

WhACKY! – What Anyone Could Know About You from Twitter

Student Name: Denzil Correa

IIITD-MTech-CS-12-PhD0903

April 13, 2012

Indraprastha Institute of Information Technology
New Delhi

Thesis Committee
Ashish Sureka (Chair)
Gaurav Gupta
Ashutosh Saxena

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science

©2012 IIITD-MTech-CS-12-PhD0903
All rights reserved

This research was partially funded by Department of Information Technology (DIT), India.

Keywords: Privacy, Social Media, Twitter, Profile Linking, PII Leakage

Certificate

This is to certify that the thesis titled “**WhACKY! – What Anyone Could Know About You from Twitter** ” submitted by **Denzil Correa** for the partial fulfillment of the requirements for the degree of *Master of Technology in Computer Science & Engineering* is a record of the bonafide work carried out by him under my guidance and supervision in the Information Management & Data Mining group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Professor Ashish Sureka

Indraprastha Institute of Information Technology, New Delhi

Abstract

Twitter is a popular micro-blogging website which allows users to post 140-character limit messages called tweets. Twitter users (also called *Twitterers*) post activity messages about their daily lives, opinions on current events and news, and even have conversations with other users. In addition, *Twitterers* also share various other information like photographs, videos and visited locations hosted on other external services like Flickr, YouTube and Foursquare. Therefore, tweets contain variety of information obtained from a combination of multiple sources. We demonstrate a cheap and elegant solution – *WhACKY!* – to harness this multi-source information 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 develop a web application based on *WhACKY!*¹ for perusal by *Twitterers* which can help them better understand unintended leakage of their PII.

¹<http://whackyapp.appspot.com/>

Acknowledgments

I would like to thank Raghav Sethi for lending a generous hand in the web application and feedback on various drafts of this thesis.

I am grateful to many individuals who have directly or indirectly helped me achieve this goal. No amount of words written on this page would be sufficient to quantify their advice, prayers, love and support for me. I would specially like to dedicate this work to – *Dadda* who couldn't live this day but would be immensely proud of, *Mumma* for being the best supermom the world doesn't know about, *Nikki* for the most loving sister a brother can get, *Mai* who probably wouldn't comprehend a word in this thesis but still appreciate what I do, '*qdLA4agkEzB316XFeKfQrQ==*' who has been my silent, emotional and moral support and *Thomas Uncle* for being there at the slightest inch of cloud surrounding me and my family.

Finally and importantly, this thesis is dedicated to *Ashish* who has been more than an advisor and a mentor. Your creativity amazes me, your work ethic inspires me and your attitude infects me. You are the man, I want to be.

Contents

1	Research Motivation and Aim	2
2	Related Work	6
2.0.1	Profile Information Based Methods	6
2.0.2	Network Based Methods	7
2.0.3	Folksonomy Based Methods	8
3	Research Contributions	9
4	Solution Approach	10
4.0.4	Problem Statement	10
4.0.5	Proposed Approach	10
5	Experimental Setup	12
5.0.6	Filter	12
5.0.7	Extract	14
5.0.8	Connect	14
6	Results	16
6.0.9	Social Profile Identity Mapping	16
6.0.10	Unintended Personal Information Leakage	16
6.0.11	How unique are usernames?	18
6.0.12	Duplicate Profile Detection	19
6.0.13	Evaluation	20
7	Discussion	21
7.0.14	Foursquare API – Privacy Issues	21
7.0.15	Advantages	21
7.0.16	Limitations	22
7.0.17	Web Application	22

List of Figures

1.1	The screenshot illustrates the <i>activity feed</i> sharing feature provided by YouTube. A user can choose to connect his YouTube account to Orkut, Twitter or Facebook. YouTube also provides the user multiple options to share specific activities such as upload, like, favorite) on the connected network.	3
1.2	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.	4
4.1	The figure represents the three step framework – <i>Filter</i> , <i>Extract</i> and <i>Connect</i> – for our proposed solution approach.	11
5.1	The figure shows the <i>activity feed</i> patterns observed in our dataset. We analyze these patterns to formulate our queries.	13
5.2	The pie chart depicts the distribution of sources from which the tweets in our dataset were generated.	14
6.1	The screenshot demonstrates how PII collected from different social networks can be aggregated to create an aggregated social footprint of the user.	18
6.2	Figure indicates the <i>normalized attribute leakage</i> before and after the Twitter profiles are mapped. The x-axis contains a subset of sensitive PII and the y-axis indicates the <i>normalized attribute leakage</i> . The difference in the two bars indicate the increase in PII leakage for the particular attribute after profile mapping. We observe an increase in PII leakage after profile mapping for all sensitive PII attributes.	19
7.1	The figure shows screenshots of our web application <i>WhACKY!</i> with the standard OAuth flow and linked profiles. The application is accessible at - http://whackyapp.appspot.com/	23

