2019 22nd International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2019

Automatic Individual Information Aggregation using Publicly Available Social Media Data

Sibendu Sarker
University of Manitoba
Winnipeg, MB, Canada
sibendu@cs.umanitoba.ca

Wayne Franz
University of Manitoba
Winnipeg, MB, Canada
umfranzw@cs.umanitoba.ca

Mir Tafseer Nayeem
Ahsanullah University of Science and Technology, Dhaka, Bangladesh
mir.nayeem@alumni.uleth.ca

***Abstract*—In this paper, we make use of freely available data on public social media sites in two ways. First, we develop a search application capable of aggregating information about an individual, using only their name as input. Second, we investigate the feasibility of mining public data and linking this information across multiple social media sites in an attempt to produce an information profile for an individual. The inspiration of our work is to allow a person to see firsthand how much information about them exists online, and how this information could potentially compromise their privacy whereas the objective of our research is to analyze the feasibility of building a profile by gathering and linking information of an individual across different social media sites. This is done in the hope of perhaps inspiring an individual pay additional heed to the privacy settings on social media accounts, and to be more vigilant about what information they choose to share online.**

I. Introduction

Social media platforms have enjoyed a great surge in popularity in recent years. According to a 2016 study by the USA-based Pew Research Center, approximately 68% of American adults use Facebook, 28% use Instagram, 21% use Twitter, and 25% use LinkedIn [1]. The user-bases for these platforms are even larger because they span multiple countries (and even continents). While there are many reasons that could be put forward for the popularity these sites enjoy, the strongest attraction arguably remains the basic ability they provide to connect with other people.

In a social networking site, most users are online friends with people with whom they either interact with on a daily basis, live nearby, are family, or went to school or work with. On sites like Facebook, posting such information allows the platform to automatically search for points of intersection between profiles and make suggestions about people whom users may know but have not yet befriended. These features are often surprisingly accurate, and are designed to enhance the user's social experience by increasing the number of people they can interact with.

From a privacy standpoint, it is interesting to note that the same information that may be used to link *different user profiles within a site* may also be used to link *the same user between sites*. This enables an attacker to build a profile about an individual that contains details gleaned from multiple social media sites. Using multiple sites is advantageous because the types of details people provide may differ in accordance with the purpose of the site. For example, people on Facebook may choose not to share their current workplace. However, if such people also have a LinkedIn account, it is likely they will share this information there, since purpose of LinkedIn is to connect job seekers with potential employers.

The feasibility of this type of attack is enhanced by the fact that an alarming proportion of social media users do not alter the privacy settings that their platforms enable by default, leaving their information accessible to strangers.

In this research, we make use of freely available data on public social media sites to build an attack model that can retrieve sophisticated information about an individual which might be threatening for his/er privacy. We also try to find out the challenges to building such an attack model and measure the feasibility of the model, considering the current security architecture of different social media platforms.

II. Related Work

Table I summarizes the different approaches to social media linkability-related assessments and attacks. Some of the works more relevant to our research are summarized below.

Goga *et al.* [7] demonstrated that it was possible to identify users of social media websites by linking together accounts belonging to a single user. Doing so only required considering three attributes to be mined from posts: a geographic location, a time-stamp, and the user's writing style (captured using language modeling techniques).

Along similar lines, Balduzzi *et al.* [5] constructed a social media profile crawler that was capable of using publicly available email addresses to link together information from different social media sites (automatically building a user profile). They found they were able to uniquely match 1.2 million email addresses to social media profiles.

Almishari and Tsudik [3] explored several techniques for linking together anonymous Yelp reviews. Given a sample of users who had submitted at least 40 anonymous reviews, they were able to accurately cluster 70% of the reviews into groups matching their authors. They speculated that this technique could be used to link anonymous reviews between Yelp and other sites providing reviews (which may contain additional user information).

