WhACKY! – What Anyone Could Know About You from Twitter

Denzil Correa, Ashish Sureka, Raghav Sethi
Indraprastha Institute of Information Technology - Delhi (IIITD)
New Delhi, India.
Email : {denzilc, ashish, raghav09035}@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). We evaluate our solution against a commercial social identity mapping service – *FlipTop* – and demonstrate the efficiency of our approach. *WhACKY!* guarantees that the mapped profiles are 100% true-positive and helps quantify the unintended leakage of Personally Identifiable Information (PII) attributes. During the process, *WhACKY!* is also able to detect duplicate Twitter profiles connected to multiple external services.We also develop a web application based on *WhACKY!*[1] for perusal by *Twitterers* which can help them better understand unintended leakage of their PII.**

I. Research Motivation and Aim

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 [1]. 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 [2]. *Twitterers* frequently share content hosted on external services like LastFM, YouTube and Foursquare.

[1] http://whackyapp.appspot.com/

To demonstrate why a user would choose to exhibit the above-described behaviour, consider the following example. Alice creates an account on Twitter with the screen name *alice* to post her daily life activities and engage in conversations with fellow *Twitterers*. Alice feels the need to share interesting videos with her Twitter followers and discovers that Twitter does not allow users to directly share videos with each other. Alice is, however, aware that YouTube is a popular video sharing service. Therefore, Alice registers for an account on YouTube with the username *allie* after she discovers that the username *alice* is already in use by another user. Alice uploads her videos on YouTube and shares links to these videos with her followers on Twitter by manually copying links to her Twitter profile. However, Alice soon starts to find the task of manually updating her Twitter profile every time she uploads a YouTube video to be an arduous and mundane activity. Alice searches around the YouTube website for a solution and finds that YouTube provides a feature to connect her YouTube profile with her Twitter profile to automatically share her YouTube *activities* (uploads, favorites, likes) on Twitter. Figure 1 shows the snapshot of the *activity feed* sharing feature provided by YouTube. Alice eagerly utilizes this feature and connects her Twitter and YouTube accounts. This allows Alice to automatically share her YouTube activities like video uploads with her Twitter followers. Alice need not update her Twitter profile manually as an auto-generated tweet is posted automatically every time she uploads a YouTube video. Figure 2 shows a tweet automatically generated as a result of a video upload on YouTube via the *activity feed* sharing feature. Similarly, Alice uses other external services like LastFM, Foursquare and Flickr to share music, interesting locations and photographs. Therefore, her profile contains diverse information from multiple external services like YouTube, LastFM and Foursquare.

*Twitterer*s can explicitly share links to content on external services via Twitter and enhance their experience. They may also leverage features on external services to easily connect their profiles to allow frictionless cross-network sharing. Hence, there exists an eco-system of cross-syndication and data flow between multiple social network websites like YouTube, Foursquare, LastFM and Twitter [3]. However, the

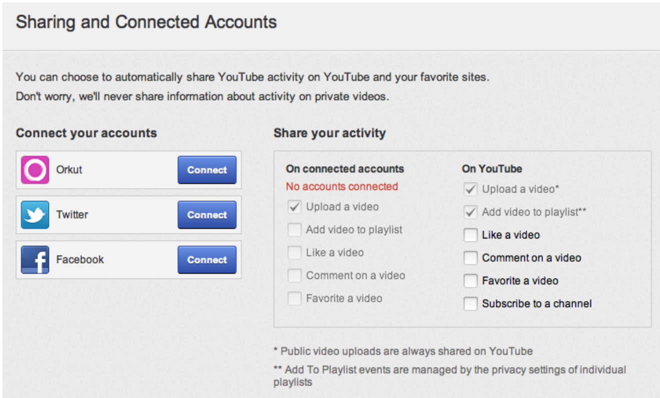

Fig. 1. **Screenshot illustrates the *activity feed* sharing feature provided by YouTube. A user can choose to connect his YouTube account to Orkut, Twitter or Facebook.**

