Methods for User Profiling Across Social Networks

Rishabh Kaushal[1,2], Vasundhara Ghose[1], and Ponnurangam Kumaraguru[2]

[1]Department of Information Technology, Indira Gandhi Delhi Technical University for Women, Delhi, India
[2]Precog Research Group, Indraprashtha Institute of Information Technology, Delhi, India

***Abstract*—Users have their accounts on multiple Online Social Networks (OSNs) to access a variety of content and connect to their friends. Consequently, user behaviors get distributed across many OSNs. Collection of comprehensive user information referred to as user profiling; an essential first step is to link user accounts (identities) belonging to the same individual across OSNs. To this end, we provide a detailed methodology of five methods useful for user profiling, which we refer to as Advanced Search Operator (ASO), Social Aggregator (SA), Cross-Platform Sharing (CPS), Self-Disclosure (SD) and Friend Finding Feature (FFF). Taken together, we collect linked identities of 208,120 individuals distributed across 43 different OSNs. We compare these methods quantitatively based on social network coverage and the number of linked identities obtained per-individual. And also perform a qualitative assessment of linked user data, thus obtained by these methods, on the criteria of completeness, validity, consistency, accuracy, and timeliness.**

***Index Terms*—User Profiling, Social Media Analysis, Online Social Networks.**

I. Introduction

The popularity of Online Social Networks (OSNs) is increasing by the day with more and more people joining multiple OSNs to share information about themselves, connect to other users, and receive updates from them. Each OSN offers a unique service or ecosystem which attracts users to join more than just one of them. Facebook for personal friends, LinkedIn for the professional network, YouTube for viewing & sharing videos, and Twitter to get quick updates are the best options. The average number of social media accounts per online user has risen from 4.3 to 7.6 during 2013 to 2017[1]. To collect user information in a comprehensive manner, an essential *first step* is to gather user accounts (identities) of the same individual across multiple OSNs, which we refer to as *linked identities*. And the systematic approach to performing a large scale collection of user behaviors across OSNs is referred to as *user profiling* [1].

User profiling has many advantages and applications. Users tend to provide incomplete information on a single social network, either with purpose or otherwise. Knowing the same user's identity on other social networks would help in the comprehensive profiling of the user in terms of user's profile, user's content, user's behavior, user's preferences, and user's friends. In the advertising world, it enables targetted advertisement [2] and improved recommendations. Researchers have studied most of the problems in the domain of social networks like information propagation, link prediction, algorithmic biases, discrimination studies, and community detection in the realm of a single social network, which we can study across multiple social networks. In social media crimes and cybersecurity problems like cyberbullying, fake accounts, and spamming, we are often looking for user footprints within the same social network in which incident occurred. If the user's identities on other social networks are known, then it is only going to help in the investigation [3]. From a user's privacy standpoint, individuals can be shown their comprehensive profiles and likelihood of linkage of their identities and nudged to control their online behavior so that their digital footprint decrease [4]. Lastly, there is no agreed benchmark dataset in the problem domain of identity resolution. So, large scale data collection of linked identities would help researchers compare and evaluate their proposed solutions.

Given the significance of user profiling, a lot of emphases has been given in the research community to solve the first step in user profiling which involves linking user identities belonging to the same person, referred to as *identity resolution* (or *identity linkage*).[2] A data-driven approach to solve the identity resolution problem has two key steps. Firstly, we collect a large number of user identity pairs belonging to linked identities and non-linked identities. Secondly, we construct a machine learning-based model over the user behavioral features extracted from user identities. In this paper, we focus our attention on the first step, which involves the collection of linked identities across OSNs. Figure 1 depicts the before and after stages involved in linked identity collection for user profiling. In this paper, we explain five methods to obtain linked identities namely Advanced Search Operator (ASO), Social Aggregator (SA), Cross-Platform Sharing (CPS), Self-Disclosure (SD) and Friend Finding Feature (FFF). Taking all these methods together, we collect linked identities of 208,120 individuals across 43 different OSNs, which is by far the most comprehensive coverage, towards user profiling, refer at http://precog.iiitd.edu.in/resources.html for dataset details. Subsequently, we present a detailed quantitative and qualitative assessment of these methods. For quantitative assessment, we evaluate the number of social networks covered by a method and number of linked identities obtained per-individual across OSNs. For qualitative assessment, we leverage standard

[1]https://www.statista.com/statistics/788084/number-of-social-media-accounts/

[2]This problem is known in literature by multiple names such as Social Identity linkage [5], User identity linkage [6], user Identity Resolution Social Network Reconciliation [7] , User Account Linkage Inference [8], Profile Linkage [9], Anchor Link prediction [10] and Detecting me edges [11].