List of Tables

1.1	Table shows publicly available <i>Personally Identifiable Information</i> (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.	5
5.1	The table shows the queries given as input to the Twitter Search API for each social network and the number of tweets collected in our dataset. The query patterns were formulated after a one-time manual trial and error experiment.	13
5.2	The table shows the profile information available in the URL for different social networks.	14
6.1	The table shows the uniquely identified Twitter profiles across external services like Flickr, Foursquare, YouTube, LastFM and Facebook. . . .	16
6.2	The table shows the number of Twitter users mapped to the number of social networks : Flickr, Foursquare, YouTube, LastFM and Facebook. . . .	17
6.3	The table shows the percentage of publicly available PII attributes present across each service in our dataset. Blank cells indicate that the PII attributes were not publicly available.	17
6.4	The table shows the number of Twitter profiles which have matching usernames across different social networks in our dataset.	19
6.5	Table shows the number of duplicate Twitter profiles in our dataset categorized according to the mapped external service.	20
6.6	Table shows the number of social profiles mapped on each external service for <i>WhACKY!</i> and <i>FlipTop</i>	20

Chapter 1

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 doesn't 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 [5], and *Twitterers* often share content hosted on external services like LastFM, YouTube and Foursquare.

To demonstrate why a user would choose to exhibit the above-described behaviour, consider the following example. Dom Cobb creates an account on Twitter with the screen name *doco* to post his daily life activities and engage in conversations with fellow *Twitterers*. Dom feels the need to share interesting videos with her Twitter followers and discovers that Twitter doesn't allow users to directly share videos with each other. Dom is, however, aware that YouTube is a popular video sharing service. Therefore, Dom registers for an account on YouTube with the username *domc* after he discovers that the username *doco* is already in use by another user. Dom uploads his videos on YouTube and shares links to these videos with his followers on Twitter by manually copying links to his Twitter profile. However, Dom soon starts to find the task of manually updating his Twitter profile every time he uploads a YouTube video to be an arduous and mundane activity. Dom searches around the YouTube website for a solution and finds that YouTube provides a feature to connect her YouTube profile with his Twitter

¹<http://blog.twitter.com/2011/06/searchphotos.html>

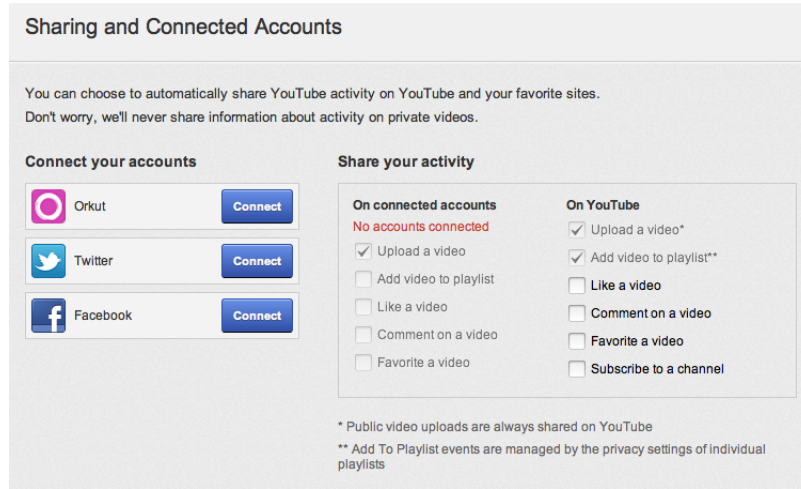


Figure 1.1: The screenshot illustrates the *activity feed* sharing feature provided by YouTube. A user can choose to connect his YouTube account to Orkut, Twitter or Facebook. YouTube also provides the user multiple options to share specific activities such as upload, like, favorite) on the connected network.

profile to automatically share his YouTube *activities* (uploads, favorites, likes) on Twitter.² Figure 1.1 shows the snapshot of the *activity feed* sharing feature provided by YouTube. Dom eagerly utilizes this feature and connects his Twitter and YouTube accounts. This allows Dom to automatically share his YouTube activities like video uploads with his Twitter followers. Dom need not update his Twitter profile manually as an auto-generated tweet is posted automatically every time he uploads a YouTube video. Figure 1.2 shows a tweet automatically generated as a result of a video upload on YouTube via the *activity feed* sharing feature. Similarly, Dom uses other external services like LastFM, Foursquare and Flickr to share music, interesting locations and photographs. Therefore, his profile contains diverse information from multiple external services like YouTube, LastFM and Foursquare.

Twitterers 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 [4]. However, the 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 [23]. 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 [7,8]. 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

²http://youtube-global.blogspot.in/2009/06/share-youtube-videos-on-facebook_11.html

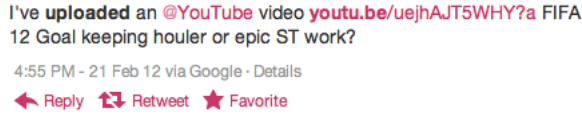


Figure 1.2: The screenshot illustrates an example of an auto-generated tweet posted on Twitter for a video upload on YouTube due to connection of a user’s YouTube account to his Twitter profile.

users to provide certain personally identifiable information (PII) in order to cater to legal issues as well as to ensure *hygiene* [13, 17]. 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 1.1 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.³

2. **Digital Marketing** – Twitter is an immensely popular social media for marketing and advertising [10]. 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.