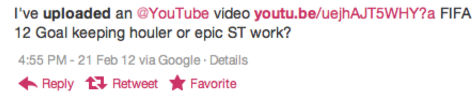

Fig. 2. **The screenshot illustrates an example of an auto-generated tweet posted on Twitter for a video upload on YouTube due to connection of a user's YouTube account to his Twitter profile.**

mapping of such social profile connections are not publicly available due to privacy issues and is a non-trivial problem as users could enter different information (both attributes and values) to different networks [4]. In the aforementioned example, Alice has different usernames on Twitter and YouTube. Nonetheless, it is clear that identifying such social network profile mappings could reveal the social footprint of a user [5], [6]. This information about the social footprint of a user can be of significant use to the concerned user as well as third party businesses. We present two real-world use cases to demonstrate our argument:

1) **User Data Privacy** – Internet users register on multiple social networks to avail unique features of each network. At the time of account creation, social networks generally ask users to provide certain personally identifiable information (PII) in order to cater to legal issues as well as to ensure *hygiene*. Users connect their social network profiles to facilitate ease of sharing and enhance their social experience on the Internet. Such cross-network syndication runs the risk of sharing of PII between the connected websites. For example, a user U using a social network X would like to avail the features of another social network Y to enhance his social experience on X. During registration, X and Y may have both asked for some PII from U which may or may not be the same. Due to privacy concerns, U may not want X and Y exchange her PII, explicitly or implicitly. Table I shows the PII attributes available with different social networks. Therefore, a social identity mapping service which maps U on X and Y could help U identify his own social footprint. Such a service would also increase privacy awareness about user's data to prevent him from PII leakage threats like identity thefts [7].
2) **Digital Marketing** – Twitter is an immensely popular social media for marketing and advertising [8]. Digital marketing teams create advertisement plans depending on customer requirements and generate marketing roadmaps based on customer engagement and demographics. Hence, it would be important for businesses to have access to as much customer information as available. Mapping users' social identities across networks can help businesses access and harness significantly more customer information. For example, a user U connects his Twitter profile to his Foursquare profile to enhance his location-sharing experience on Twitter. Mapping U from Twitter to Foursquare could aid businesses to provide location-sensitive advertisements and services. Therefore, social profile identity mapping can aid businesses for various purposes like targeted or contextual advertisements.

	Flickr	Four Square	You Tube	Last FM	Twit ter	Face Book
Username	✓	✓	✓	✓	✓	✓
Name	✓	✓	✓	✓	✓	✓
Gender		✓	✓	✓		✓
Profile Image	✓	✓	✓	✓	✓	✓
Relationship			✓			
Location		✓	✓	✓	✓	
School			✓			
Company			✓			
Occupation			✓			
Hobbies			✓			
Music			✓	✓		
Movies			✓			
Books			✓			
Contacts	✓	✓	✓	✓	✓	
Likes, Favorites	✓	✓	✓	✓	✓	
Photos	✓					
Age			✓	✓		
Videos			✓			
Description			✓		✓	
Last Web Access			✓			

TABLE I
TABLE SHOWS PUBLICLY AVAILABLE *Personally Identifiable Information* (PII) WITH DIFFERENT SOCIAL NETWORKING WEBSITES. THIS INFORMATION CAN BE ACCESSED BY UTILIZING THE API OF THE RESPECTIVE SOCIAL NETWORK. THE BLANK CELLS INDICATE THAT THE INFORMATION IS NOT PUBLICLY AVAILABLE.

The specific 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.

II. Research Contributions

In this section, we present the novel contribution of our work in context of existing literature on social profile identity mapping :

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 [9], [10], [11], [12], [13], [14]. *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 [15]. 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 [11]. In contrast, we focus on the *activity feed* sharing patterns in tweets which are generated due to profile connections as illustrated in Figure 2. Mining tweets to identity Twitter profiles on other networks is a novel contribution in context to previous work.

With respect to the above points, we provide a fresh perspective to the problem of using Twitter for *social profile identification of users across other social networks*.

III. Solution Approach

In this section, we first define our problem statement and then discuss our novel solution approach for social profile identity mapping by exploiting *activity feed* using Twitter.