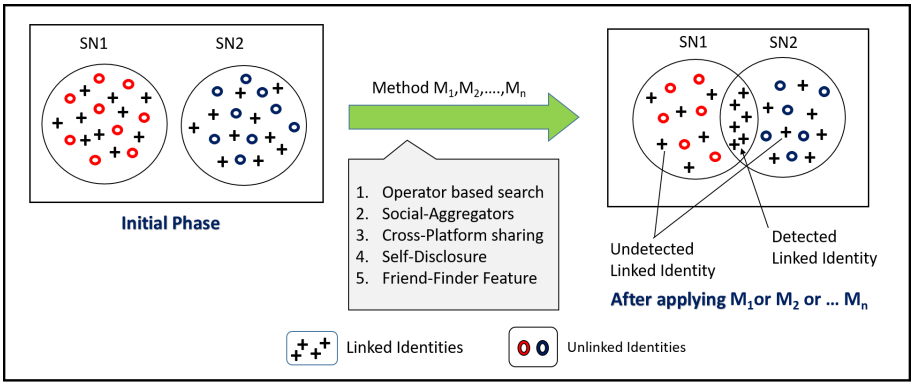

Fig. 1. Visual depiction of progressive stages in which linked identities are collected starting from no linked identities and gradually progressing to collect as many of them as possible by applying methods for user profiling.

parameters from ISO 9000:2015[3] namely data completeness, consistency, accuracy, validity, availability and timeliness.

Collecting user data from online social networks have always been a challenge and given that the data is related to users, there are privacy issues as well. Application Programmer's Interfaces (APIs) offered by OSNs have been dwindled their capabilities over the years owing to data privacy concerns. The recent data breach[4] involving Facebook and Cambridge Analytica would have an adverse implication on data collection by academics for research purposes.[5] With all this happening, users are becoming even more privacy aware which would dissuade them from mentioning all the details in their accounts, resulting in *missing values* when data is collected. To make things worse, there are social network platforms like Twitter which allow users to change their account handles, thereby, complicating the data collection process.

Regardless, this is the first work to the best of our knowledge which focuses exclusively upon the methods for collecting linked identities, which is the the de-facto first step for user profiling. Key contributions of our work are:-

- Detailed description of data collection methods to retrieve linked identities, thereby facilitating user profiling.
- Comprehensive evaluation of data collection methods both qualitatively and quantitatively.
- Creation of a comprehensive dataset that can be used as benchmark dataset for identity resolution research.

II. RELATED WORK

One of the earliest works is that of Perito et al. [12], who conducted various data collection methods to obtain the usernames belonging to the same individual across many sites. Prominent sites used were Google profiles and eBay to collect 3.5 million usernames and 6.5 million usernames, respectively. For ground truth, they relied on Google profile users who have listed their usernames on other sites. Malhotra et al. [13] used Google API to retrieve a list of user accounts declared by users on their Google profile. They collected ground truth from Twitter, YouTube, and Flickr; however, there were many missing fields. Twitter and LinkedIn were other platforms to obtain 29,129 pairs of user accounts. Zafarani et al. [14], in their work, collected usernames of the same real world user across 32 different sites. Three primary sources to obtain these matching usernames were namely social network sites like Facebook or Google+ where online users mention their usernames on other websites, blogs, or blog advertisement portals. Oana Goga et al. [15] collected identical usernames on three of the social network being considered Twitter, Yelp and Flickr using 'Friend Finder' mechanism present on these sites. Further, in their work [16], they focused on Twitter, Facebook, Google+, Flickr, and Myspace. User features from these social networks were obtained using their APIs. However, for ground truth, 3 million Google+ profiles were randomly crawled to exploit the fact that user list down their social network accounts on Google+ profile pages. Iofciu et al. [17] used Social Graph API to crawl 421,188 public profiles of users while considering Flickr and Delicious and StumbleUpon (FDS dataset). Kong et al. [10] considered Twitter and Foursquare as the two real-world networks for their investigation. For ground truth, they used users who had mentioned in Twitter ID's in their Foursquare account pages. Xin Mu et al. [6] considered Chinese social networks Weibo, Renren, 36.cn, and Zhaopin for studying identity resolution. For ground truth, they annotated 2,186 pairs of user accounts across these social network pairs. Man et al. [18] took their first dataset from crawling of Facebook users, after deleting the users who have less than five friends, there remains 40,710 users with 766, 519 connections. The second data set was co-author network formed from papers published in conferences in the domain of Data Mining and Artificial Intelligence. Peled et al. [19] collected data using web crawling from two social networks, Xing and Facebook. They manually obtained seed profiles for crawling pairs of user profiles, one from each network, that belong to the same individual. A tool was