The specific research aim of this thesis 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.

³<https://freedom-to-tinker.com/blog/harlanyu/bad-phorm-privacy>

	Flickr	FourSquare	YouTube	LastFM	Twitter	Facebook
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 1.1: 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.

Chapter 2

Related Work

In this section, we review the closely related literature and position our work with respect to them. We divide the related work into three classes depending on the methodology proposed viz. *Profile Information*, *Network* and *Folksonomy* based methods. We observe that *Profile Information* (or PII) based methods are the most popular methods in literature.

2.0.1 Profile Information Based Methods

Motoyama and Varghese propose a machine learning based method to link profiles across Facebook and MySpace [15]. They scrape HTML pages of user profiles on Facebook & MySpace and extract attributes of users like name, email, education, gender, age, country and city if available. They use these profile attributes to query the Facebook & MySpace search engines and simultaneously iteratively refine them to generate a set of candidate profiles. In order to find the final profile matches, Motoyama and Varghese train a machine learning classifier using *boosting* on a ground truth of 900 profiles and tested on a collection of 500 profiles. They are able to achieve a false positive rates of lesser than 5% on the test set. They also observe that ‘name’ and ‘name’ with ‘educational records’ lead to a large number of profile matches across Facebook and MySpace.

Vosecky *et al.* demonstrate a vector based similarity matching approach to match profiles across OSNs [21]. Each profile is considered a vector consisting of the profile fields such as name, data of birth and age where each field is assigned a weight. To match vectors an exact, partial or fuzzy approach is used to calculate the similarity between profiles. They experiment their approach on users across Facebook and StudiVZ and report 83% accuracy.

Balduzzi *et al.* exploit the commonly provided ‘e-mail search feature’ by social networks to discover and map profiles across networks [1]. They start with 10.4 Million seed e-mail addresses left on a dropzone of a compromised machine. They query eight social networks including Facebook, MySpace, Twitter, LinkedIn, Friendster, Badoo, Netlog and Xing with these e-mail addresses and discover 1.2 Million profiles. They observe that Facebook-MySpace, Facebook-LinkedIn and Facebook-Twitter combinations contain a large number of profiles. They also

report that users provide conflicting PII information across OSNs like relationship status, age and sex.

Carmagnola *et al.* use a PII attribute based weighted matching method to find candidate profiles to be mapped [3]. They query the OSNs and generate a set of candidate profiles with a score depending on the number of matched profile attributes. In order to narrow down to one profile, they make use of a conditional probability distribution based heuristic approach to disambiguate the candidates. They evaluate their approach on a ground truth dataset of MySpace and Flickr.

Perito *et al.* use usernames to connect profiles across multiple social networks [18]. They use *Information Surprisal* to study the uniqueness of usernames for more than 10 Million profiles. They observe that usernames are unique enough to identify profiles across networks. In order to link profiles with non-unique usernames, they train a machine learning classifier using two approaches – Markov Chains and TF-IDF. They demonstrate that Markov Chains perform better than the tf-idf approach. Zafarani and Liu use also use usernames to link profiles across online social network websites [22]. They evaluate their approach on 12 different OSNs like Del.icio.us, Digg, Flickr, Furl, LastFM, MySpace, Reddit, StumbleUpon, Twitter, YouTube, Technorati and MyBlogLog & achieve 66% accuracy.

It must be noted that approaches based on profile information, are dependent on a subset or the complete set of profile (or PII) attributes. Therefore, they are sensitive to conflicting PII across networks. Some approaches take these conflicts into consideration and try to provide the most accurate guess. 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. In addition, there’s no guess work involved in the profile mapping.

2.0.2 Network Based Methods

Narayanan and Shmatikov propose a network topology based algorithm to *de-anonymize* user profiles across social networks [16]. They consider the social network as a graph and map unknown profiles in the network by exploiting the knowledge of known profiles and their auxiliary information obtained from the home network as well as external networks. They experiment on three social networks – Twitter, Flickr and LiveJournal – and show that they are successfully able to *de-anonymize* user profiles with a low error rate.

Labitzke *et al.* use the publicly available friend network to match social profiles across networks [19]. They use multiple metrics to calculate the overlap between friend networks for profiles across external services. They show that a small overlap in friend lists is sufficient to link profiles across OSNs.

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 don't rely on the network information and therefore, require no guess work.

2.0.3 Folksonomy Based Methods

Szomszor *et al.* use tag-clouds from multiple folksonomies to link social profiles across networks [20]. They compare the tag distribution patterns of users across social networks in addition to other profile information like age, gender, sex and name. They perform experiments on two social networks, *del.icio.us* and *flickr*, and show that tag distribution overlaps can help linking social profiles across networks.

