Clusters and constellations: Untwisting shortened links

Emma Tonkin
UKOLN
c/o The Library, University of Bath, Bath
e.tonkin@ukoln.ac.uk

Heather D. Pfeiffer
New Mexico State University
c/o DACC, Business and Information Systems
hdp@cs.nmsu.edu

ABSTRACT
Frequent users of social services such as Twitter are now familiar with the use of URL shortening services to produce compressed versions of actionable URLs. Although the typical user motivation for these is often taken to be convenience – particularly in the matter of reducing the amount of space taken by a link during a tweet, these shortening services are also used for various other purposes, including the collection of analytics for marketing purposes. In this poster we present the initial findings from our analysis of 350,000 tweets from Twitter on a subject close to the hearts of many ASIS&T conference attendees – the TSA (transport security administration). These tweets, gathered over a period of six weeks, primarily collect together reactions to a number of events and announcements of both positively and negatively nature, and as such, contain a large number of encoded URLs. We show the result of back-tracking URLs to their origin, demonstrating that it is now commonplace for URLs to be redirected through more than one redirection service. From analysis of the shortening services used we demonstrate that the majority of shortened URLs make use of one of a very small number of services, although these may be identified via an alias. Finally, we discuss the implications of these findings, both in terms of preservation and our ability to access the context of older Twitter conversations, and in terms of the implications for developers of user applications or content analysis platforms.

Keywords
Twitter, Preservation, URL redirects, URL shortened links

INTRODUCTION
The increasing prominence of Twitter as a social site in the last years has led to a great deal of interest in the way in which the site is used, as well as the technical enablers that underlie that use. A particularly important tool for Twitter users in the past has been the URL shortener (Carmody, 2011) – a tool, often web based or built in to the application, used by the individual to post their remarks to Twitter. These are conceptually simple: a URL is provided to the shortening tool, which assigns to it a unique key; when presented with that key, the tool will then present the browser with some form of redirect (often a 301) to return the user to the original long URL.

Benefit: Shortened URLs
The primary benefit for Twitter users was simply that a shortened URL does not eat significantly into the limited space available for each tweet (Twitter's famous 140 characters or less), leaving the user with more space to present their own ideas or opinions. There are also secondary benefits, of course, such as relative opacity (i.e. it is not usually possible to guess at the destination of a shortened URL), making it possible for users to forward readers to unexpected URLs, providing potential for practical jokes and for malicious reuse as well as fulfilling the more general purpose of compressing information.

Rationale: Construction and Maintenance-relative Costs?
Since URL shorteners are not technically complicated, they are relatively easy to set up, and indeed a site that tracks URL shorteners has identified over a thousand individual services (Yi.tl, 2012). However, like many other such initiatives the attrition rate of URL shorteners over time appears to be quite high – according to yi.tl, the majority of shortening services identified have since closed. As we will discuss in this poster, the majority of shortened URLs from a given US-centric discourse during the spring of 2012 make use of one of a few major service providers, either directly or via aliases run by those providers. One clear advantage of making use of a URL shortener is the opportunity to gain information about the number of click-throughs – how many people accessed the link that was posted, when, and from which broad geographic region. This is particularly useful to those for whom the distribution of links in a given venue forms part of a marketing strategy – a group in which Higher Education institutions are increasingly likely to count themselves, as market forces penetrate ever more deeply.

This reasoning also leads the construction of URL shorteners in some domains – indeed, it is not uncommon for parent enterprises to sell social media analytics services or provide free or paid analytics services. Yet, as with sentiment analysis, much of this activity deals with short-term, transitory events. Such analysis is typically bound to a relatively brief timescale – a few hours to a few days. Little financial benefit may exist in long-term provision of a 'long tail' of older redirects.