Acquisti, *et al.* [2] used facial recognition software together with publicly available online data (from sites like Facebook)

978-1-7281-5842-6/19/$31.00

TABLE I: Summary of the social media linkability-related approaches

Paper	Type of Work	Features used	Datasets Used
Acquisti *et al.*, 2014 [2]	Account Linking / Identification	Photos (Facial Recognition), location	Facebook, Unnamed Dating Website
Almishari and Tsudik, 2012 [3]	Account Linking	Language Analysis, Rating, Category	Yelp Reviews
Backes *et al.*, 2016 [4]	Risk Analysis	Language Analysis	Reddit Comments
Balduzzi *et al.*, 2010 [5]	Account Linking	E-mail address, age, sex, location, job	Facebook, MySpace, Twitter, LinkedIn, Friendster, Badoo, Netlog, and XING Profiles, leaked email addresses
Becker *et al.*, 2012 [6]	Data Aggregation (for a social event)	Language Analysis,, Time	Last.fm, EventBrite, LinkedIn, Facebook event pages
Goga *et al.*, 2013 [7]	Account Linking	Language Analysis,, Geo-locations attached to posts, timestamps, tags	Yelp, Flickr, Twitter
Krishnamurthy and Wills, 2009 [8]	Risk Analysis	Cookies passed in HTTP headers	Bebo, Digg, Facebook, Friendster, Hi5, Imeem, LiveJournal, Twitter, LinkedIn etc.
Liu *et al.*, 2011 [9]	Event detection	Geo-tags, Venue, timestamp, media-sharing between users	EventMedia (Flickr, Last.fm, Eventful, and Upcoming combined dataset)

to identify people and gather information about them. They even developed a mobile application capable of doing this in real time (using only a photo provided by the mobile device).

One way in which our research differs from the above mentioned works is that we are seeking to use publicly available information that does not require any authentication to access. Many of the above works used web-based APIs for crawling the social network graphs of various sites and collecting information. However, currently these types of web APIs almost universally require authentication to be granted by the user before they can be used.

Many of the related works noted above do not mention this. It is possible that, as most of these works are dated prior to 2014, it may have still been possible to use Facebook's graph API (a particularly popular data source in the table above) to collect public data at the time. Facebook continues to improve its privacy-preserving measures, and, as far as we can tell, this is no longer possible. In fact some of the most recent works didn't attempt to consider this aspect [10], [11], [12], [13]. Several other works use sites that provide reviews, which are all publicly accessible, but are not as relevant when attempting to build a personal profile (one of our goals). It is also worth noting that one of the works (Liu *et al.* [9]) made use of the *EventMedia* dataset, which is a large anonymized collection of events pulled from social media sites. Again, this dataset is less relevant to our project, as it is event-focused rather than individual-focused.

Instead, we approach data collection problem using a technique known as "screen scraping." More details on this technique are provided in the Methodology section below.

III. Methodology

The first part of our research involved developing a web-based interface that collected data from Facebook. A screen-shot of the application in operation is also shown in figure 1. We were able to retrieve information about users such as schools they had attended, the year in which they had graduated, current and past workplaces, profile pictures, current and past cities they'd lived in, and more.

Secondly, we created a mechanism for discovering URLs of public Facebook and Twitter profiles, and extracting data from

Fig. 1: User interface

Facebook and Twitter (note: our reasoning for abandoning LinkedIn as a data source is discussed in the Data Collection section below). This information was then inserted into a local SQLite database [14]. Finally, we attempted to manually link this data using SQL queries, and examined the distribution of different attributes. These results are presented in the Results section of this paper.

A. Design Decisions

Facebook provides a number of APIs that web applications can make use of. Their *Graph API* [15] is of particular interest to us. In this API, nodes represent user profiles and edges are the relationships between friends. It is also possible to trace links between people who write on each other's walls or comment on each other's posts. Unfortunately, using this API requires an access token. Obtaining an access token for a particular profile generally requires that user to authenticate (login to Facebook through your application) as noted earlier. This makes the Graph API difficult to use for crawling public information. This is likely intentional on Facebook's part. This

API is intended to allow mobile phone developers to create native Facebook Apps. In such applications people generally only need to access their own data and that of their friends.