[3]International Standards Organization: https://www.iso.org/standard/45481.html

[4]https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

[5]https://scroll.in/article/872770/cambridge-analytica-scandal-could-hurt-legitimate-researchers-using-facebook-data

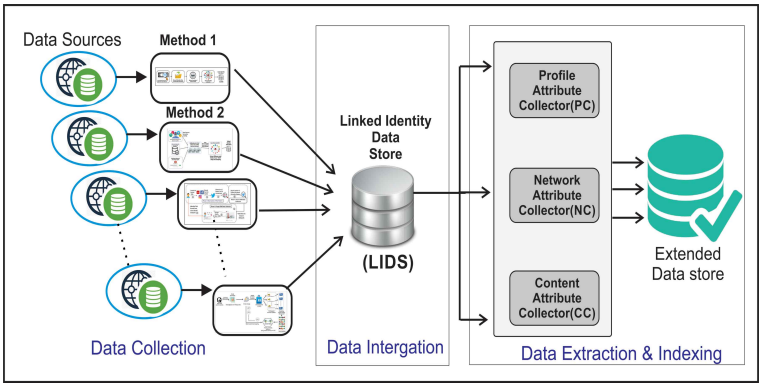

Fig. 2. Generic Framework for User Profiling.

developed to extract user data like gender, name, education, professional experience, friend list, etc. Labitzke et al. [20] collected 110,000 Facebook profiles, more than 43,000 profiles of StudiVZ, more than 25,000 MySpace profiles and more than 10,000 profiles of Xing. Liu et al. [21] collected the dataset by performing a survey based on 153 respondents and an analysis of 75,472 users on About.me website. For the collection of ground truth, they hired one human annotator who considered user profile content pages with more details related to users and their posts to mark true positive identities. Riederer et al. [22] collected datasets from location-based data extraction from OSN. Most interestingly, they included Call Data Records (CDRs) as well for making use of cell phone location tracking. OSNs considered were Foursquare-Twitter-Instagram cellphone-credit card records. They collected ground truth from the publicly available dataset and cell-bank. Zhang et al. [9] used a synthetic network based on Renren, Facebook, Sima, WeChat, and Twitter. They collected ground truth by using page crawling and open API wherever possible as per the platform is chosen.

Our work is different from prior works in the sense that we focus on methods to collect linked identities to enable the user profiling and present comparative assessment of these methods on qualitative and quantitative parameters.

III. Methodology

A generic framework for user profiling (Figure 2) comprises three steps, namely data collection, data integration, and data extraction & indexing. The first step is data collection, in which we identify a source of data followed by a selection of data collection methods. We follow it by data integration in which we store user identities collected from all methods at a single data store point, which we refer to as Linked Identity Data Store (LIDS). Finally, data extraction and indexing involve collecting the three components of user identity namely, profile, content, and network. Next, we describe each data collection method in detail. Next, we present a detailed methodology adopted to perform data collection using five methods, which is the focus of this paper.

A. Advanced Search Operator (ASO)

Search engines typically provide advanced search operators using which users use to obtain more detailed and specific information. In this work, we leverage Google's advanced operator search, also referred to as *google hacking* or *google dorking* (Figure 3). For instance, the search query *intext:facebook.com,twitter.com filetype:xlsx* would locate all web documents that have `facebook.com` or `twitter.com` written as text anywhere in the record with the additional constraint that these documents must be of *xlsx* file type. As per Figure 3, we first run a script which searches using pre-configured search queries on a specific search engine. Downloaded files are filtered and subsequently read through automated scripts which store linked identities in LIDS.

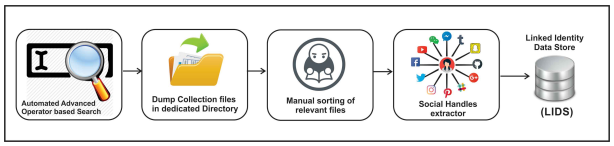

Fig. 3. Pipeline for Advanced Search Operator (ASO) method.