PRESERVATION OF SHORTENED URLS

Shortened URLs, once identified, can (if the underlying service is still available), trivially be resolved into the original destination URL. This is a useful step for many forms of analysis (e.g. content/contextual analysis of tweets on Twitter). The half-life of social services is often short, but a URL shortener is more intimately bound into our ability to follow a conversation than, for example, a news aggregation service might be. The loss of the news aggregation service potentially compromises our ability to identify the trigger for transitory interest in a given subject or resource. The loss of the redirect service means that the key resources referenced during a conversation can no longer be referenced, compromising our ability to understand the social or political context and underlying framing of the discussion. URL redirection increasingly offers a further challenge, for although the number of discrete services in popular use appears to be reducing, the penetration of these services into the user experience continues to increase. Twitter itself did not initially impose the use of a domain redirection service. Later, the service began to 'wrap' popular (frequently retweeted/referenced) URLs into Twitter's own domain redirection service, t.co. In late 2011 Twitter made this mandatory for all URLs (dev.twitter.com, 2012); therefore, any URL published through the Twitter service will be published in the form of a t.co/key alias. Since users' choice of URL redirection service typically relates to their choice of application (for example, HootSuite users will find that they are minting ow.ly URLs, which are inbuilt), this means that a user making use of HootSuite will have a characteristic 'fingerprint': t.co → ow.ly (→ previous source of link).

There are many reasons to look into URL redirection other than preservation, such as the need to identify spam (Thomas et al, 2011), or an interest in conversation/ discourse analysis and information propagation (Rodrigues et al, 2011).

IN CHAINS: UNWRAPPING THE URL

The implementation of various services and applications leads to the 'wrapping' of existing URLs into one or more URL redirects. The effect is similar to taking a postcard, and placing it into an envelope addressed to the initial receiver care of an intermediary. Then that envelope is passed on to a courier service who insist on placing the mail into their own brand of envelope and addressing it to 'Original recipient, care of initial intermediary, care of the courier service's posting office'. By this means, each agency is able to collect statistics about visitors to that URL.

Short URL	Response code	Redirect	Chain ID
t.to/example	301	ow.ly/example	1
ow.ly/example	301	Bbc.in/example	1
Bbc.in/example	301	http://news.bbc.co.uk/example	1

Table 1: A sample HTTP response chain

For the user, this carries the penalty that URLs are both opaque and somewhat slower to resolve. It also implies that the user is providing considerable information about their interests and activities to each agency in the redirect chain. However, for the researcher at least, it provides us with additional information about the pathway that this information took on its way from the originator to the author of the tweet.

Backtracking the Trackers

A simple URL redirect tracker was developed for the purpose of tracking each step of URL redirection, using Perl's LWP libraries to extract information about each step of domain resolution. This 'traceroute' application is able to generate information about a shortened URL by backtracking through each step and documenting each redirect. A sample result is given (see Table 1).

Each of these intermediate references may also be in use elsewhere. For example, news.bbc.co.uk/example may also have a t.to/example identifier that does not pass through the intermediate stages. The result is that a series of different identifiers is minted with a similar endpoint (i.e. they could arguably be identified as sameAs, in RDF terms, although best practice states that the final destination should be used as the identifier), but that travel a network of different intermediate routes. An open research question is the fragility of longer chains of redirects. It would appear that longer chains of redirects are statistically more likely to break, but this conclusion follows only in the case that longer redirect chains are equally, or more likely, to include redirect services with a higher risk of downtime or de-commissioning.