Iofciu *et al.* investigate multiple ways to link profiles using tags, username and tags + username [6]. They employ the TF, TFIDF, BM25, BM25 specific IDF strategies to link profiles via tags & ExactMatch, Jaccard, SmithWaterman, Levenshtein, LCS strategies to link profiles via usernames. In the case of using tags + usernames, they appropriately combine the strategies as well. They show that profiles could be linked across OSNs by exploiting their tagging behavior using the BM25 site specific IDF strategy.

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 don't allow the use of tags and therefore, *folksonomy* based approaches can't be used.

Chapter 3

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 [6, 15, 16, 18–20]. *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.¹ 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 [16]. Some other study uses Twitter to understand unintentional PII leakage [11, 14]. In contrast, we focus on the *activity feed* sharing patterns in tweets which are generated due to profile connections as illustrated in Figure 1.2. Mining tweets to identify 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*.

¹http://www.youtube.com/t/press_statistics

Chapter 4

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.

4.0.4 Problem Statement

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

4.0.5 Proposed Approach

Our solution consists of a three-step framework – *Filter*, *Extract* and *Connect*. Figure 4.1 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. Figure 1.2 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 4.1 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 4.1 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 4.1 shows how profile information embedded in URLs can be used to link Twitter profiles to external services.

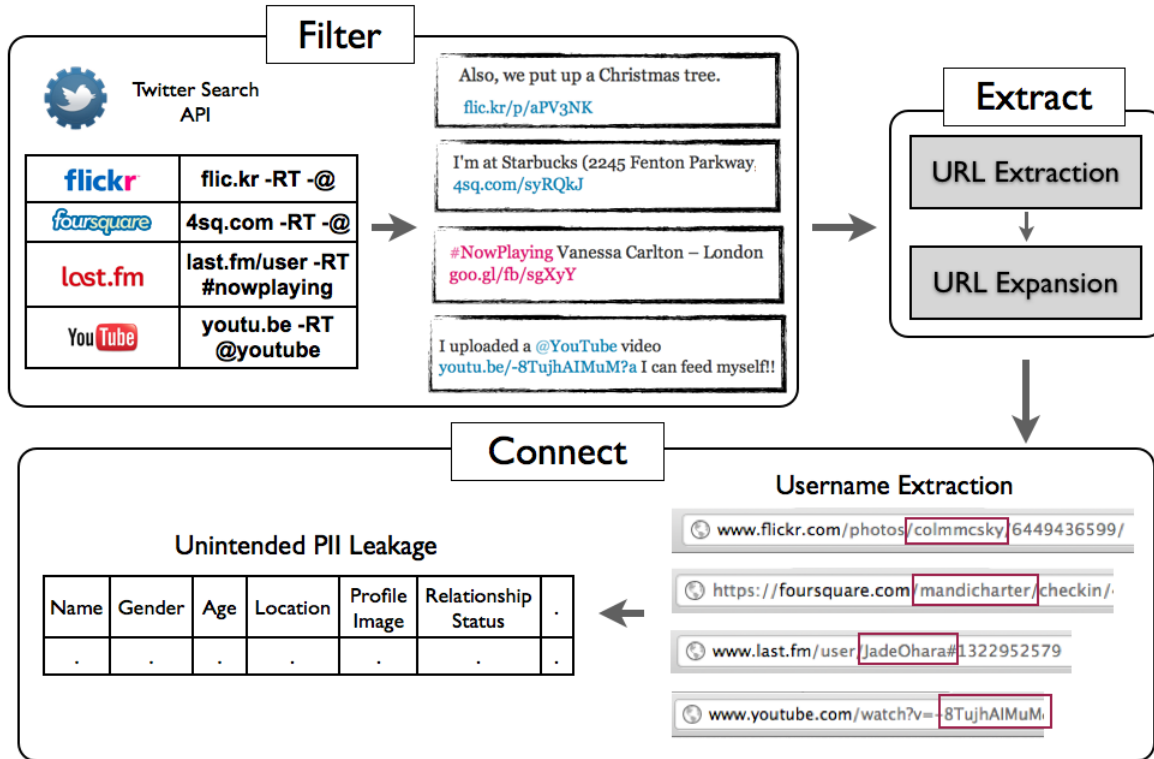


Figure 4.1: The figure represents the three step framework – *Filter*, *Extract* and *Connect* – for our proposed solution approach.

Chapter 5

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 2011.¹ 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.

5.0.6 Filter

A twitter user can post a tweet about his own activity (such as uploading a video on a Video sharing website or liking a photo on a social networking website) or about the activity of another user (such as a video uploaded by another user). The filter step applies regular expression and string matching to identify only those tweets that mention the activity of the respective twitterer on another website and not the activities of another user.

We analyze the auto-generated tweets generated by *activity feeds* for four external services – Flickr, Foursquare, YouTube and LastFM, and observe 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. Figure 5.1 show the *activity feed* sharing patterns we observe for each of the social network services – Flickr, Foursquare, YouTube and LastFM. Table 5.1 shows the query patterns used to *Filter* common tweet patterns and the number of tweets downloaded.

