WhACKY![*] - What Anyone Could Know About You from Twitter

Denzil Correa
IIIT-Delhi
denzilc@iiitd.ac.in

Ashish Sureka
IIIT-Delhi
ashish@iiitd.ac.in

ABSTRACT

Twitter is a popular micro-blogging website which allows users to post 140 character limit messages called tweets. We demonstrate a cheap and elegant solution – *WhACKY!* – to harness the multi-source information from tweets to link Twitter profiles across other external services. In particular, we exploit *activity feed* sharing patterns to map Twitter profiles to their corresponding external service accounts using publicly available APIs. We illustrate a proof-of-concept by mapping 69,496 Twitter profiles to at least one of the five popular external services : Flickr (photo-sharing service), Foursquare (location-based service), YouTube (video-sharing service), Facebook (a popular social network) and LastFM (music-sharing service). *WhACKY!* guarantees that the mapped profiles are 100% true-positive and helps quantify the unintended leakage of Personally Identifiable Information (PII) attributes.

1. INTRODUCTION

Due to the advent of Web 2.0 technologies, there has been a swift rise in the number of social networking services. Internet users utilize these social networks to connect and share information and diverse kinds of media with each other. Twitter is one such immensely popular micro-blogging website which allows users to share short 140-character messages with each other. *Twitterers* connect with other users via a subscription feature called *Follow*. Twitter provides its registered users with other features to interact with each other, such as: reply or mention (@-message), repost (Retweet or RT), private messages (direct messages or DM), favorites and lists (categorization of users). Twitter has recently added capabilities to natively post images within the Twitter web interface. However, Twitter does not provide users with built-in options to share diverse kinds of media, such as video and music. Nonetheless, these features are a *Unique Selling Point* of akin niche social networks like YouTube, LastFM and Foursquare. Several popular web services such as YouTube, Foursquare and LastFM are designed to allow users to share different kinds of information and media. Studies show that social networks that combine information from multiple sources enhance user social experience. *Twitterers* frequently share content hosted on external services like LastFM, YouTube and Foursquare.

2. AIM AND RESEARCH CONTRIBUTION

The research aim of this paper is to exploit *activity feed* sharing patterns on Twitter to infer a user's social identity mapping across multiple social media services like Flickr, YouTube, LastFm and Foursquare in order to assist real-world applications like user data privacy awareness and digital marketing. The main contributions of our work are –

1. *Investigation of activity feed sharing patterns for social profile identity mapping* – The investigation of mining *activity feed* sharing patterns for social profile identity mapping is a unique contribution in context to previous approaches. *Activity Feed* sharing is a popular feature utilized by users on various social media. YouTube reports that nearly 17 million people connect their YouTube accounts to another social network and over 12 million people share their YouTube activity on at least one social network. We mine this information flow to demonstrate an extremely low-cost, elegant and efficient technique to map social profiles across different networks.

2. *First focussed study on social profile identity mapping on Twitter* – To the best of our knowledge, this is the first empirical study to focus on mapping Twitter profiles to other networks. We acknowledge that there are generic solutions which are applicable to social networks like Twitter. But, these solutions use Twitter as a test-bed for experiments and do not consider specific properties of Twitter as a whole. In contrast, we focus on the *activity feed* sharing patterns in tweets which are generated due to profile connections. Mining tweets to identity Twitter profiles on other networks is a novel contribution in context to previous work.

3. METHODOLOGY

Our solution consists of a three-step framework – *Filter*, *Extract* and *Connect*. Figure 1 illustrates the framework used in our proposed approach. We now discuss this three step framework.

1. *Filter* – Due to sharing of *activity feeds* from other social networks like Flickr and YouTube, auto-generated tweets contain common patterns. The first step of our framework requires identification of tweets with common text patterns for the respective service. The *Filter* block in Figure 1 shows the common text patterns occurring in tweets for external services like Flickr, Foursquare, LastFM and YouTube. We filter tweets

[*] Published in Proceedings of the 10th Annual International Conference on Privacy, Security and Trust (PST), 2012. `http://bit.ly/1ezlcWu`

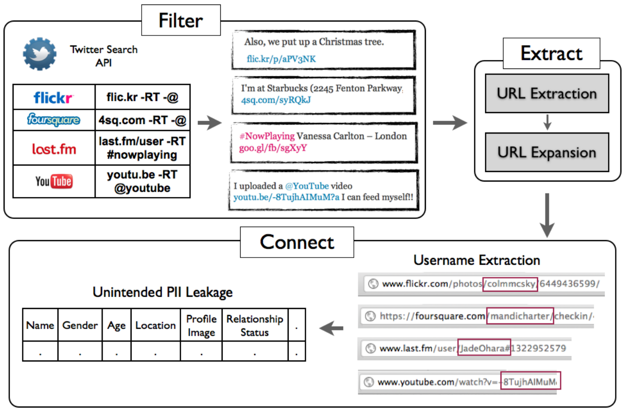
Filter
Twitter Search API
flickr
flic.kr -RT -@
foursquare
4sq.com -RT -@
last.fm
last.fm/user -RT #nowplaying
YouTube
youtu.be -RT @youtube
Also, we put up a Christmas tree. flic.kr/p/aPV3NK
I'm at Starbucks (2245 Fenton Parkway. 4sq.com/syRQkJ
#NowPlaying Vanessa Carlton – London goo.gl/fb/sgXyY
I uploaded a @YouTube video youtu.be/-8TujhAIMuM?a I can feed myself!!
Extract
URL Extraction
URL Expansion
Connect
Username Extraction
www.flickr.com/photos/colmmcsky/6449436599/
https://foursquare.com/mandicharter/checkin/
www.last.fm/user/JadeOhara#1322952579
www.youtube.com/watch?v=-8TujhAIMuM
Unintended PII Leakage
Name
Gender
Age
Location
Profile Image
Relationship Status