Luckily, Facebook also provides a page that allows one to search profiles by name, using request parameters in the URL. This allows us to discover profiles and determine whether or not they are public.

Data on Twitter are more easily accessible than other social media platforms. Most content users submit, post, or display is public by default so that it can be viewed by other users. It is worth noting that in their privacy policy [16], Twitter states that they have the right exchange information with certain third parties such as advertisers. Technically, our information gathering process still operates within the boundaries laid out in this document because our aim is to collect only those data which are publicly accessible. In spite of the policy, most Twitter users still choose to share a lot of sensitive information like their date of birth, current location, and profile pictures, which become public by default. There is an API available for accessing user information on Twitter. However, much like Facebook's, this API requires an access token and a secret key (granted to an application that has been approved by Twitter).

Unlike some other social media platforms, the Twitter search page is publicly available. Therefore we used it to locate users' pages by name, all of which are public.

Ideally, there are a number of factors that should be weighed in order to determine which types of information to collect from social media sites. One wants to link using attributes that would give the highest accuracy. Therefore, linking attributes should be chosen such that they act as a sort of quasi-identifier [17], [18]. As we learned from studying attacks on K-Anonymity [19] , in general, the larger the number of attributes in the composite quasi-identifier, the smaller the number of records each unique combination tends to map to.

However, in reality, not all attributes are universally available on social media sites, and those that are may be given at different levels of granularity. For example, a location attribute may be very helpful to use within a composite quasi-identifier if it is specified as a city, but may be less helpful to use if it is simply given as a country. In other words, it is not only combinations of attributes that matter, but also their availability and the granularity at which they are expressed. We discuss additional challenges we faced identifying attributes to use for linking in the (following) Data Collection & Account Linking sections.

In the end, we found that the attributes that seem to be most frequently available on public profiles include the username, schools the user has attended, cities and locations the user has lived in (past and present), current and past workplaces, the user's profile picture, and their "bio" (a short paragraph describing themselves). We therefore started by collecting this information.

B. Data Collection

In general, there are two different ways to go about collecting data from social media sites.

As noted earlier, one method (arguably the easiest) involves using the publicly available web APIs offered by the platforms. For example, Facebook offers a Graph API [15] in which nodes represent profiles and edges are the connections between friends. Twitter and LinkedIn have similar functionality.

Unfortunately, there is problem with this approach. Accessing web APIs requires user authentication. This means that there is a certain amount of tension between the web API data collection approach and our goals. We want to collect publicly available data in order to give the user of our system an idea of how much information is freely accessible on the Internet. If we obtained user authentication and made use of the access key, we would have access to not only public, but also private information, with no way to tell the difference. On Facebook, for instance, the level of privacy can be controlled - to some extent - by the user, meaning that what constitutes publicly available information may differ between profiles.

A second method for obtaining data from social media sites is known as *scraping*. In this approach, one writes a script that makes a request to the social media web server and collects the HTML response that is sent back [20], [21]. This response can then be parsed, and the relevant data can be extracted.

Extracting the data in this manner can be time-consuming, as it requires the programmer to write code tailored to the format of each social media site being scraped [22], [23]. Some sites, such as Facebook, contain page structures the user can customize (eg. we found that some users may have elected not to update from the "old" style profiles to the newer style ones), further complicating the parsing.

Finally, most modern social media websites make extensive use of Asynchronous JavaScript and XML (AJAX) Requests, which fetch additional data on demand - typically when the user performs some action on the page. For example, when one views a Twitter profile, only a small number of tweets are actually loaded in the initial HTTP GET request. Additional tweets are transferred from the server by a secondary request that is made only when the user attempts to scroll down below the currently viewable list.

In some cases, it is possible to replicate AJAX requests by examining the network activity that results when a human performs an interaction with the website. We were able to do this for the aforementioned Twitter tweet loading scenario. In other cases, some profile data is embedded in "meta" tags in the HTML which do not render. Instead, the meta tags appear to be used by the JavaScript running on the page, which can create HTML elements on the fly. We were able to pull some data attributes from the meta tags in responses from Facebook to avoid making unnecessary and time consuming AJAX requests.