A. Problem Statement

Let $u^i_{\mathcal{S}_j}$ denote a registered user on a social network $\mathcal{S}_j$. Let $\{\mathcal{P}_1 \ldots \mathcal{P}_n\}$ be the text patterns, originated due to *activity feeds*, for social networks $\{\mathcal{S}_1 \ldots \mathcal{S}_n\}$. Given a tuple $(u^i_T, \mathcal{T}, p)$ where u^i_T is a twitter user who has posted a set of tweets $\mathcal{T}$ containing a set of patterns $p \subseteq \{\mathcal{P}_1 \ldots \mathcal{P}_n\}$. Our goal is to find $(u^i_{\mathcal{S}_1} \ldots u^i_{\mathcal{S}_n})$ which are the profiles mappings of user u^i_T for the social networks $\{\mathcal{S}_1 \ldots \mathcal{S}_n\}$.

B. Proposed Approach

Our solution consists of a three-step framework – *Filter*, *Extract* and *Connect*. Figure 3 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. Figure 2 shows the common pattern in tweets generated as a result of sharing YouTube *activities* on Twitter. The first step of our framework requires identification of tweets with common text patterns for the respective service. The *Filter* block in Figure 3 shows the common text patterns occurring in tweets for external services like Flickr, Foursquare, LastFM and YouTube. We filter tweets 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 3 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 3 shows how profile information embedded in URLs can be used to link Twitter profiles to external services.

IV. Experimental Setup

In order to build our dataset, we leverage the Twitter Search REST API to collect a random sample of tweets matching our filters repeatedly during the period of 1st December 2011 to 31st December 2011 [16]. The Twitter Search API takes a keyword query as input and returns a maximum of 1500 tweets per day matching to the query. In this section, we detail our experimental setup according to the framework detailed in our solution approach.

A. Filter

We analyzed the auto-generated tweets generated by *activity feeds* for four external services – Flickr, Foursquare, YouTube and LastFM, and observed that there exists a common pattern to tweets generated via *activity feeds* for each external service. We leverage these patterns to create search queries which we then pass to the Twitter Search API. We repeatedly reformulate these queries until we are certain that the tweets retrieved for each query are a 100% match with the observed patterns. As described above, we then proceed to retrieve tweets matching these patterns at regular intervals via the RESTful Twitter API to build a database of tweets generated via activity feed sharing. We only need to identify the correct query string once for each service. Table II shows the query patterns used to *Filter* common tweet patterns and the number of tweets downloaded.

Apart from the text of the tweet, the Twitter Search API also returns meta-data like time, tweet id, source of the generated tweet and other related user information. We see that the *source* field id as a very good indicator for identification of auto-generated tweets. For example, tweets manually entered by users via the website are appended with the meta-information – *via Web*. Figure 2 also shows the source of the *activity feed* auto-generated tweet from YouTube as *via*

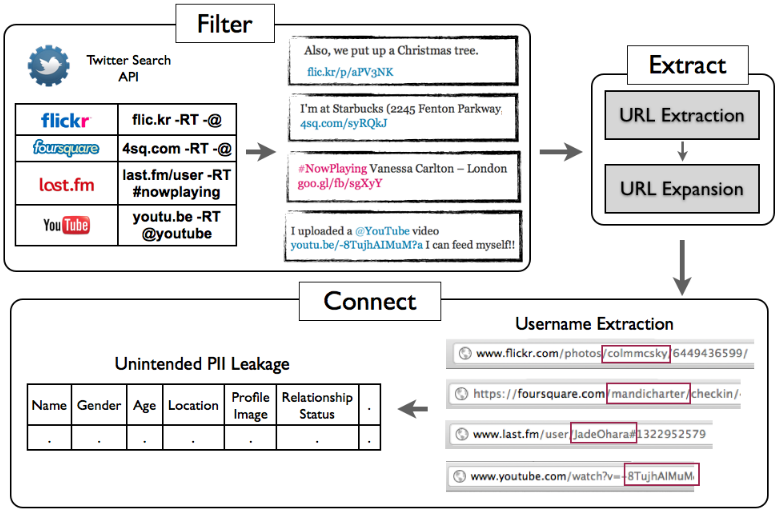

Fig. 3. **Three step framework – *Filter*, *Extract* and *Connect* – for our proposed solution approach.**

Social Network	Input Query	Number of Tweets
Flickr	flic.kr -RT -@	43438
Foursquare	4sq.com -RT -@	43245
YouTube	youtu.be -RT @youtube	43319
LastFM	last.fm/user -RT #nowplaying	13037
Total		**143039**

TABLE II
QUERIES GIVEN AS INPUT TO THE TWITTER SEARCH API FOR EACH SOCIAL NETWORK AND THE NUMBER OF TWEETS COLLECTED IN OUR DATASET.