B. Social Aggregator (SA)

There are several social aggregating websites on which users create an account and provide details of their multiple OSN accounts. One such site that we investigate is *about.me*[6] which is a website that offers its users with a platform to mention numerous user identities, external websites, and well-known social networking websites such as Facebook, Flickr, Google+, Pinterest, LinkedIn, Twitter, Tumblr, and YouTube. Users put their one-page descriptions giving details of their social media profiles along with their background image and abbreviated biography. Initially, when we started data collection using this method, *about.me* provided an option to search user profile using the topic-based search (referred to as *discovery feature*). Given an interest-topic as input, it would return all the user profiles having that interest. After one month of data

[6]About.me: https://about.me/

collection, in March 2018, we found that this discovery feature of *about.me* got discontinued. Subsequently, on exploring other options, we found a public dataset[7] containing *about.me* profiles which we use in this work.

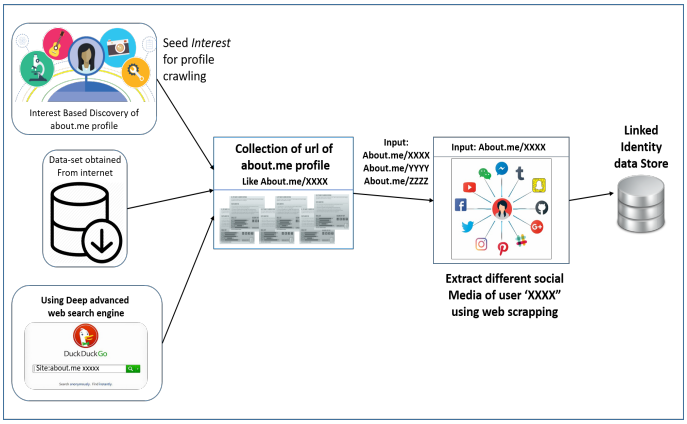

Fig. 4. Pipeline for Social Aggregator (SA) Method.

Besides, we leverage the previous ASO method using interests as *intext* and *site* as *about.me* to obtain more user profiles. Figure 4 explains the three data sources employed in this method.

C. Cross-Platform Sharing (CPS)

Many OSNs provide an option to share content across other (target) OSNs which we refer as *cross-platform sharing* (CPS). As depicted in Figure 5, a user makes a post on the source network (say Zomato, Facebook or Instagram) and then subsequently shares the same post on the target network (Twitter). Such shared content on the target network appears with a specific pattern. In our work, when we took Twitter as the target network and Instagram as the source network, then the pattern that appears on the shared post is $\backslash instagram.com\backslash p\backslash$. Using the API provided by the source network, we search for posts that contain such patterns. Besides this pattern, we also specifically check for the *source* field present in the Tweet JSON object and make sure that it has the name of the source network (in our case, Instagram). This check ensures that we filter out those scenarios in which a user might have copy-pasted the URL pattern because of such situations are not guaranteed to link to the same individual across the two networks. We parse the collected posts from the target network, identify the URL and expand the URL to reach the desired content on the source social network. On reaching the source social network, we either use source social network API or scrap the post page to obtain the tagged user (mentioned user) in the post on the source social network. In this way, we obtain linked identity pair between source and target social network.

[7]http://scholarbank.nus.edu.sg/bitstream/10635/137403/2/about_me.sql

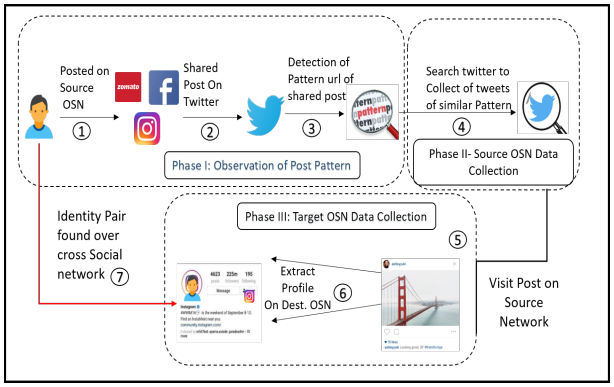

Fig. 5. Pipeline for Cross Platform Sharing (CPS) method.