Subject Matter

For the purposes of this analysis, we analyzed the subject of each destination website (i.e. the resource referred to by the user's intended destination link) by making use of DMOZ

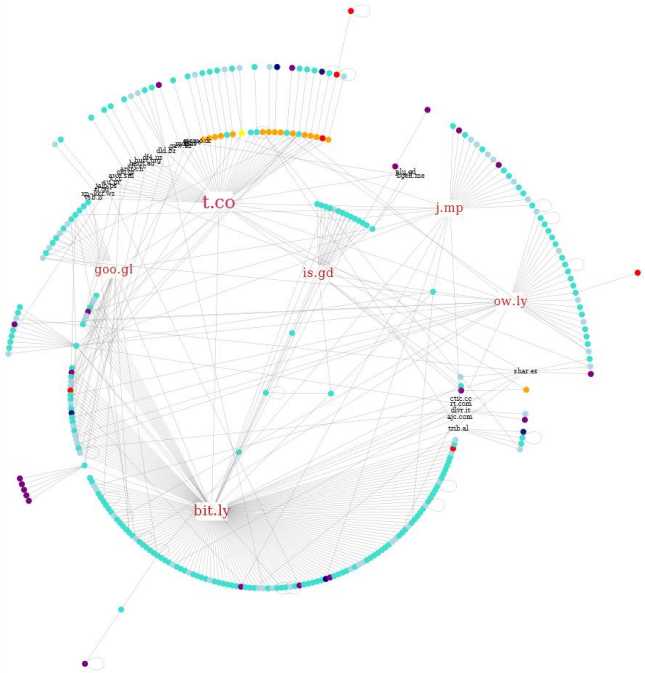

Figure 1: The TSA web of redirects

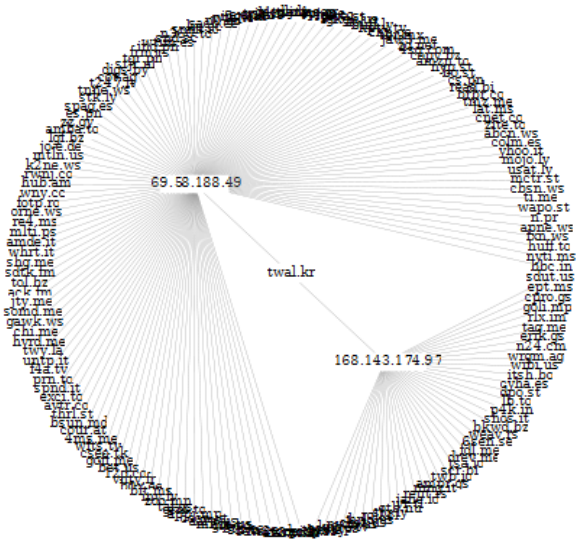

Figure 3: The bit.ly infrastructure constellation

classification data, as well as collecting information from the semantics implied by the URL itself (i.e. 'blogspot.com' or 'wordpress.com' extensions, or the use of the .gov TLD).

Internet Infrastructure

In order to determine the extent to which URL redirect services share infrastructure, we used a number of UNIX tools to collect information about the services themselves.

MAPPING THE REDIRECT WEB

Figure 1 demonstrates the initial findings from our analysis. 'Leaf node' sites (e.g. the sites to which the end-user intended to refer) appear towards the edge of the graph, and are color-coded according to function; purple dots represent blogs, whilst green dots represent news sites and red dots represent governmental resources – in the context of discussion, that of airport security, government resources represent an authoritative information resource.

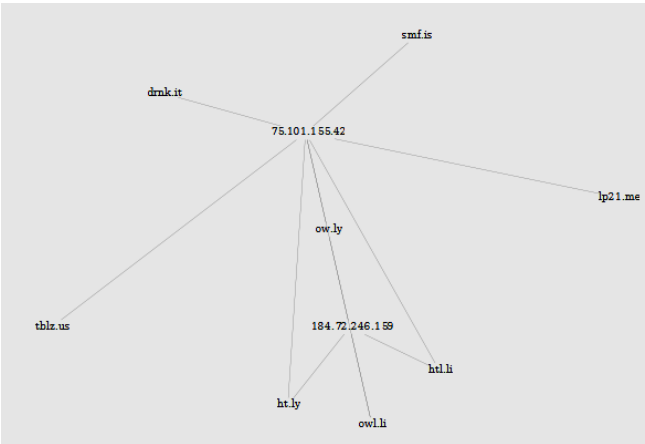

Figure 2: A simple constellation of redirect services

As can be seen, bit.ly redirects dominate the landscape in terms of variety and number of unique links, both directly and through vanity redirects provided by bit.ly for third-party organizations (color-coded in orange). Many other services also make use of redirects that appear to be provided in-house.

