2014 IEEE Joint Intelligence and Security Informatics Conference

A Service-independent Model for Linking Online User Profile Information

Matthew Edwards Awais Rashid Paul Rayson
Security Lancaster, School of Computing and Communications
Lancaster University
Lancaster, UK
Email: m.edwards7@lancaster.ac.uk

***Abstract*— Public user profile information is a common feature of modern websites. These profiles can provide a valuable resource for investigators tracing digital artefacts of crime, but current approaches are limited in their ability to link identities across different platforms. We address this through a service-independent model of user profile information, grounded in the details visible on a number of the most-frequented sites on the web. Building on this, we report the details most widespread across platforms and the number of features visible on each site, thus highlighting details of use to both privacy researchers and investigators attempting to cross-link profiles.**

I. Introduction

Users of the modern web are no longer visible only to the owners of the websites with which they interact. Web services do not just collect information about users, they publicly reveal some portion of this information on a user's profile page within their website. Identifying details which could once be found on a web user's personal homepage are now replicated across a number of websites, and visibly linked to public logs of their activities on that site.

For practitioners of open-source intelligence (OSINT), this presents an unprecedented data-gathering opportunity. But while a growing number of technologies exist for deep exploration of the information made available by particular services, current methods are still not capitalising on the potential value of profiles constructed from identities correlated across numerous services. A major hindrance to such linkage is the lack of a common, service-independent model for understanding the information exposed in user profiles.

In constructing such a model, it is not sufficient to merely examine the few most-dominant social networking sites. Critical information for investigations can be buried in many rarely-considered websites which carry surprising amounts of user information. In this paper, we present a model of publicly-visible user profile information as it is presented in the 100 most-frequented sites on the web, thus drawing upon the common elements of user profile pages from sources as diverse as Wikipedia, Imgur, Reddit and even popular pornographic sites like XHamster. Our intent is that this model can be used to identify common ground for collating, linking and comparing the user profiles visible from nearly any web service which reveals them.

While record linkage as a field predates the advent of the Internet [1], there is a sparsity of reapplication in online contexts. Chandler [2] provides a detailed discussion of identity construction on personal web pages, along with a structured list of the key features of such pages. While instructive, their 1998 feature set may not fit modern, service-defined user profile pages. In a different vein, a Semantic Web effort at self-published friend-of-a-friend (FOAF) user profiles [3] has already allowed some success in linking user identities across social networks [4]. While FOAF covers a range of user identity items, it does not capture types of information which social networking sites may reveal, such as markers of fame and different forms of user relationship.

As well as our model of user profile information, we present a comparison which assesses how much of a user's potential online profile can be directly extracted from each of the included services, surveying the privacy risks posed by maintaining a profile on each site as well as the potential value for legal and ethical investigations involving OSINT.

This paper proceeds as follows: Section II outlines the method by which websites were chosen; Section III details the model constructed from the contents of these websites; Section IV summarises the proportion of possible information which is directly obtainable from each service.

II. Data Sources

To build an accurate model of the information contained in user profiles, we make use of a grounded theory approach, using examples visible on the web. The category of website we are interested in to inform our model is simply restricted to those which contain publicly-accessible user profiles, but a manual review of the structure of all such sites would be impractical. An alternative approach would be to focus on the websites which contain the most user profiles, but information on the size of online communities is difficult to obtain, perhaps because such information is increasingly considered commercially sensitive.

Instead, we draw on rankings of a website's overall traffic, as provided by Alexa (http://alexa.com). In the case of sites like Twitter, Facebook or Reddit, traffic volume will usually correlate with user population, but this does not necessarily hold for sites such as Wikipedia or the BBC, where typical traffic is consumption-oriented and few visitors create public

978-1-4799-6364-5/14 $31.00 © 2014 IEEE
DOI 10.1109/JISIC.2014.55

IEEE computer society

profiles. See also sites such as Adobe, where user profiles are only for a small number of users in a discussion forum located on a site visited for other reasons.

Working from the Alexa ranking of worldwide websites, we examined the top 100 sites for evidence of user profile pages. We found that 65 of these sites held some form of publicly-accessible user profile page for a member willing to fill out details and/or publicise activity information. The Alexa rankings included many highly-ranked sites which were effective duplicates of each other: for example, google.com and google.co.uk both appear on the list, but refer to essentially the same service and user profile content[1]. Discounting such duplicates, there are 39 services containing individual profiles. Of these, four proved difficult to translate and were discounted, leaving a final selection of 35 services, which we examined to form our model of available user information.

III. Service-independent Model

For each component identified, a description is accompanied by the number of services which reveal this information. We list only information which is structurally supported by a service's platform, not incidental user adaption of fields, nor information which may be inferred from profile content.

A. Contact

Types of contact information are noted in Table I. We did not include means of contacting the user such as platform-specific messaging options, as these details do not aid in identification from outside the platform. Profile links were surprisingly common, and could be considered highly useful for re-identification purposes. Web links were the most broadly-available contact identifier.