Our query patterns are formulated to ensure that all filtered tweets are always useful for the other two process blocks *Extract* and *Connect*. Hence, these patterns may not capture all the useful tweets and contain false negatives. However, our aim is to achieve a 100% accuracy on

¹<https://dev.twitter.com/docs/using-search>

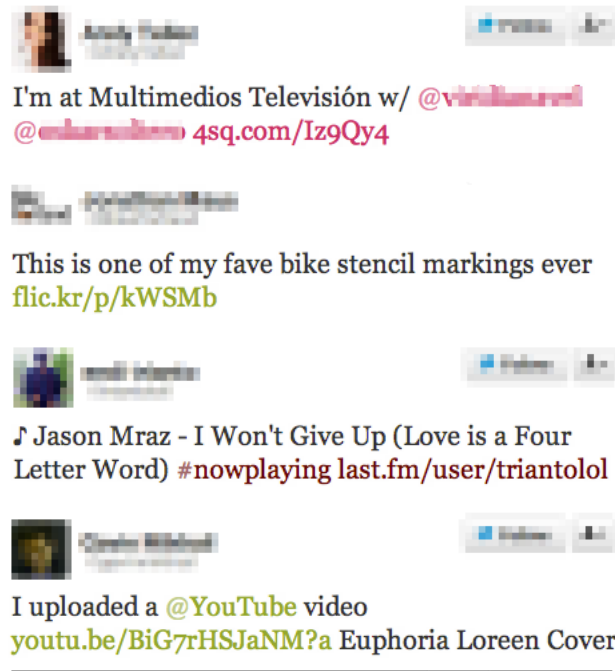


Figure 5.1: The figure shows the *activity feed* patterns observed in our dataset. We analyze these patterns to formulate our queries.

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 5.1: The table shows the queries given as input to the Twitter Search API for each social network and the number of tweets collected in our dataset. The query patterns were formulated after a one-time manual trial and error experiment.

linked profiles and therefore, we adopt such a conservative approach by making a trade off. The conservative natures of our experiments also reveal insights on the ease to which an adversary can gather PII.

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 1.2 also shows the source of the *activity feed* auto-generated tweet from YouTube as *via Google*. Figure 5.2 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 TweetDeck² and HootSuite³ 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 5.2 these external applications make up a small distribution of our dataset.

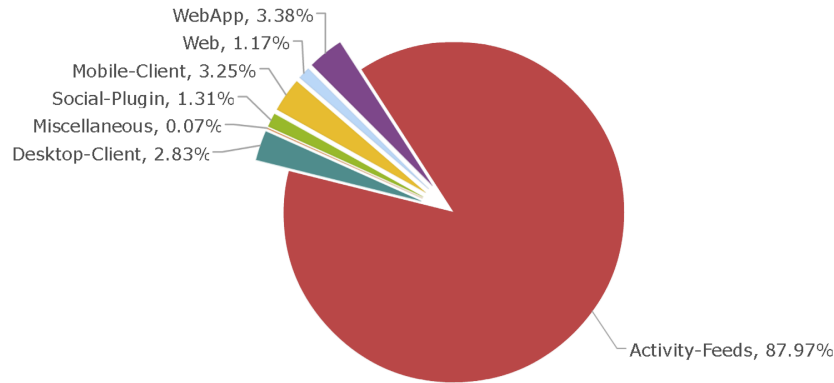


Figure 5.2: The pie chart depicts the distribution of sources from which the tweets in our dataset were generated.

5.0.7 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 5.2 shows the information available in the URL for the different social networks in our dataset.

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

Table 5.2: The table shows the profile information available in the URL for different social networks.

5.0.8 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

²<http://www.tweetdeck.com/>

³<http://hootsuite.com/>

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 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.

⁴<https://developer.foursquare.com/docs/responses/user>

Chapter 6

Results

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

6.0.9 Social Profile Identity Mapping

Table 6.1 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 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 6.1: **The table shows the 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 6.2 shows the number of Twitter profiles our solution approach could map to the number of services.

6.0.10 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 [2, 7–9, 12,

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

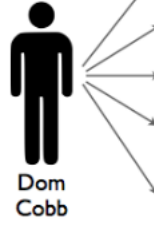
Table 6.2: The table shows the number of Twitter users mapped to the number of social networks : Flickr, Foursquare, YouTube, LastFM and Facebook.

24]. Table 6.3 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 don't count that attribute towards the final percentage.

	Flickr	FourSquare	YouTube	LastFM	Facebook
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%		

Table 6.3: The table shows the percentage of publicly available PII attributes present across each service in our dataset. Blank cells indicate that the PII attributes were not publicly available.