We used the scraping approach to collect data from Facebook and Twitter. We also tried to collect information from LinkedIn, but found that it seems to block public profiles after about 5 views. We therefore abandoned LinkedIn as a data source.

We tried two different approaches when searching for user profiles by name. These are illustrated in the flowchart in figure

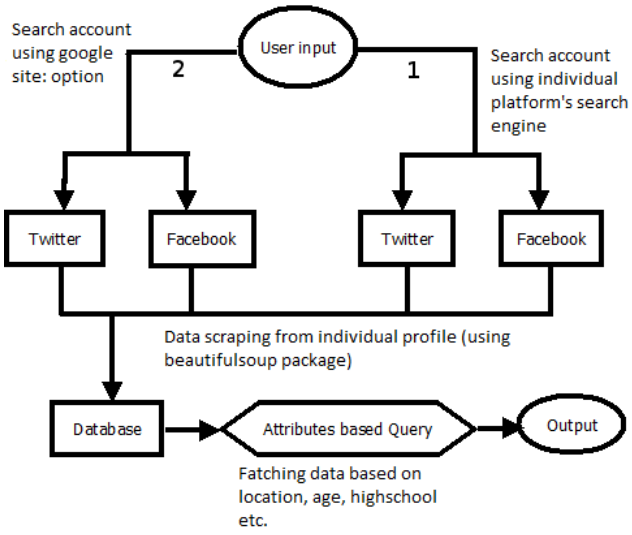

Fig. 2: Flowchart of our approach.

2. Each approach is represented as a numbered path emerging from the (initial) *user input* state at the top of the diagram.

Path 1 involves searching for accounts by name using the social media platform's search features. We found that both Facebook and Twitter seem to limit the number of results that are returned when searching in this manner. Facebook seems to return links to 100 profiles (not all of which may be public). Twitter returns many more, but seems to leave out some that are returned when searching manually via Google. We collected a small dataset (186 profiles; 78 from Facebook and 108 from Twitter) by executing a search for profiles with the input name "Wayne Franz" using this method. We will hereafter refer to this dataset as *dataset 1*.

Path 2 involves using Google's "site search" feature to search for profiles with the given name. In this approach, our code first executes a Google search that operates only on the social media network's domain, then extracts the links from the first 100 pages of results. We then filter the links down to those that look like profiles (rather than posts with references to the name) and follow them. The downside to this approach is that Google may block IP addresses that make too many requests in too short a time period. In our experience, we were able to run though all 100 pages of results before getting blocked. Any more than this, however, resulted in a ban lasting several hours in duration. In spite of this, we were able to collect a slightly larger dataset using this method (269 profiles; 199 from Facebook and 70 from Twitter). As before, we searched for profiles with the input name "Wayne Franz." We will refer to this dataset as *dataset 2*.

C. Account Linking

In both of the approaches mentioned above, once we obtained links to the social media profiles, we followed them and attempted to extract information. On Facebook, we were interested in names, locations, workplaces, schools, and the text in the person's "biography" section. We also extracted additional information (links related to their workplaces, blogs, etc.), but found that these varied too much between profiles to be of much use (they are not consistently labelled and are relatively sparse). On Twitter, we were able to pull a profile's name, location, the text in their "biography" section, and the top 3 tweets in the profile's feed. Finally, on both Facebook and Twitter, we collected the URLs of the profile pictures to use in our web-based interface.

It should be emphasized that the attributes mentioned above were the ones we *looked for* in the HTML response - the number of these that were actually present varied from profile to profile. Table III displays the counts of the number of profiles containing each attribute for dataset 2.

The sparsity of the attributes makes it difficult to link the profiles using simple SQL queries. There were also a number of other complicating factors.

We realized that many companies and organizations maintain a social media presence. This means that searching for the name "Wayne Franz" turns up profiles for organizations like the "Fort Wayne Police Department" in Indiana. Similarly, there are many accounts dedicated to particular bands, athletes, and politicians with the name "Wayne."