Social Network	User Information in URL
Flickr	username or user id
Foursquare	username
YouTube	video id
LastFM	username

TABLE III
PROFILE INFORMATION AVAILABLE IN THE URL FOR DIFFERENT SOCIAL NETWORKS.

Google. Figure 4 shows the distribution of sources for the tweets in our collected dataset. We see that the major distribution of the tweets in our dataset are auto-generated by using the *activity feed* sharing features by external services like Flickr and YouTube. A small percentage of tweets are generated from other sources like mobile clients, desktop clients, social plugins and web applications. Due to immense popularity of Twitter, a large number of external applications and clients like TweetDeck[2] and HootSuite[3] have sprung up. Such external applications and clients provide a host of additional features to *Twitterers* including support of *activity feed* sharing. However, as is visible in Figure 4 these external applications make up a small distribution of our dataset.

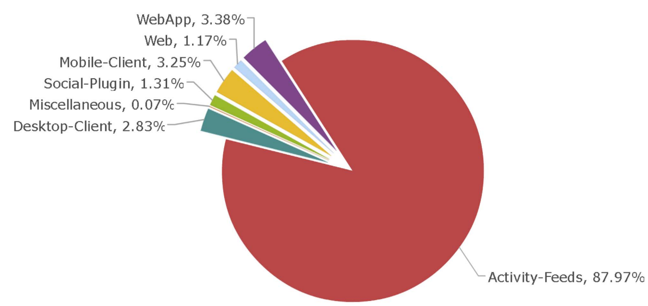

Fig. 4. **Distribution of sources from which the tweets in our dataset were generated.**

B. Extract

In this step, we extract the short URL from the tweets and expand these short URLs. We then extract the available profile information from the URL. Table III shows the information available in the URL for the different social networks in our dataset.

C. Connect

In the final step, we utilize the available profile information to connect users to their respective external services. In addition, we use the publicly available APIs for YouTube, LastFM, Foursquare and Flickr and extract publicly available information for each of the mapped profiles. In total, we were able to map 69,496 Twitter profiles to at least one other social network. The Foursquare API (in addition to retrieval of Foursquare user information) also allows access to a given user's usernames on Twitter and Facebook, if available.[4] For example, if Alice is registered on Foursquare and has connected her Twitter and/or Facebook accounts to her Foursquare

[2] http://www.tweetdeck.com/
[3] http://hootsuite.com/
[4] https://developer.foursquare.com/docs/responses/user

account; the Foursquare API returns the usernames/user-id of Alice on Twitter and Facebook. Therefore, we map these Twitter profiles to their Facebook accounts, if available, in addition to their Foursquare profiles.

V. Results

In this section, we outline our experimental results and analyze these results.

A. Social Profile Identity Mapping

Table IV 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 1500 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.

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

TABLE IV
Uniquely identified Twitter profiles across external services like Flickr, Foursquare, YouTube, LastFM and Facebook.

Our solution approach is also capable to map Twitter profiles to more than one service. Table V shows the number of Twitter profiles our solution approach could map to the number of services.

Number of Social Networks	Number of users mapped
2	86430
3	17216
4	97

TABLE V
Number of Twitter users mapped to the social networks – Flickr, Foursquare, YouTube, LastFM and Facebook.

B. 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 [5], [6], [17], [18], [19], [20]. Table VI 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.

Irani *et al.* propose a measure named "Normalized Attribute Leakage" in order to quantify the PII attribute leakage of users [5], [6]. We use the same measure to quantify the PII attribute leakage in our dataset under two settings – (1) with only information from Twitter profiles viz. without *WhACKY!* , (2) with the social profiles mapped to external services by *WhACKY!*. Figure 5 shows the "Normalized Attribute Leakage" for a subset of sensitive PII attributes under both the settings. Similar to our results, previous studies also see an increase in attribute leakage for all PII due to social profile mapping [5], [6]. The difference in the two bars indicate the increase in PII leakage for the particular attribute after profile mapping. We notice the highest increase in PII leakage for the Gender, Age and Location attributes. Therefore we can conclude that mapping Twitter profiles across other social networks can reveal more PII about a user.

