A Cross-Platform Collection of Social Network Profiles

Maria Han Veiga
Inst. of Computational Science
University of Zurich, Switzerland
hmaria@physik.uzh.ch

Carsten Eickhoff
Dept. of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

ABSTRACT

The proliferation of Internet-enabled devices and services has led to a shifting balance between digital and analogue aspects of our everyday lives. In the face of this development there is a growing demand for the study of privacy hazards, the potential for unique user de-anonymization and information leakage between the various social media profiles many of us maintain. To enable the structured study of such adversarial effects, this paper presents a dedicated dataset of cross-platform social network personas (*i.e.*, the same person has accounts on multiple platforms). The corpus comprises 850 users who generate predominantly English content. Each user object contains the online footprint of the same person in three distinct social networks: Twitter, Instagram and Foursquare. In total, it encompasses over 2.5M tweets, 340k check-ins and 42k Instagram posts. We describe the collection methodology, characteristics of the dataset, and how to obtain it. Finally, we discuss a common use case, cross-platform user identification.

CCS Concepts

•**Information systems** → *Test collections;*

Keywords

Collection; Data set; Online Social Networks

1. INTRODUCTION

The field of computational social science and data-driven research is growing in importance [9], and with this trend, there is a need for common academic benchmarking collections to facilitate a robust and reproducible research environment. In practice, however, datasets are often obtained via ad-hoc collection, or on the basis of proprietary data.

While originally Online Social Networks (OSNs) focused on allowing users to communicate, connect with others and share content, nowadays the term includes platforms which are primarily user-centric, allowing members to broadcast

personal thoughts and content. [8] finds that OSNs are among the most frequently visited Web sites for a large population of users. In consequence, they can be used to study human behaviour at a large scale. Furthermore, because many OSNs are public by default and provide APIs to access their content, they have become good candidates for data collection to be used to study problems such as user/ topic modelling, user identification or information leakage.

In this paper, we introduce a collection of 850 users and their online footprint (part of their generated content and user profiles) spread across three social networks: Twitter, Instagram and Foursquare. It is our objective to provide a dataset on which privacy-sensitive methods and the defense against them can be tested. The construction of this dataset was initially carried out during a research project studying cross platform privacy loss arising from public information sharing on social networks [4]. In Table 1, a comparison of our dataset with other existing corpora is given.

Our data collection method relies on linking users across three popular OSN platforms. Twitter is a microblogging platform whose main content comes in *tweets*, posts limited to 140 characters which can contain text, video or images, links to external Web sites, references to other users and *hashtags* (terms starting with the # symbol used to mark keywords or topics in a tweet). Instagram is a photo sharing platform. Its main content are photos or videos along with optional textual descriptors. Foursquare is a location service platform concentrating on the notion of *check-ins*.

The paper makes the following novel contributions (1) We describe the methodology and release the code to construct a user-centric cross-OSN dataset. (2) We release a dataset of 850 profile triples across the aforementioned OSNs.

The remainder of this paper is structured as follows: in Section 2, we present the methodology used to create the collection. Section 3 presents key statistics and qualitative aspects of the dataset. Section 4 discusses the task of cross-platform user identification as an example use case before sketching a range of further conceivable use cases and tasks.

2. METHOD

We use Twitter, Instagram and Foursquare mainly due to the ease of crawling and publicly accessible APIs, but also because the content generated by the users in these distinct social networks is diverse and representative of a comprehensive range of OSN use cases. In [11], the authors find that on Twitter, the top types of content users share are: personal information, random thoughts, opinions/complaints and facts (*e.g.* news). While on Instagram, the ma-

Table 1: An overview of existing datasets containing cross-OSN user profiles.

Features	MNA [7]	MAH [12]	About.me [10]	Cross OSN
OSNs	Twitter, Foursquare	Twitter, BlogCatalog	Flickr, Google+, Instagram, Tumblr, Twitter, Youtube	Instagram, Twitter Foursquare
# Users	500	2710	15,595	850
Content type	User IDs, posts, friends graph	User IDs, friends graph	User IDs, post IDs	User IDs, post IDs
Availability	Unknown	Unknown	Available	Available

jority of the posted pictures can be put in 1 of 8 categories, with the leading categories being *selfies* and *friends* [5]. On Foursquare, users share their location in terms of venues, which carries not only information in the form of raw geographic coordinates, but often also the venue's name and function.