Irani *et al.* propose a measure named “Normalized Attribute Leakage” in order to quantify the PII attribute leakage of users [7,8]. “Normalized Attribute Leakage” is a metric to measure information that one can uncover about a specific attribute or feature given a web users social foot print. We use the ‘Normalized Attribute Leakage’ to quantify the PII attribute leakage in our dataset under two settings – (1) with only information from Twitter profiles viz. without



	Username	Name	Gender	Description	Relationship Status	Location	Age
foursquare	dom.cobb	--	Male	--	--	Paris, France	--
YouTube	domc	Dom Cobb	--	an engineer...	married	--	--
last.fm	dcobb	--	Male	--	--	--	32
twitter	doco	Dom K C	--	loves cricket...	--	France	--
facebook	dom86	Dom Cobb	Male		--	--	32

Figure 6.1: The screenshot demonstrates how PII collected from different social networks can be aggregated to create an aggregated social footprint of the user.

WhACKY!, (2) with the social profiles mapped to external services by *WhACKY!*. Figure 6.2 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 [7, 8]. 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. Figure 6.1 shows an example of how PII can be aggregated from different social networks to create an aggregated social footprint.

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 ¹, these attributes have been made non-public by users purposely opting-out. This strongly suggests that the PII leakage we observe is indeed unintended.

6.0.11 How unique are usernames?

We study the uniqueness of Twitter profile usernames to their mapped networks. Table 6.4 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 [18]. 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.

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

¹<https://www.eff.org/deeplinks/2010/04/facebook-timeline>

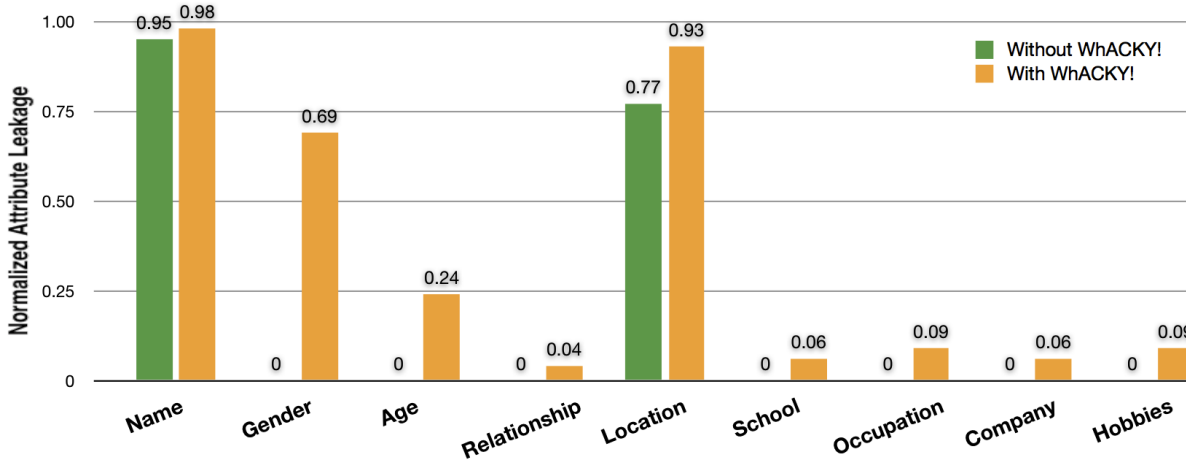


Figure 6.2: Figure indicates the *normalized attribute leakage* before and after the Twitter profiles are mapped. The x-axis contains a subset of sensitive PII and the y-axis indicates the *normalized attribute leakage*. The difference in the two bars indicate the increase in PII leakage for the particular attribute after profile mapping. We observe an increase in PII leakage after profile mapping for all sensitive PII attributes.

Social Network	Number of matching usernames
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 6.4: The table shows the number of Twitter profiles which have matching usernames across different social networks in our dataset.

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.

6.0.12 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 6.5 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 6.5: Table shows the number of duplicate Twitter profiles in our dataset categorized according to the mapped external service.

6.0.13 Evaluation

We compare our solution with that provided by a commercial service – *FlipTop*² – 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 6.6 shows the number of mapped users across each external service for our proposed solution approach versus *FlipTop*.

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

Table 6.6: Table shows the 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 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.

²<http://www.fliptop.com/>

Chapter 7

Discussion

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

7.0.14 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.¹ 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.² Hence, an adversary could collect a huge number of profiles mapped across Foursquare, Facebook and Twitter with minimal effort. This Foursquare API endpoint reveals user data privacy concerns and could be exploited by an adversary to cheaply gain access to huge amount of personally identifiable information.

7.0.15 Advantages

All our experiments were run on one machine with 4GB memory and 2.4GHz processor. Therefore, our solution approach does not require a high amount of computing resources. In addition, our solution approach 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 don't clearly fit the pattern identified. The proposed solution is not limited to the number of tweets posted by a user. The amount of time taken to perform profile linking is a function of the number of tweets as naturally more tweets will increase the processing time but the number of tweets per user does not influence accuracy.