While these results are interesting, they are seldom related to the particular *user* who has entered their name into our application. Unfortunately, manually (without a machine learning model), it is difficult for us to tell the different types of profiles apart. Thus, our search wastes a lot of time and expends a number of useless queries following the links to these accounts.

In order to emphasize the extent of this problem, we have provided a breakdown types of accounts present in dataset 2 (these were manually classified) in Table IV, also expressed as a pie chart in Figure 3. In the table, "Interest Groups" refers to groups dedicated to a particular activity that does not have a business interest (eg. a chess club), "Businesses / Orgs." refers to commercial, government, and non-profit organizations (eg. Restaurants, Police Departments, self-employed Artists, Charities, etc.), "People" refers to the types of profiles we are interested in (regular profiles for individuals with no business interests), and "Public Figures" refers to people like politicians and athletes who are using the social media platform as a means to disseminate or collect information. The values in the "Total" row add up to 268 (one less than reported in Table III) because there was a single profile that we were unable to manually classify due to its removal from social media since our crawler collected data on it. Note that a full 38% of the profiles appear in the "Businesses / Orgs." section of figure 3, while only about 39% represent "regular" people with no apparent ulterior motives for creating a profile.

Another difficulty with linking arises from the fact that there is no standard way to label the data. On Facebook, for example, users may enter free-form text into the "location" box when they create their profiles, leading to (entertaining but) unhelpful colloquialisms such as "Farm Fresh New York" in place of "New York City". In addition, abbreviations can cause problems (eg. "NYC") without some form of natural language processing (or an exhaustive list that can be searched) to recognize them. Finally, as noted earlier, the granularity at

which some attributes are expressed can differ significantly between accounts. For example, while some people may list the city they reside in, others may only list the country, making it difficult to match these values.

IV. Experimental Results

Using dataset 1, we were able to match 3 accounts (out of a total of 186) by linking on name alone. We then manually verified that these accounts referred to the same individual. The results of our join operation are shown in Table II. In the column headings, Facebook is abbreviated "FB" and Twitter "TW". Facebook allows users to specify multiple locations, each of which may be given one of several predefined labels (such as "Home Town" or "Current City"). Similarly, our database stores multiple locations for each profile. The ID columns show which rows correspond to the same profile.

Table II illustrates many of the problems mentioned above regarding linking based on location. First, note that while the first individual (FB ID = 25) has specified their full location (in the form "City, Province") on both Facebook and Twitter, on Twitter their province name has been abbreviated.

The second individual (FB ID = 34) has given two locations of Facebook, but none on Twitter. Also notice that the Facebook location labels differ slightly from those the previous individual used, specifying "Hometown" instead of "Home Town". This appears to be the result of changes to Facebook's labels over time, which further complicates linking.

Finally, notice that the third individual (FB ID = 36) has listed three cities on Facebook in three different formats ("City, Country", "City, Province" and "City, Province, Country", respectively).

TABLE II: Results of join on name (Dataset 1)

FB ID	FB Location	FB Label	TW ID	TW Location
25	Saint John, New Brunswick	Current city	90	Saint John, N. B.
25	Saint John, New Brunswick	Home Town	90	Saint John, N. B.
34	Los Angeles, California	Current city	98	(unspecified)
34	California City, California	Hometown	98	(unspecified)
36	Pretoria, South Africa	Current city	159	Pretoria, South Africa
36	Centurion, Gauteng	Hometown	159	Pretoria, South Africa
36	Benoni, Gauteng, South Africa	Moved here	159	Pretoria, South Africa

We were unable to match any rows based on name in dataset 2. This is likely due to the high percentage of companies and organizations in this dataset, as mentioned earlier. In addition, it is likely that our datasets are simply too small to expect many matches. If we had been able to somehow circumvent the problem of getting blocked from sites like Google, we expect that we would be able to perform additional matching.