	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%	
Relationship			14.52%		
Location		97.67%	99.81%	97.33%	
School			20.01%		
Company			21%		
Occupation			32.46%		
Hobbies			31.8%		
Music			26.69%	50%	
Movies			20.33%		
Books			18.56%		
Contacts	89.87%	99.37%	80.74%	89.47%	
Likes Favorites	75.17%	87.85%	25.22%	93.06%	
Photos	99.86%				
Videos			99.78%		
Description			56.27%		
Last Web Access			98.07% 98.07%		

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

In order to investigate if the PII leakage is unintended by the user, we manually inspect Facebook profile pages for 100 random users in our dataset for whom age, location and relationship status are available in one or the other linked service (except Facebook). We observe that 68 users do not list their age, 77 users hide their current location and 80 hide their relationship status on their public Facebook profiles. As Facebook makes this information available to all Facebook users by default, these attributes have been made non-public by users purposely opting-out [21]. This strongly suggests that the PII leakage we observe is indeed unintended.

C. How unique are usernames?

We study the uniqueness of Twitter profile usernames to their mapped networks. Table VII shows the percentage of Twitter profiles which have matching usernames on other social networks. Similar to previous studies, we notice that there is a significant amount of overlap in the usernames used by *Twitterers* on external services [12]. We also observe

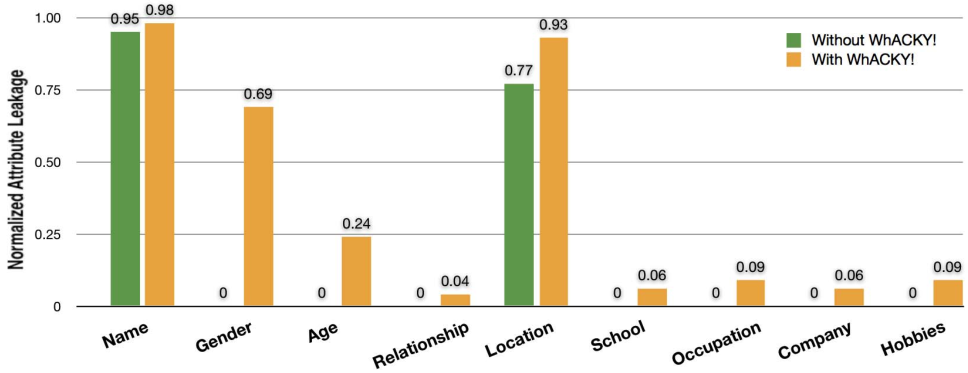

Fig. 5. ***Normalized attribute leakage* before and after the Twitter profiles are mapped. The x-axis contains a subset of sensitive PII and the y-axis indicates the *normalized attribute leakage*.**

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.

Social Network	Number of matching usersnames
Twitter – Flickr	3085 (21.88%)
Twitter – Foursquare	31610 (96.83%)
Twitter – YouTube	6883 (30.36%)
Twitter – LastFM	31 (40.79%)
Twitter – Facebook	4702 (27.77%)
Total	**46311/69496 = 66.63%**

TABLE VII
NUMBER OF TWITTER PROFILES WITH MATCHING USERNAMES ACROSS DIFFERENT SOCIAL NETWORKS IN OUR DATASET.

It must be noted that our approach is not a function of *usernames* (or any other profile attribute) and hence, is able to detect a large proportion of profiles which have no *usernames* in common. As explained earlier, our methods rely only on the *activity feed* sharing patterns of a user, which provide explicit links to the user's profile on other networks. Therefore, our solution is able to identify social profile mappings despite having non-matching usernames and other attributes without compromising compromising on accuracy.