¹<https://developer.foursquare.com/docs/responses/user>

²<https://foursquare.com/about/>

As shown in Figure 4.1, the proposed solution is a generic framework consisting of three main components: Filter, Extract and Connect (key sequential steps in the processing pipeline). One of the advantages of the proposed solution is scalability (ability to handle load without performance degradation) as some of the tasks are independent and can be executed in parallel. The filter step can be performed in parallel as each tweet is independent of the other tweets in a stream. Similarly, the URL extraction step, URL expansion and Connect can also be executed in parallel for different tweets. While for the same tweet, the various steps need to perform in sequence, the steps can be executed in parallel for different tweets (as each tweet is independent of each other in the content of the problem). The proposed solution does not implement parallel processing as the main focus of the work was to investigate the feasibility and accuracy of the approach. However, parallel execution and load balancing techniques can be employed to bring scalability in the overall processing pipeline.

In a nutshell, our proposed solution approach is *Cheap, Elegant*, requires *No Evaluation*, *Scalable* and guarantees *100% Accuracy*.

7.0.16 Limitations

One of the inherent limitations of the proposed solution is that it can link profiles of only those users that use activity sharing functionality. The solution will not be able to link profiles if there are no data flows (auto generated activity feed) between two Web 2.0 platforms as the first step in the process is to filter activity sharing tweets. The proposed solution will work for only those cases where the activity sharing tweets are available. Another 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.

We don't consider manual tweets posted by users as evident from our query patterns. While it is true that there can be other tweets (manually posted) that can be exploited to link profiles, the research motivation and aim of the work is to focus on automatically generated activity feed (this is focus of the research). Currently manually generated tweets are out of the scope of the work even though it is a good idea and logical extension of the work. We also don't make use of semantics to enhance the filtering technique to access more tweets and net more users. However, processing manual tweets or incorporating natural language processing techniques leads to a trade off on the elegance and computational expense of the solution.

7.0.17 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. We host our application on a cloud service provided by *Google App Engine*.³ We use *Twitter Bootstrap* to design the UI elements of our application.⁴ Figure 7.1 shows screenshots of the working of our web application. The application implements the standard *OAuth* flow to ensure that we don't store any user information. The application also demonstrates that our approach is computationally cheap and can link a Twitter account to four external services — Flickr, Foursquare, Facebook and YouTube — within seconds even on a limited Platform as a Service(PaaS) cloud (Google App Engine). The use of *OAuth* ensures that 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. We plan to add more features and accessibility options to the application in the near future.

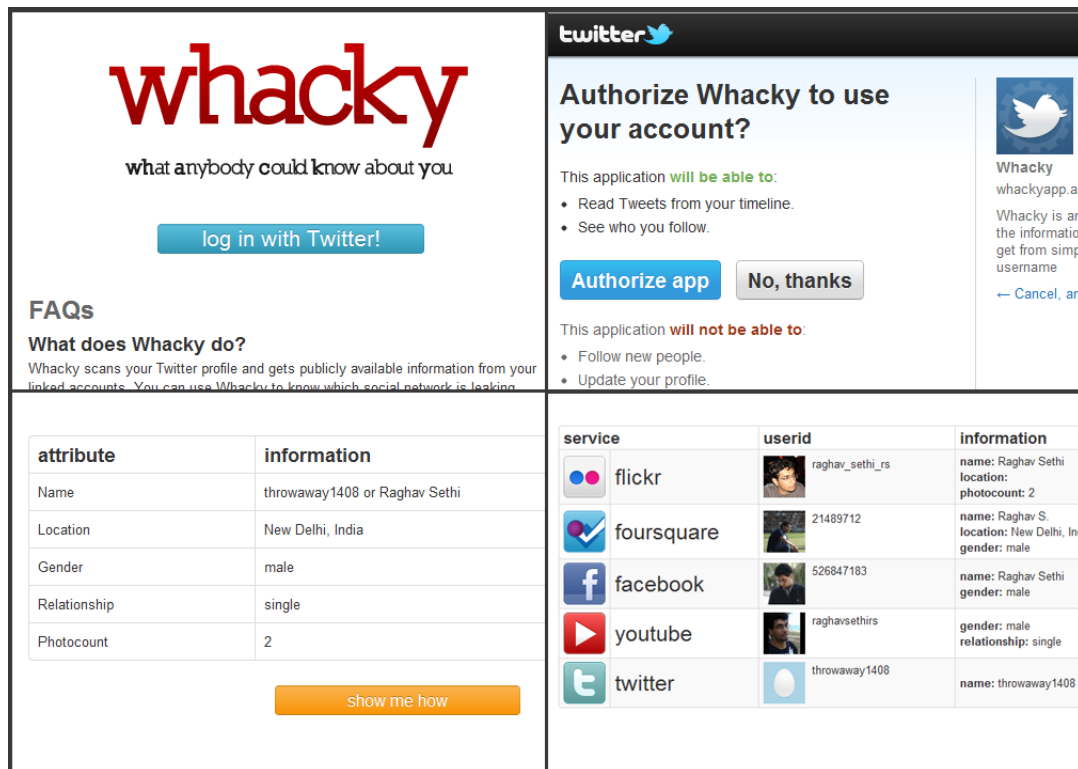


Figure 7.1: The figure shows screenshots of our web application *WhACKY!* with the standard *OAuth* flow and linked profiles. The application is accessible at - <http://whackyapp.appspot.com/>.