Table I Contact information

Name	Description	#Services
Web links	URL designated for the user's web site or page.	11
Profile links	Direct links to this person's profile on other services, explicitly particular to those services.	9
Email address	An email address for the user. Potentially service-provided or partially anonymised.	5
Phone number	A visible phone number for the user.	3

B. Biographical

Various fields from services provide biographical information about users. This includes the only feature common to all services — the username. As Table II shows, usernames were rarely verified as the user's actual name, but in practice it may well be a useful indicator. The most common field after the `Username` was `Self-description` areas, which contain free text about the user.

Table II Biographical attributes

Name	Description	#Services
Username	A human-readable identifier for the user.	35
Self-description	A free text field for user self-identification.	21
Age	The age or birthdate of the user.	10
Tags	Short textual labels self-applied by the user.	10
Education	The user's educational history, marked by school names. (Sometimes dated).	8
Occupation	The user's current employment.	8
Gender	An explicit marker for the user's gender.	6
Relationship status	A field denoting the user's relationship status.	6
Sexual orientation	A field denoting the user's sexual orientation.	4
Verified	Whether or not the platform indicates if the username given matches the user's real name.	3
Religion	A field for the user's faith.	3
Physical description	An physical self-description of the user.	2
Habits	Fields denoting other personal habits.	2

C. Visual

Visual identification information is defined here as images useful to visually identify the user, outlined in Table III. The use of avatars or profile images is widespread, but they are not always photographic representations of the user. Images uploaded by the user (Section III-I) may contain pictures of them, but this not assured.

Table III Visual identifiers

Name	Description	#Services
Profile image	A user's chosen photo or avatar representation.	23
Banner image	A background or banner image, usually chosen by the user to complement their profile image.	13
Tagged photos	Photographs labelled as containing the user, usually as identified by the service or other users, but accessible via the user's profile.	3

D. Opinion

Opinion markers (e.g. favourites, star-ratings). Textually expressed opinion is not included in this set, as it requires additional inference. As Table IV shows, opinion markers are most common for in-platform content.

Table IV Opinion markers

Name	Description	#Services
Content	This user's opinion of on-platform user content.	12
Brand	This user's opinion of a product, company or item on the platform.	7
Other	The user's opinion of off-platform items.	2

E. Temporal

As Table V shows, activity timestamps are common, with only two services not revealing these. The two other information items can be partially inferred from these logs. Not noted is the specificity of time recording – some sites present detailed timestamps, while others provide vaguer information as items age.

[1] Though the user populations and demographics will differ

Table V Temporal information

Name	Description	#Services
Activity timestamps	Date and time stamps for user posts and activity, accessible via the profile page.	33
Membership date	When a user created their account.	17
Last seen	The date and time at which the user was last recognised by the service.	9

F. Geographical

Geographical information seems more closely guarded by services than temporal information. As Table VI shows, the most common form is an indicator of the user's location in a city or country, with varying specificity. In interpreting this table, note that location history simply extends a location set, so both are counted where the first is available.

Table VI Geographical Information

Name	Description	#Services
Current Location	The user's address or advertised location.	12
Location Set	A set of locations associated with the user.	6
Location History	A timestamped set of user locations.	3

G. Degree

Degree information is information about the user's popularity, dedication or influence. Most commonly this is linked to social connectedness within a network's population, particularly the number of other users following[2] a user.

Table VII Degree metrics

Name	Description	#Services
Subscribers	The number of other users following this user.	21
Subscribed	The number of other users that this user follows	16
Contributions	The number of contributions that this user has made to the network (links, posts, videos, etc.)	16
Visibility	The number of views of this user's profile.	9
Reputation	This user's reputation, as rated by other users or other performance metrics.	9
Trophies	Markers for a user's special achievements.	9
Rank	The user's rank within the community, as one of a number of possible tiered levels.	5

H. Relationships

The relationship information presented in Table VIII is broken down into two components: relationships the user has with other people in a platform's network and relationships that the user has with brands within the network, where brands are either companies, products, or media items which are not uploaded by other users. Where relationships in a platform's network are bilateral friendships, both `Follower` and `Following` are considered fulfilled.

Table VIII Relationships

Name	Description	#Services
	People	
Interacted	A user this user has interacted with.	26
Follower	A user following this user.	20
Followed by	A user this user follows.	14
Grouped	A user this user shares a group with.	10
	Brands	
Follower	A product, company or item this user follows.	13
Contributor	A product, company or item this user has interacted with or is marked as affiliated with.	6

I. Content

As Table IX shows, a text corpus of the user's contributions is available for most platforms, with user-submitted images also common. Table X shows the attributes which each item of content may have. Only content directly reachable from profile pages was considered.

Table IX Content types

Name	Description	#Services
Text	Typed text content, not necessarily exclusive of images or other media	29
Image	Still images.	17
Video	Video content.	11
Links	Links to content from outside the platform	5

Table X Content attributes

Name	Description	#Services
Temporal	Timestamp for a post.	31
Impact	The content item's popularity, in terms of any of various vote or viewing measures.	31
Category	A categorisation or tag for the content item.	12
Spatial	Geographical locator for the content	3

IV. Vulnerability

In constructing this model, we have begun evaluation of a means of assessing the vulnerability – or value, from an investigator's perspective – of certain categories of information revealed in user profiles. The value of personal information varies according to the context of an investigation, but the level of support for individual information items suggests that certain pieces of information are more pervasive.