D. Duplicate Profile Detection

We observe that a few *Twitterers* have multiple Twitter profiles but connect their Twitter profiles to the same external service. For example, Alice wants to create two Twitter profiles to demarcate her professional and personal interests. However, she just has one Flickr profile and shares the same *activity feeds* to both her Twitter profiles. We notice the presence of such profiles and link the external service to both of the Twitter profiles. Table VIII shows the number of duplicate Twitter profiles on each external service. We see that **0.54%** of the Twitter profiles in our dataset are duplicate profiles. Therefore, *activity feed* patterns could play a key complimentary role in solutions to detect duplicate users who operate multiple Twitter profiles.

Social Network	Number of duplicate Twitter profiles
Flickr	255
Foursquare	30
YouTube	85
LastFM	5
Facebook	0
Total	**375/69496 = 0.54%**

TABLE VIII
NUMBER OF DUPLICATE TWITTER PROFILES IN OUR DATASET CATEGORIZED ACCORDING TO THE MAPPED EXTERNAL SERVICE.

E. Evaluation

We compare our solution with that provided by a commercial service – *FlipTop* [5] – which provides a feature for social identity mapping. *FlipTop* is a leading social intelligence service which provides social information like social profile mapping about customers to businesses. According to their website, *FlipTop* collects this data from business data partner services and web crawls. We queried the Twitter usernames in our collected data to *FlipTop* API to evaluate the effectiveness of our approach. Table IX shows the number of mapped users across each external service for our proposed solution approach versus *FlipTop*.

Social Network	Number of Users Mapped	
	WhACKY!	*FlipTop*
Flickr	14064	2416
Foursquare	32646	4570
YouTube	22672	1217
LastFM	76	0
Facebook	16934	3403

TABLE IX
NUMBER OF SOCIAL PROFILES MAPPED ON EACH EXTERNAL SERVICE FOR ***WhACKY!*** AND ***FlipTop***.

WhACKY! is able to map more users for every external service than *FlipTop*. We argue that the use of *activity feed* patterns for mapping helps *WhACKY!* collect more and up to date information than traditional web crawls. It must be noted that the tweets resulting due to shared *activity feeds* are

[5] http://www.fliptop.com/

auto-generated and contain implicit links to a users profile on external social networks like YouTube, Flickr and Foursquare. Hence, the nature of our proposed solution requires no evaluation and is deterministic rather than probabilistic. The accuracy of our solution approach is directly dependent on the implementation of the matching patterns $\mathcal{P}$. Our experiments show that identification and utilization of these patterns is not only cheap but also easy. Therefore, our proposed solution approach guarantees 100% true positive mappings of profiles. Also, *WhACKY!* can be used to enhance the accuracy of previous approaches and alleviate the *guesswork* required in these approaches.

VI. Related Work

In this section, we review the closely related literature and position our work with respect to them.

A. Profile Information Based Methods

Profile Information or PII based methods are the most popular methods in literature. Motoyama and Varghese use web crawling techniques and scrape HTML pages linked to the respective profile to collect PII attributes for every user [10]. Vosecky *et al.* map profiles by matching PII attributes like username, name, hometown across social networks [22]. Balduzzi *et al.* exploit the commonly provided e-mail search feature by social networks to discover and map profiles across networks [23]. Carmagnola *et al.* use a PII attribute based weighted matching method to find candidate profiles to be mapped [24]. Perito *et al.* investigate the uniqueness of usernames of a user across multiple social networks [12].

In all of the above profile information based approaches, the solution is dependent on a subset or the complete PII attributes. Therefore, they are sensitive to conflicting PII across networks. In contrast to these approaches, *WhACKY!* does not utilize profile information and hence, is not PII sensitive. *WhACKY!* is able to map social profiles with little or no matching PII across social networks.

B. Network Based Methods

Narayanan and Shmatikov propose a network topology based algorithm to *de-anonymize* user profiles across social networks [11]. Labitzke *et al.* use the publicly available friend network to match social profiles across networks [13].

Network based solutions utilize the friend structure of the network to disambiguate the user across social networks. However, network based approaches may not be feasible on social networks where the *friend network* of a user is not publicly available. Moreover, due to the rise of niche social networks like LinkedIN, the friends of users on both networks may not overlap. Thus, such networks require some amount of probabilistic guessing. In contrast, our solutions do not rely on the network information and therefore, require no guess work.