TABLE III: Counts of profiles containing various attributes (Dataset 2)

Platform	Total Profiles	Location	Schools	Workplaces	Bio
Facebook	199	34	28	24	16
Twitter	70	70	0	0	70
Total	269	104	28	24	86

Some of the related works were able to solve the location linking problems mentioned above by relying on "Geo-tags."

TABLE IV: Counts of profile types (Dataset 2)

Platform	Bands	Interest Groups	Businesses / Orgs.	People	Public Figures
Facebook	30	16	93	52	8
Twitter	4	1	9	53	2
Total	34	17	102	105	10

These are coordinates specified in latitude and longitude that users can attach to posts. In our experience, the number of posts containing these tags was limited. In addition, the fact that someone has tagged themselves in a location does not necessarily mean that they live in the vicinity, as there are a plethora of vacation-related posts on social media. Still, there may be some value in taking this approach, and if we had had more time, we would have attempted to explore it further.

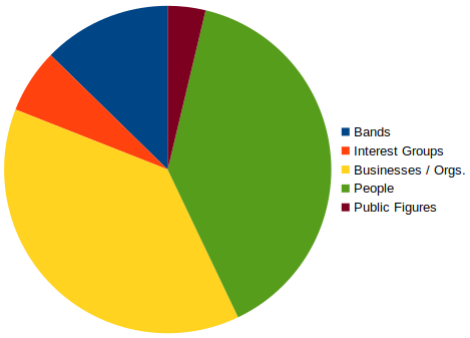

Fig. 3: Breakdown of profile types (Dataset 2)

Workplaces and schools, though plentiful on Facebook, were more difficult to collect from Twitter (this can be seen in Table III), since it has no dedicated profile entries for these attributes. Generally, if they are present on a Twitter profile, they are indirectly specified either in the "biography" section or the tweets themselves. Both of these sections of the profile contain free-form text, which means the best way to extract schools and workplaces is probably to use some form of natural language processing package. These can be quite complicated, and unfortunately we did not end up having the time to incorporate them into our research.

V. Discussion & Drawbacks

While we were ultimately unable to collect a large enough dataset to accomplish our goal, we would to take the opportunity to communicate a number of things we investigated over the course of this research.

Goga *et al.* [7] emphasize that username is the most valuable attribute that can be used for linking. They were able to develop a system that could predict whether Flickr and Twitter accounts belonged to the same user with a true positive rate of approximately 77%. Users, it seems, do not often go to the trouble of thinking up different usernames for different accounts (and as a side note, this suggests that the same is likely true of their passwords). We were not able to procure a large enough dataset to be able to match effectively on username alone. However, we found that we were able

to obtain profile pictures for almost all users in both of our datasets.

We believe that the inconsistency of the labels on the data presented in profiles (such as the location) makes natural language processing a necessity in any tool that is to attempt to link profiles accurately. *Linking information, it turns out, is not as hard a problem as locating and classifying it.*

One of the best ways to see this is to consider the "About" section on a Facebook profile. We were able to extract much more information from this section than we could insert into our database. These data included things like links to personal web pages, email addresses, phone numbers, and more. This information is often specified in a free-form manner, and, while it might be feasible to pull some of it out using simple regular expressions (eg. phone numbers in particular), it is impossible to gauge its relevance without knowing something about the semantic context it appears in (eg. distinguishing a work phone number from a personal one, or a link to a personal web page from a link to a company website). In other words, we were able to obtain a great deal of *information* that we could not convert into *data*.

The uniqueness of a particular piece of information among other public profiles is also worth considering. This is reminiscent of the concept of quasi-identifiers in K-Anonymity. If many users list their hometown, then hometown actually becomes a less useful attribute to link on (because there will tend to many matches for the same value). However if few users provide their hometown, or there is some combination of hometown and other attributes that is sufficiently unique, it may be possible to link with greater accuracy. Based on our limited experiments on our datasets and our literature survey, it seems reasonable to expect that features used for linking might need to be weighted depending upon their uniqueness.