³<https://developers.google.com/appengine/>

⁴<http://twitter.github.com/bootstrap/>

Chapter 8

Conclusion

We present a cheap and elegant solution to link Twitter profiles across external social network services. We observe that *Twitterers* connect their Twitter profiles to other social networks like Flickr, YouTube and Foursquare to enhance their social experience. We exploit the text patterns in tweets which are auto-generated as a result of such connections, also called as *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. Moreover, our solution 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 – *WhACKY!* – based on our solution approach to help *Twitterers* easily detect attribute leakage caused by activity feed sharing and increase user data privacy awareness amongst *Twitterers*.

Bibliography

- [1] BALDUZZI, M., PLATZER, C., HOLZ, T., KIRDA, E., BALZAROTTI, D., AND KRUEGEL, C. Abusing social networks for automated user profiling. In *Recent Advances in Intrusion Detection*, Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, pp. 422–441.
- [2] BILGE, L., STRUFE, T., BALZAROTTI, D., AND KIRDA, E. 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* (New York, NY, USA, 2009), WWW '09, ACM, pp. 551–560.
- [3] CARMAGNOLA, F., OSBORNE, F., AND TORRE, I. 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* (New York, NY, USA, 2010), HetRec '10, ACM, pp. 9–15.
- [4] GERLITZ, C., AND HELMOND, A. Hit, link, like and share. organizing the social and the fabric of the web in a like economy. In *DMI mini-conference* (2011).
- [5] GUY, I., JACOVI, M., SHAHAR, E., MESHULAM, N., SOROKA, V., AND FARRELL, S. 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* (New York, NY, USA, 2008), CHI '08, ACM, pp. 1017–1026.
- [6] IOFCIU, T., FANKHAUSER, P., ABEL, F., AND BISCHOFF, K. Identifying users across social tagging systems. In *ICWSM* (2011).
- [7] IRANI, D., WEBB, S., LI, K., AND PU, C. Large online social footprints—an emerging threat. In *Computational Science and Engineering, 2009. CSE '09. International Conference on* (aug. 2009), vol. 3, pp. 271 –276.
- [8] IRANI, D., WEBB, S., PU, C., AND LI, K. Modeling unintended personal-information leakage from multiple online social networks. *Internet Computing, IEEE 15*, 3 (may-june 2011), 13 –19.

- [9] KRISHNAMURTHY, B., AND WILLS, C. E. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks* (New York, NY, USA, 2009), WOSN '09, ACM, pp. 7–12.
- [10] LACY, K. *Twitter Marketing for Dummies*, vol. 2nd. 2009.
- [11] MAO, H., SHUAI, X., AND KAPADIA, A. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society* (2011), ACM, pp. 1–12.
- [12] MATTHEWS, B., AND ESTERLINE, A. 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.
- [13] MCCALLISTER, E. *Guide to Protecting the Confidentiality of Personally Identifiable Information*. DIANE Publishing, 2010.
- [14] MEEDER, B., TAM, J., KELLEY, P., AND CRANOR, L. Rt@ iwantprivacy: Widespread violation of privacy settings in the twitter social network. In *Web 2.0 Privacy and Security Workshop, IEEE Symposium on Security and Privacy* (2010).
- [15] MOTOYAMA, M., AND VARGHESE, G. I seek you: searching and matching individuals in social networks. In *Proceedings of the eleventh international workshop on Web information and data management* (New York, NY, USA, 2009), WIDM '09, ACM, pp. 67–75.
- [16] NARAYANAN, A., AND SHMATIKOV, V. De-anonymizing social networks. *CoRR abs/0903.3276* (2009).
- [17] NARAYANAN, A., AND SHMATIKOV, V. Myths and fallacies of personally identifiable information. *Communications of the ACM* 53, 6 (2010), 24–26.
- [18] PERITO, D., CASTELLUCCIA, C., KAAFAR, M., AND MANILS, P. How unique and traceable are usernames? In *Privacy Enhancing Technologies, Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011, pp. 1–17.
- [19] S. LABITZKE, I. TARANU, H. H. 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* (San Diego, CA, USA, 2011), SNA KDD '11, ACM, pp. 51–60.
- [20] SZOMSZOR, M. N., CANTADOR, I., AND ALANI, H. Correlating user profiles from multiple folksonomies. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia* (New York, NY, USA, 2008), HT '08, ACM, pp. 33–42.
- [21] VOSECKY, J., HONG, D., AND SHEN, V. User identification across multiple social networks. In *Networked Digital Technologies, 2009. NDT '09. First International Conference on* (july 2009), pp. 360–365.

- [22] ZAFARANI, R., AND LIU, H. Connecting corresponding identities across communities. In *ICWSM* (2009).
- [23] ZHANG, C., SUN, J., ZHU, X., AND FANG, Y. Privacy and security for online social networks: challenges and opportunities. *Network, IEEE* 24, 4 (july-august 2010), 13–18.
- [24] ZHELEVA, E., AND GETOOR, L. 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* (New York, NY, USA, 2009), WWW '09, ACM, pp. 531–540.