C. Folksonomy Based Methods

Szomszor *et al.* use tag-clouds from multiple folksonomies to link social profiles across networks [14]. Iofciu *et al.* use tags to match social identities across different social networks [9].

Folksonomy based methods rely on the tags generated by users across social networks. They hypothesize that tag behaviors are signatures of users and hence, profiles across social networks which have similar tag distributions across networks are likely to be the same. *Folksonomy* based methods are only applicable on social networks which allow users to define and use tags. Social Networks like Twitter, Facebook and Foursquare do not allow the use of tags and therefore, *folksonomy* based approaches can not be used.

VII. Discussion

In this section, we further discuss our observations and outline the advantages & limitations of our solution approach.

A. Foursquare API – Privacy Issues

During our experiments, we observe that Foursquare API provides an API endpoint to retrieve the *twitter username*, *facebook username*, *e-mail* and *phone number* given a *foursquare user-id*, if available publicly. Foursquare *user-id*'s are n-digit serial numbers assigned (without choice) to users apart from the usernames they choose at the time of account creation. An adversary could serially input user-ids starting from 1 to 15 Million (the number of users on Foursquare as per January 2012) and collect the corresponding *twitter username*, *facebook username*, *e-mail* and *phone number* of all the users, if available [25]. Hence, an adversary could use this to map users across Foursquare, Facebook and Twitter with minimal effort to collect huge amount of PII. Therefore, this Foursquare API endpoint raises user data privacy concerns.

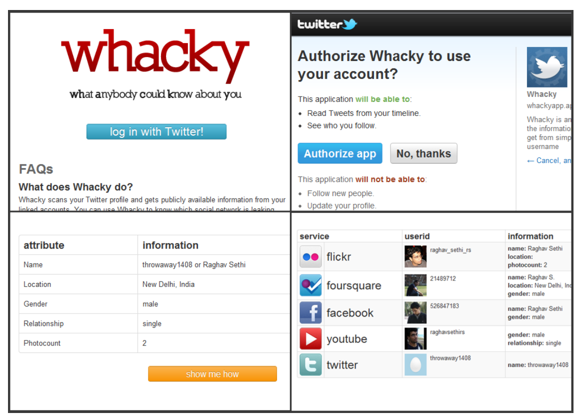

Fig. 6. **Screenshots of our web application *WhACKY!* with the standard OAuth flow and linked profiles.**

B. WhACKY! – Web Application

We developed a web application *WhACKY!* (acronym of *What Anyone Could Know about You*) to help increase data privacy awareness amongst *Twitterers*. The application can be accessed at *http://whackyapp.appspot.com/*. It helps users understand which of their linked accounts leak attributes such as age, location and relationship status. The application also demonstrates that our approach is computationally cheap and can link a Twitter account to four external services within seconds even on a limited Platform as a Service(PaaS) cloud (Google App Engine) [26]. The app uses OAuth flow to ensure users can only see external services linked to their own Twitter accounts (as we do not want to contribute to privacy violations) but there is no technical reason why the leaked attributes for any twitter username cannot be displayed. Figure 6 shows screenshots of our web application.

C. Advantages and Limitations

All our experiments were run on one machine with 4GB memory and 2.4GHz processor and therefore, is computationally cheap. In addition, 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*. Our proposed solution approach can be used as a complimentary solution to previous approaches and can help increase their overall accuracy.

VIII. Conclusion

We present a cheap and elegant solution to link Twitter profiles across external social network services. 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. We compare our approach to a popular commercial social profile mapping service and demonstrate the efficiency of our approach. 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. We also develop a web application to help increase user data privacy awareness amongst *Twitterers*.

References

[1] search+photos. [Online]. Available: http://blog.twitter.com/2011/06/searchphotos.html

[2] I. Guy, M. Jacovi, E. Shahar, N. Meshulam, V. Soroka, and S. Farrell, “Harvesting with sonar: the value of aggregating social network information,” in *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ser. CHI ’08. New York, NY, USA: ACM, 2008, pp. 1017–1026.

[3] C. Gerlitz and A. Helmond, “Hit, link, like and share. organizing the social and the fabric of the web in a like economy,” in *DMI mini-conference*, 2011.