Figure 1: Three step framework – *Filter*, *Extract* and *Connect* – for our proposed solution approach.

according to these text patterns and pass them to the next block.

2. *Extract* – The auto-generated tweets obtained from the previous step contain explicit short URLs to the content hosted by the same user on another social network. We extract such short URLs from the tweets obtained in the previous step and expand them. The *Extract* block in Figure 1 represents this step of our framework. These expanded URLs are passed to the next block.

3. *Connect* – In the final step, we obtain the URLs obtained from the previous step and extract uniquely identifiable profile information on the external service like username or user id. We link the user's Twitter profile to these external services. We now extract PII from the external social network and gain access to more information about the user. The *Connect* block in Figure 1 shows how profile information embedded in URLs can be used to link Twitter profiles to external services.

4. RESULTS AND EVALUATION

4.1 Social Profile Identity Mapping

Table 1 shows the number of unique Twitter profiles mapped to other social networks. Foursquare mappings contained the highest number of users while LastFM contained the least, indicating that there exists a small subset of users who generate many auto-generated tweets. Note that the Twitter Search API returns only the 1,500 most relevant results to the input query per day. Hence, these numbers only place a lower-bound on the number of users who connect their Twitter profiles to other social networks.

4.2 Unintended Personal Information Leakage

The mapping of social profiles across multiple networks leads to increase in access of PII of the user and various approaches have been proposed in literature to collect this PII. Table 2 shows the percentage of publicly available PII attributes observed in each social network of our dataset. These percentages reflect a conservative estimate of the attributes in each social network. For example, if the contacts of a user were available but 0 in number, we do not count that attribute towards the final percentage.

Social Network	Number of unique users mapped
Flickr	14102
Foursquare	32646
YouTube	22672
LastFM	76
Facebook	16934
Twitter (total)	**69496**

Table 1: Uniquely identified Twitter profiles across external services like Flickr, Foursquare, YouTube, LastFM and Facebook.

4.3 Username Uniqueness & Profile Duplication

We notice that there is a significant amount of overlap in the usernames used by *Twitterers* on external services. We also observe a high overlap of usernames between Twitter profiles and their Foursquare profiles showing that one could predict a *Twitterer*'s Foursquare profile by a simple lookup.

We observe that a few *Twitterers* have multiple Twitter profiles but connect their Twitter profiles to the same external service. We see that **0.54%** of the Twitter profiles in our dataset are duplicate profiles.

	Flickr	Four Square	You Tube	Last FM	Face Book
Total Users	**14064**	**32646**	**22672**	**76**	**16934**
Name	64.5%	98%	66.63%	86.84%	100%
Profile Image	100%	97.91%	89.29%	98.68%	100%
Gender		100%	95.44%	98.68%	100%
Age			84.02%	98.68%	
Contacts	89.87%	99.37%	80.74%	89.47%	
Likes Favorites	75.17%	87.85%	25.22%	93.06%	

Table 2: Percentage of publicly available PII attributes present across each service in our dataset. Blank cells indicate that the PII attributes were not publicly available.

5. DISCUSSION AND CONCLUSION

Our algorithm is *elegant* and requires *no manual evaluation* as the mapped social profiles are 100% accurate. In order to achieve this accuracy, we adopt a conservative approach and discard tweets which do not clearly fit the pattern identified. A major limitation of our solution approach is that it is restricted to social networks like *Twitter*. However, we argue that similar *activity feed* sharing patterns are observed on other social networks like *Facebook* albeit to a lesser degree. Our approach is applicable to all social networks which allow *activity feed* sharing. In a nutshell, our proposed solution approach is *Cheap*, *Elegant*, requires *No Evaluation* and guarantees *100% Accuracy*.

We exploit the text patterns in auto-generated tweets as a result of such connections called *activity feeds*. We also demonstrate a proof-of-concept of our solution approach by connecting Twitter profiles to the social networks Flickr, Foursquare, Facebook, LastFM and YouTube. Our solution is also able to detect duplicate Twitter profiles in the process, requires no manual evaluation and gives 100% accuracy. We also show that mapping of Twitter profiles to external services leads to an increase of unintended leakage of sensitive personally identifiable information.