In order to obtain profiles from different OSNs belonging to the same person, we first use the Twitter Search API to search for specific post patterns (*e.g.*, https://instagr.am/p/*), in order to identify users who cross-post content from other platforms on Twitter. A graphical depiction of the collection process is shown in Figure 1.

The data was collected in January and February 2016. While we were able to find over 5000 profile triples, only approximately 20% of these triples fulfill our criteria of actively using the three selected OSNs, posting predominantly in English and sharing content publicly. After enforcing these requirements, our dataset contains a total of 850 distinct user profiles.

2.1 English predominance

We focused on profiles with predominantly English content as the initial study for which the data was collected was a natural language processing task that would have suffered from excessive amounts of cross-language content.

In order to guarantee that the majority of the content is posted in English, we use the method described in Algorithm 1. We set the ratio to be 0.1 and $K = 100$ for our collection. The purpose is not to exclude users that occasionally post in a non-English language, but to make sure the data set does not contain too many strictly non English-speaking users.

> **Data:** twitter timeline
> **Result:** true or false
> initialization;
> **for** K *posts* **do**
> | count(english posts);
> **end**
> **if** *count.total*/K < *ratio* **then**
> | disregard user;
> **else**
> | crawl timeline;
> **end**

Algorithm 1: English check

2.2 Spam detection

To increase the confidence that users are authentic personal accounts instead of spammers that merely redistribute content from other users, we have a simple heuristic described in Algorithm 2. We take the timeline of the user from the respective OSN and we check whether the majority of the posts come from the same user name. In our collection, the threshold is set to 30%. Suspected spammer triples are flagged but, for completeness, remain in the dataset.

> **Data:** Instagram or Foursquare timeline
> **Result:** spammerFlag
> initialization;
> **for** *posts* **do**
> | count(author of post);
> **end**
> **if** *counter.max/total posts* < *threshold* **then**
> | spammerFlag=1;
> **else**
> | spammerFlag=0;
> **end**

Algorithm 2: Identity check

3. DATASET

Our 850 authentic English-speaking users produced approximately 2.5M tweets, 340k check-ins and 42k Instagram posts[1]. Four users were flagged as spammers under the rule described in Section 2.2.

Figure 2 shows the distribution of tweets per user. Figures 3 and 4 show the distribution of number of check-ins and Instagram posts per user, respectively.

The Twitter API restricts access to at most 3200 tweets per profile (including re-tweets) [13]. Because we exclude direct re-tweets from our data set, the majority of the Twitter profiles we collect contain between 3000 and 3200 tweets. For each Instagram profile, we recover the most recent posts through the Instagram API and complement them with existing content which has been posted on Twitter. For each Foursquare profile, we recover those check-ins that were cross-posted on the retrieved Twitter timeline.

In this dataset, we observe check-ins from 111 countries, spawning 669 venue types. The most visited venue types are shown in Figure 5.

3.1 Properties

For each user we store Twitter, Instagram and Foursquare IDs along with the tweet and post IDs of all shared content. The Foursquare data is represented as a list of tuples $(venue\_id, tweet\_id)$, where $tweet\_id$ is the tweet announcing the check-in. An example of the collected user data is shown in Table 2. For copyright reasons, we do not distribute the content but instead provide the scripts to crawl the data [1]. Using these scripts, the following can be retrieved:
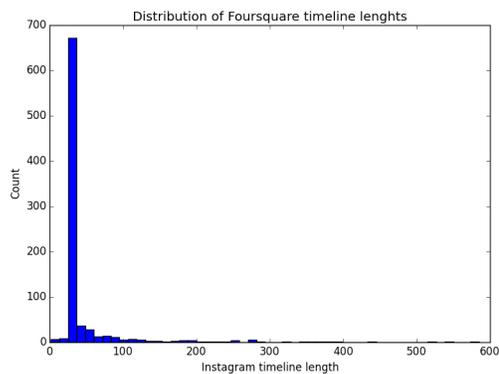