In some cases there may not be enough unique information to narrow down the linkage between profiles. In these cases, we had planned to request additional information from the user (eg. perhaps they know the province that the person they are looking for resides in). However, our limited ability to classify location information hampered this. Currently, humans are still some of the best classifiers around so it stands to reason that an interactive "incremental" search process would likely enhance both the accuracy and performance (by reducing the size of the search space) of the linking process.

VI. Conclusion

This research has investigated the feasibility of aggregating publicly available information from social media sites to build an information profile for a given individual. We were able to achieve our first goal: building a web-based application that pulls data from Facebook and Twitter. In the end, we ran into some road blocks procuring the dataset that was necessary to link the gathered data. However, we were able to achieve a slightly scaled down version of our second goal: the automated linking of profile data.

References

[1] Greenwood, S. and Perrin, A. and Duggan, M., "Social Media Update 2016," Nov. 2016.

[2] Acquisti, A. and Gross, R. and Stutzman, F., "Face Recognition and Privacy in the Age of Augmented Reality," *Journal of Privacy and Confidentiality*, vol. 6, no. 2, 2014.

[3] Almishari, Mishari and Tsudik, Gene, *Exploring Linkability of User Reviews*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[4] Backes, Michael and Berrang, Pascal and Goga, Oana and Gummadi, Krishna P. and Manoharan, Praveen, "On Profile Linkability Despite Anonymity in Social Media Systems," in *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, ser. WPES '16. ACM, 2016, pp. 25–35.

[5] Balduzzi, Marco and Platzer, Christian and Holz, Thorsten and Kirda, Engin and Balzarotti, Davide and Kruegel, Christopher, "Abusing Social Networks for Automated User Profiling," in *Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection*, ser. RAID'10, 2010, pp. 422–441.

[6] Becker, Hila and Iter, Dan and Naaman, Mor and Gravano, Luis, "Identifying Content for Planned Events Across Social Media Sites," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12. New York, NY, USA: ACM, 2012.

[7] Goga, Oana and Lei, Howard and Parthasarathi, Sree Hari Krishnan and Friedland, Gerald and Sommer, Robin and Teixeira, Renata, "Exploiting Innocuous Activity for Correlating Users Across Sites," ser. WWW '13. New York, NY, USA: ACM, 2013, pp. 447–458.

[8] Krishnamurthy, Balachander and Wills, Craig E., "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.

[9] Liu, Xueliang and Troncy, Raphaël and Huet, Benoit, "Using Social Media to Identify Events," in *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media*, ser. WSM '11. New York, NY, USA: ACM, 2011, pp. 3–8.

[10] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explorations Newsletter*, vol. 18, no. 2, pp. 5–17, 2017.

[11] R. Zafarani and H. Liu, "User identification across social media," Jan. 10 2017, uS Patent 9,544,381.

[12] Y. Li, Y. Peng, Z. Zhang, Q. Xu, and H. Yin, "Understanding the user display names across social networks," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017.

[13] P. Jain, P. Kumaraguru, and A. Joshi, "Other times, other values: leveraging attribute history to link user profiles across online social networks," *Social Network Analysis and Mining*, p. 85, 2016.

[14] SQLite, Available: https://www.sqlite.org/.

[15] Facebook, "Graph API," Available: https://developers.facebook.com/docs/graph-api/overview/.

[16] Twitter, "Twitter Privacy Policy," Available: https://twitter.com/privacy.

[17] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[18] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005.

[19] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 106–115.

[20] M. T. Nayeem, T. A. Fuad, and Y. Chali, "Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1191–1204.

[21] M. T. Nayeem and Y. Chali, "Extract with order for coherent multi-document summarization," *TextGraphs-11*, p. 51, 2017.

[22] M. T. Nayeem, T. A. Fuad, and Y. Chali, "Neural diverse abstractive sentence compression generation," in *European Conference on Information Retrieval*. Springer, 2019, pp. 109–116.

[23] M. T. Nayeem and Y. Chali, "Paraphrastic fusion for abstractive multi-sentence compression generation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017.