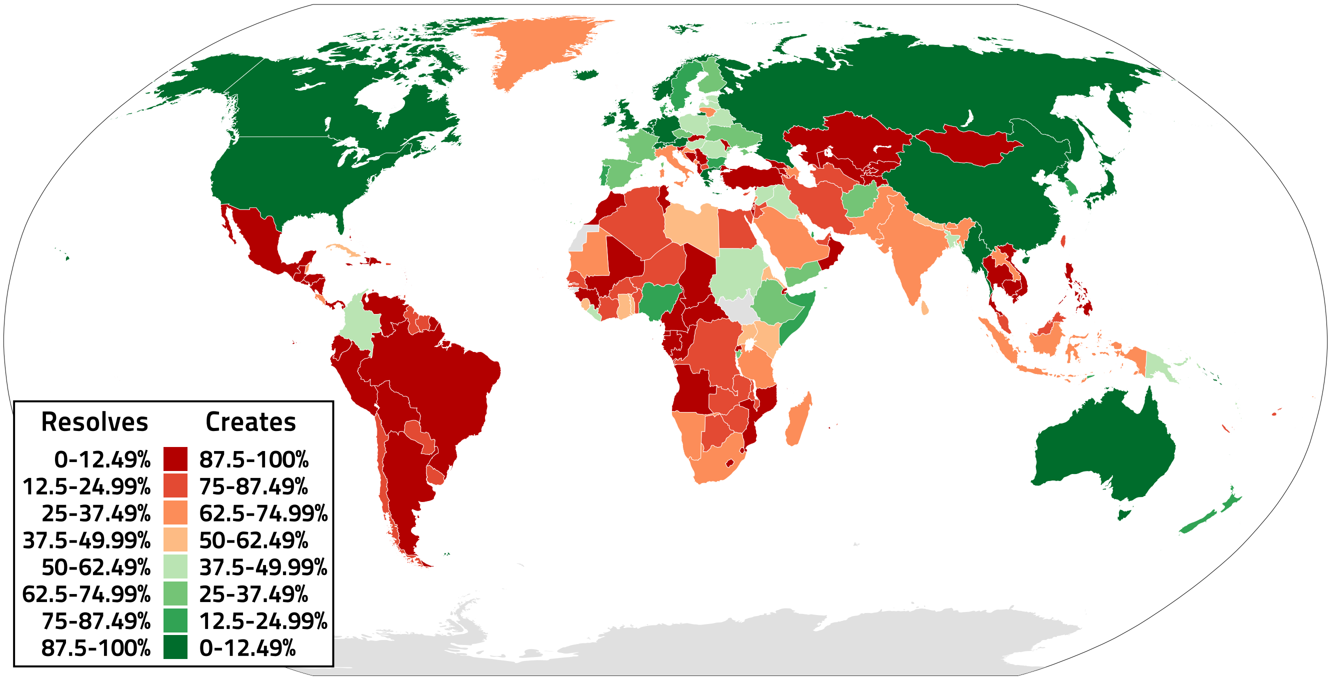


Figure 4: World map showing the ratio between *resolves* and *creates* (*RC Ratio*) by country, as shown in equation 1. Large parts of South America and Africa are identified as mostly *creators* (with small numbers of *resolves*) whereas Northern America, Asia, Australia and (to some extent) Europe are identifed as mostly *resolvers* (with small numbers of *creators*).

of URL shortener services. Based on our investigations of a URL shortener service, we find that around 80% of shortened URLs contained spam-related content. The extent of spam might be larger due to the lack of spam blocking features. Our geographical analysis reveals that this problem has an international scale, suggesting that URL shorteners play a role in spam attacks that cross national borders. The scale of this problem also suggests that in the future, sophisticated approaches and algorithm for identifying URL spam are needed. Our results indicate that different countries differ significantly with regard to the usage of URL shortener services. We find that imbalances between *creating* and *resolving* short URLs exist, and we have visualized the flow of *resolves* between countries based on URL shortener services. Our exploratory work provides first novel insights into global usage patterns of URL shortener services and warrants future research into understanding spam behavior in this new domain.

6. REFERENCES

[1] We want 'em shorter. `http://www.metafilter.com/8916/`, 2001.

[2] TinyURL Shortener. `http://tinyurl.com/`, 2002.

[3] rcdn data. `http://rcdn.info/data.html` - rcdn dataset, 2011.

[4] qr.cx usage time analysis video. `http://qr.cx/8Ctq` or `http://youtu.be/06Mhn0L23Tk`, 2012.

[5] The Coral Content Distribution Network. `http://www.coralCDN.org/`, 2012.

[6] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis. we.b: the web of short urls. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 715–724, New York, NY, USA, 2011. ACM.

[7] D. G. Bonett. Pearson chi-square estimator and test for log-linear models with expected frequencies subject to linear constraints. *Statistics & Probability Letters*, 8(2):175–177, June 1989.

[8] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/$ocial: the phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 92–101, New York, NY, USA, 2011. ACM.

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

[10] T. Inoue, F. Toriumi, Y. Shirai, and S.-i. Minato. Great east japan earthquake viewed from a url shortener. In *Proceedings of the Special Workshop on Internet and Disasters*, SWID '11, pages 8:1–8:8, New York, NY, USA, 2011. ACM.

[11] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. *Proceeding of the 17th international conference on World Wide Web WWW 08*, 393:915, 2008.

[12] C. Wagner, S. Mitter, C. Koerner, and M. Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. In *Proceedings of the WWW'12 Workshop on 'Making Sense of Microposts' (MSM2012)*, 2012.

[13] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Number 8 in Structural analysis in the social sciences. Cambridge University Press, 1994.