Table XI lists those profile content items which were common to over 60% of the sites examined. An investigator aiming to match users across platforms – perhaps searching for extra information in order to facilitate an arrest for online harassment – may find these attributes of greater value due to their ubiquity; the combination of usernames and linguistic

[2] Following being where a follower is publicly listed as receiving updates.

fingerprints with temporal data and network graphs can be considered highly identifiable on a broad range of services.

Table XI The most widely represented information items

Item	#Services
Biographical→Username,	35
Biographical→Self-description	21
Visual→Profile Image	23
Temporal→Activity Timestamps	33
Content→Text	29
Content Attributes→Time	31
Content Attributes→Impact	29
Relationships→People→Interacted	26
Degree→Subscribers	21

It is also true that some sites have more capacity to reveal information about their users. For privacy-concerned users, these sites might be avoided or approached with caution. For the OSINT investigator, these sites could be targeted for monitoring. Figure 1 shows the number of items from our model present in each site identified. The sites are ranked by Alexa's web traffic ranking, and coloured according to rough approximations of their function.

Clear leaders in information content are Google and Facebook, as would be expected. Pornographic sites also rank high, with detailed biographical pages for members, who seem to use these pages in a similar manner to a dating site. Question-and-answer forums such as StackOverflow seem slightly above-average in their information content. Knowledge-building sites like IMDB and Wikipedia ranked low, as did news sites, blogs and video-sharing sites. Blogs and Wikipedia user pages are notable in that both may carry more structured information about the user in optional widgets, which have uneven adoption.

Investigators intent on gathering more information about a suspect can make use of this analysis by prioritising sites with greater potential levels of information when trying to link a target's profiles, or when prioritising new information-gathering technology development.

V. Conclusion and Further Work

We have presented a model of the information publicly available on a wide variety of user profiles, taken from some of the most trafficked sites on the web. This model can be used to represent the information available on these profiles in a cross-platform manner, aiding efforts at search, comparison and identity resolution.

The model presented could be extended to include information obtainable through deeper examination of a user profile. Such information may include links and email addresses mined from text fields, linguistic fingerprints trained on user comments [5] and even facial recognition models trained on videos and photos of the user [6].

We have also presented a shortlist of the information items most common across sites, which could be considered some of the most valuable items for investigators attempting to link profiles from one platform to the other. This evaluation is incomplete, however: additional factors to be considered include the size of user communities on each platform, the reliability of information provided in online profiles (and whether this varies) and the completeness of user profiles on various sites. In future work we hope to measure the influence of these factors and extend this analysis.

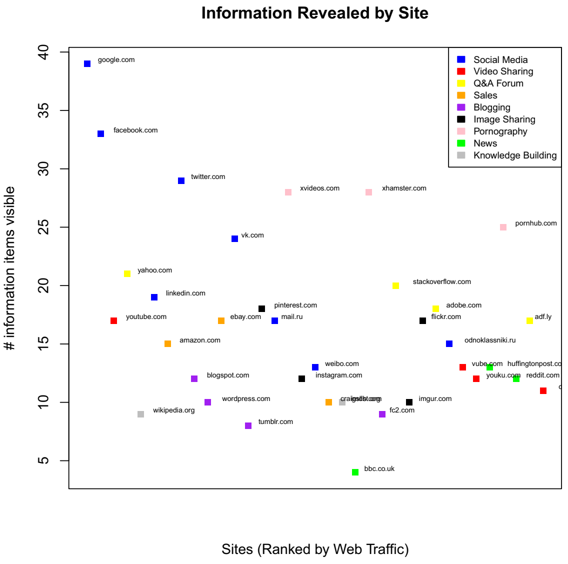

Figure 1 Sites plotted against number of public information items

References

[1] H. L. Dunn, "Record linkage," *American Journal of Public Health*, vol. 36, no. 12, pp. 1412–1416, 1946.

[2] D. Chandler, "Personal home pages and the construction of identities on the web," *URL (consulted June 2014): http://www.aber.ac.uk/˜dgc/Webident.html*, 1998.

[3] L. Ding, L. Zhou, T. Finin, and A. Joshi, "How the semantic web is being used: An analysis of foaf documents," in *System Sciences, 2005. 38th Annual Hawaii International Conference on*. IEEE, 2005, pp. 113c–113c.

[4] J. Golbeck and M. Rothstein, "Linking social networks on the web with foaf: A semantic web case study." in *AAAI*, vol. 8, 2008, pp. 1138–1143.

[5] A. Rashid, A. Baron, P. Rayson, C. May-Chahal, P. Greenwood, and J. Walkerdine, "Who am i? analysing digital personas in cybercrime investigations," 2013.

[6] M. Zhao, J. Yagnik, H. Adam, and D. Bau, "Large scale learning and recognition of faces in web videos," in *Automatic Face & Gesture Recognition, 2008. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–7.