CONSTELLATIONS OF REDIRECTS

In Figure 2, the relationship between a number of redirect services and the server IPs responsible for fulfilling user requests, is shown. Whilst not infallible (it is not inconceivable for two unrelated services to use the same hosting, especially where large-scale support infrastructure is in use), this form of analysis is able to provide useful guidance as to the administrative ownership of domains.

To demonstrate once again the broad reach of the bit.ly service in this domain, Figure 3 shows an analogous constellation of services making use of bit.ly infrastructure.

STUDYING THE CONSTELLATIONS

Preservation of data from the social web may be automated, but where resources are scarce or bottlenecks exist (as in the case of redirect resolution), priorities may be set on the basis of available data (Day, 2003). One such data point is the predicted longevity of resolution services. The lifespan of social websites is often relatively short and services may be ephemeral, but it appears likely that identification of the sponsoring agency or agencies supporting the service may be instrumental in ascertaining the stability of that service.

In service constellations, there may, under some circumstances, prove to be strength in numbers.

That being said, we recommend the conceptual separation of social websites as platforms (Clemens et al, 2007) from content encoding employed within those sites, and the treatment of each of these layers in a manner consistent with its predicted active lifespan depicted by its constellation.

CONCLUSION

In conclusion, we have seen from the analysis of tweets referring to a US-centric subject – the TSA – that discourse on Twitter has a strong dependence on an apparently broad collection of redirect services. We have also seen that the resolution of URLs into resources frequently requires the browser to travel via a number of different redirect services. The fragility of these redirect threads is difficult to estimate naively, because of this high level of infrastructural interdependence.

This preliminary investigation has left us with many open questions, amongst them the level to which URL redirection services are localized in use. Would a similar survey conducted in Japanese produce a similar result? We note that the visual evidence here suggests a strong relationship between the resolvers used and the type of information referred to – for example, news sites appear to be more likely to use custom redirect services. Finally, we also note that there is some evidence in our data that redirect behavior is a useful metric for identifying spam sites which might allow the reduction of spam with Twitter retreets.

ACKNOWLEDGMENTS

Thanks go to James Currington for his insight into Twitter analysis, to Greg Tourte for his technical expertise, and to Adam Chen for reviewing.

REFERENCES

Carmody, T. (2011), A Tangled Web of Shortened Links. A study of link shortening reveals hidden strands of the Web. *Technology Review*. Retrieved May 16, 2011 from http://www.technologyreview.com/news/423170/a-tangled-web-of-shortened-links/

Clemens, E. K., Barnett, S. and Appadurai, A. (2007). *The future of advertising and the value of social network websites: some preliminary examinations.* ICEC '07: Proceedings of the ninth international conference on Electronic commerce.

Day, M. (2003). *Preserving the Fabric of Our Lives: A Survey of Web Preservation Initiatives*. T. Koch and I.T. Sølvberg (Eds.): ECDL 2003, LNCS 2769, pp. 461–472, 2003.

dev.twitter.com (2012). *The T.co URL wrapper*. Retrieved May 15, 2012 from https://dev.twitter.com/docs/tco-url-wrapper

Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K.. and Almeida, V. (2011). On word-of-mouth-based discovery of the web. *Proceedings of IMC 11*.

Thomas, K., Grier, C., Paxson, V. and Song, D. (2011). Suspended Accounts in Retrospect: An Analysis of Twitter Spam, *Proceedings of IMC 11.*

Yi.tl (2012). *URL shorteners*. Retrieved May 15, 2012 from http://yi.tl/pages/urlshorteners.php.