[1]http://cake.da.inf.ethz.ch/OSN-sigir2016/
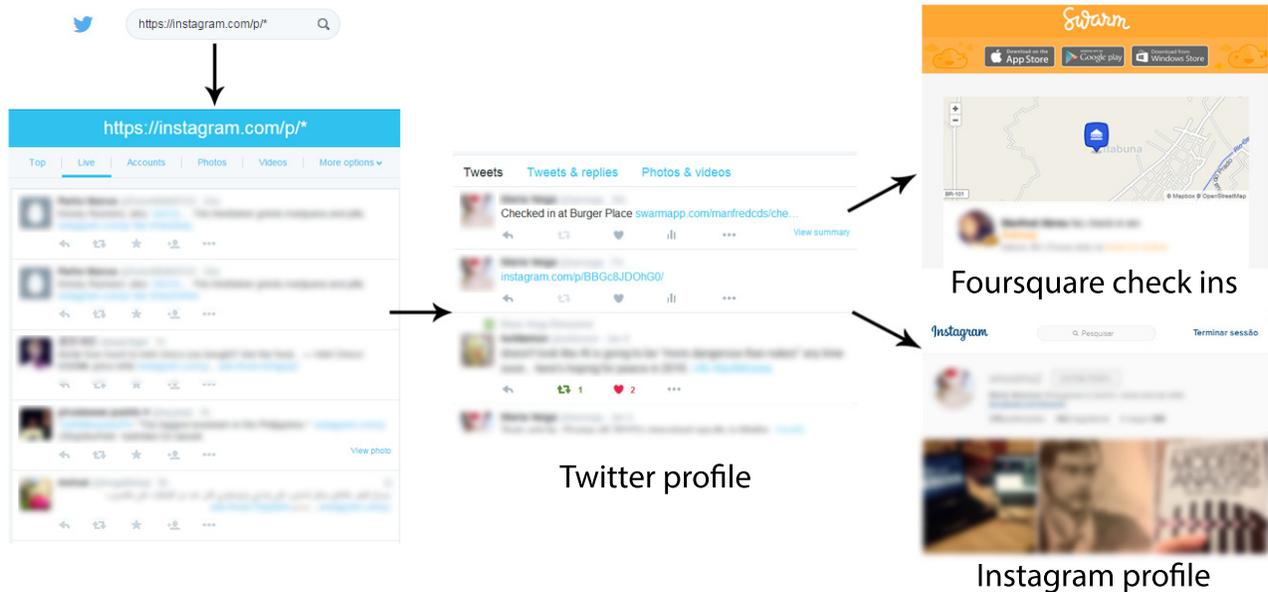
Figure 1: Dataset collection. We use Twitter's search API to find posts with the pattern of a Foursquare check-in or an Instagram post. If a profile contains both, we crawl the Foursquare check-ins and Instagram profile through their respective APIs.
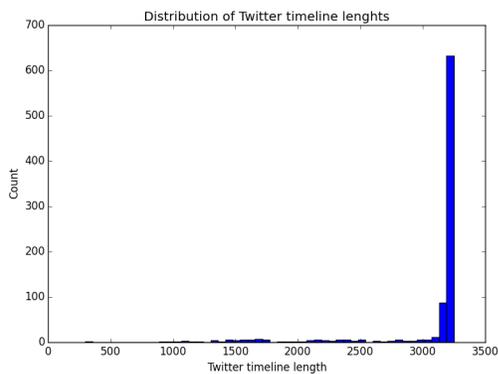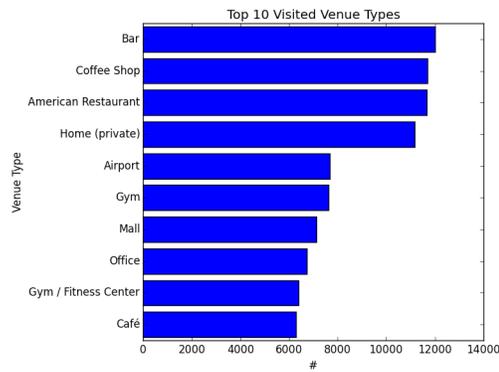

Figure 2: Histogram of tweets per user

Figure 4: Histogram of Instagram posts per user
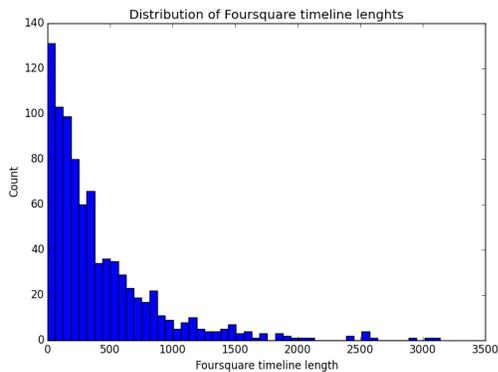
Figure 3: Histogram of Foursquare posts per user

Figure 5: Top 10 visited venue types

Table 2: Example of provided and retrievable data.