[4] C. Zhang, J. Sun, X. Zhu, and Y. Fang, “Privacy and security for online social networks: challenges and opportunities,” *Network, IEEE*, vol. 24, no. 4, pp. 13 –18, july-august 2010.

[5] D. Irani, S. Webb, K. Li, and C. Pu, “Large online social footprints– an emerging threat,” in *Computational Science and Engineering, 2009. CSE ’09. International Conference on*, vol. 3, aug. 2009, pp. 271 –276.

[6] D. Irani, S. Webb, C. Pu, and K. Li, “Modeling unintended personal-information leakage from multiple online social networks,” *Internet Computing, IEEE*, vol. 15, no. 3, pp. 13 –19, may-june 2011.

[7] H. YU. Bad phorm on privacy. [Online]. Available: https://freedom-to-tinker.com/blog/harlanyu/bad-phorm-privacy/

[8] K. Lacy, *Twitter Marketing for Dummies*, 2009, vol. 2nd, no. ISBN: 978-0-470-93057-1.

[9] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, “Identifying users across social tagging systems,” in *ICWSM*, 2011.

[10] M. Motoyama and G. Varghese, “I seek you: searching and matching individuals in social networks,” in *Proceedings of the eleventh international workshop on Web information and data management*, ser. WIDM ’09. New York, NY, USA: ACM, 2009, pp. 67–75.

[11] A. Narayanan and V. Shmatikov, “De-anonymizing social networks,” *CoRR*, vol. abs/0903.3276, 2009.

[12] D. Perito, C. Castelluccia, M. Kaafar, and P. Manils, “How unique and traceable are usernames?” in *Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, pp. 1–17.

[13] H. H. S. Labitzke, I. Taranu, “What your friends tell others about you: Low cost linkability of social network profiles,” in *5th International ACM Workshop on Social Network Mining and Analysis*, ser. SNA KDD ’11. San Diego, CA, USA: ACM, 2011, pp. 51–60.

[14] M. N. Szomszor, I. Cantador, and H. Alani, “Correlating user profiles from multiple folksonomies,” in *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, ser. HT ’08. New York, NY, USA: ACM, 2008, pp. 33–42.

[15] Youtube. (2012, May) Youtube press statistics. [Online]. Available: http://www.youtube.com/t/pressstatistics

[16] Twitter. Search rest api. [Online]. Available: https://dev.twitter.com/docs/using-search

[17] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, “All your contacts are belong to us: automated identity theft attacks on social networks,” in *Proceedings of the 18th international conference on World wide web*, ser. WWW ’09. New York, NY, USA: ACM, 2009, pp. 551–560.

[18] B. Krishnamurthy and C. E. Wills, “On the leakage of personally identifiable information via online social networks,” in *Proceedings of the 2nd ACM workshop on Online social networks*, ser. WOSN ’09. New York, NY, USA: ACM, 2009, pp. 7–12.

[19] B. Matthews and A. Esterline, “Personally identifiable information: Identifying unprotected pii using file-indexing search tools and quantitative analysis,” in *IEEE SoutheastCon 2010 (SoutheastCon), Proceedings of the*, march 2010, pp. 360 –362.

[20] E. Zheleva and L. Getoor, “To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles,” in *Proceedings of the 18th international conference on World wide web*, ser. WWW ’09. New York, NY, USA: ACM, 2009, pp. 531–540.

[21] E. F. Foundation. Facebook’s eroding privacy policy: A timeline. [Online]. Available: https://www.eff.org/deeplinks/2010/04/facebook-timeline

[22] J. Vosecky, D. Hong, and V. Shen, “User identification across multiple social networks,” in *Networked Digital Technologies, 2009. NDT ’09. First International Conference on*, july 2009, pp. 360 –365.

[23] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel, “Abusing social networks for automated user profiling,” in *Recent Advances in Intrusion Detection*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, pp. 422–441.

[24] F. Carmagnola, F. Osborne, and I. Torre, “User data distributed on the social web: how to identify users on different social systems and collecting data about them,” in *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, ser. HetRec ’10. New York, NY, USA: ACM, 2010, pp. 9–15.

[25] Foursquare, “Foursquare - about us.” [Online]. Available: https://foursquare.com/about/

[26] Google, “Google app engine.” [Online]. Available: https://developers.google.com/appengine/