Feature	Description	Retrieves
TWITTER		
ID	Integer	Twitter profile
Timeline	List of tweet IDs	Tweet object
INSTAGRAM		
ID	Integer	Instagram profile
Timeline	List of post shortcodes	Instagram object
FOURSQUARE		
ID	User unique identifier	–
Timeline	List of pairs (tweet_id,venue_id)	Venue object

- From Twitter: using the user ID, a profile object that contains information such as the name, location, date of creation of account, and with the post ID, a the tweet object, containing the tweet and metadata [13].

- From Instagram: using the user ID, a profile object containing information such as the name, location, date of creation of account, and with the post shortcode from Instagram, an Instagram object containing the link with the image and related metadata [6].

- From Foursquare: using the venue ID, the venue object can be retrieved, containing information such as venue type, venue name and location [3].

4. DISCUSSION

In this section we present a use case for this dataset in a cross-OSN user identification task. Then, we elaborate on other tasks that can find this collection useful.

4.1 User Identification

The task of user identification is concerned with matching profiles from different domains belonging to the same natural person [14]. In this scenario, we use Instagram and Twitter as our data sources. A simple, yet powerful baseline algorithm is used, which compares the similarity of the nicknames on the respective platforms.

By assigning minimal Levenshtein distance between user names and choosing the one with the lowest edit distance, we attain a matching accuracy of 70.1%. If the user names are preprocessed by converting them to lowercase (as in both these social networks the letter case does not make a difference), we attain an accuracy of 72.8%. One can think of many more advanced schemes tracking common topics or writing styles across social networks.

4.2 Content generation and spreading

In [2], the authors study how users behave across Pinterest and Twitter. Other papers study Twitter or Instagram alone [11, 5]. It would be interesting to study the behaviour of users in these OSNs while having access to their multiple profiles. This could help answer whether some topics are OSN-specific or whether the activity in one profile can indicate future activity in another profile.

4.3 Cross-OSN inference

Following the idea that an activity on one platform can indicate future activity on another one, an interesting study would be to see whether it is possible to infer information from one OSN regarding another one. For example, whether it is possible to predict topics, interests or intentions. An example of this type of work can be found in [4], where the authors use Twitter timelines to infer venue type visits.

4.4 Aggregated OSN topic modelling

Because OSNs can carry different information depending on their functionality, the access to several profiles from the same user across different platforms might give an advantage when modelling the user. These user models can be used for targeted advertisement or recommender systems. An interesting task would be to compare whether the inclusion of different profiles can benefit the models or not.

5. CONCLUSION

In this paper, we described the collection, structure and properties of a benchmarking corpus of cross-platform OSN user profiles useful for a wide range of privacy-related research questions. It encompasses hundreds of user triples, millions of tweets and thousands of Instagram and Foursquare posts. In order to reduce the inherent amount of noise in the dataset, we ensure a focus on English-speaking users and perform spammer identification. To the best of our knowledge, this dataset is the first of its kind both in nature as well as scale.

6. REFERENCES

[1] *CrossOSN-crawler code repository*. https://github.com/hanveiga/CrossOSN-crawler.

[2] R. O. et al. Of Pins and Tweets: Investigating how users behave across image- and text-based social networks. In *ICWSM 2014*.

[3] Foursquare. *Foursquare API overview*, 2015 (accessed May 2, 2015). https://developer.foursquare.com/overview/.

[4] M. I. Han Veiga. *Task Driven Information Valuation*. Zürich, 2015.

[5] Y. Hu, L. Manikonda, and S. Kambhampati. What we instagram: A first analysis of instagram photo content and user types.

[6] Instagram. *Instagram API overview*, 2015 (accessed May 2, 2015). https://www.instagram.com/developer.

[7] X. Kong, J. Zhang, and P. S. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM 2013*.

[8] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *WWW 2010*.

[9] D. Lazer, A. Pentland, and et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[10] B. H. Lim, D. Lu, T. Chen, and M.-Y. Kan. #mytweet via instagram: Exploring user behaviour across multiple social networks. In *ASONAM 2015*.

[11] M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: Message content in social awareness streams. In *CSCW 2010*.

[12] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen. Mapping users across networks by manifold alignment on hypergraph.

[13] Twitter. *Twitter Rest API overview*, 2015 (accessed May 2, 2015). https://dev.twitter.com/rest/public.

[14] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In *KDD 2013*.