D. Self-Disclosure (SD)

Whenever a user signs up on OSN, there is an option to provide a user description. At times, users provide details of their identities on other OSNs, which we refer to as *self-disclosure*. More specifically, we focus on the user's *bio* field in the Twitter network (Figure 6). We first use *Twiangulate* web tool[8] to collect all those twitter profiles which have at least one social network mentioned in their bio-field. Then, we observe various patterns in the bio-field on Twitter because a user can specify other OSN details in multiple ways. For instance, a user can mention *TV Host and Media Trainer - Instagram: @NeshanTVxyz Snapchat: @Neshaxyz* while another user can use acronyms like *TV Host and Media Trainer - IG: @NeshanTVxyz SP: @Neshaxyz FB: nashbin123*. To address these variations, we tokenize all text and check for the occurrence of URL which could lead to other OSNs.

E. Friend Finder Feature (FFF)

Whenever a user joins a new OSN, we sign up using our unique identifier, say email or phone number. This information is used by OSN to find our friends in our email contacts or phone contacts. Using this information, OSN offers a *friend finder* option to help connect to those friends who already have an account in OSN. Figure 7 depicts the entire sequence of steps that we followed in this method. In the first step, we use a deep web search engine like Duckduckgo [9] for retrieving emails present over the web. Next, we create an email account and add the extracted in its contact list. Then we sign-up in a social network to exploit friend finder feature using the created account. We use string matching on display name of users to find identity belonging to the same user.

IV. Results and Evaluation

In this section, we compare five methods, as stated before, by performing a quantitative and qualitative assessment of the linked identities obtained by them. Table I depicts total identities collected by prior works along with the OSNs covered by them. Few of them have a higher number of

[8]Twiangulate: http://twiangulate.com/search/
[9]Deep Web: www.duckduckgo.com

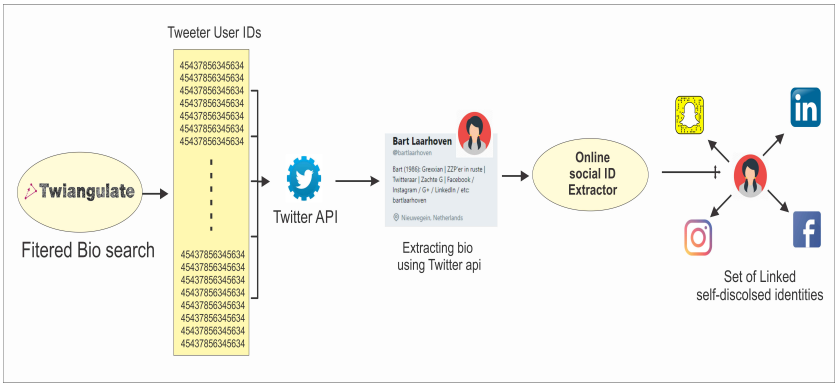

Fig. 6. Pipeline for Self-Disclosure (SD) Method.

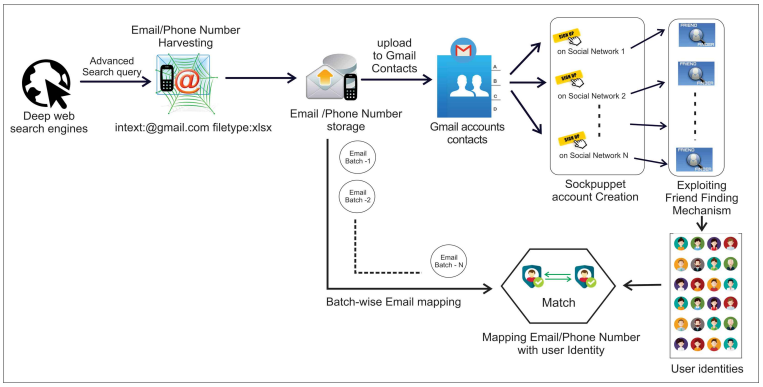

Fig. 7. Pipeline for Friend-Finder Feature method.

TABLE I
IDENTITIES COLLECTED IN PREVIOUS WORKS.

OSNs Covered with Reference	Identity collected
Twitter, Foursquare and yelp [15]	17,276
Twitter, Flicker, Facebook, Google+, Myspace, Yelp [16]	655,079
Social Graph API [17]	421,188
Facebook Twitter [10]	500
StudiVZ, Facebook, Myspace and Xing [7]	89,000
Facebook and Myspace [23]	5,296
Twitter, Flicker and Linkedin and 12 more.. [21]	75,472
Weibo, Renren, 36.cn and Zhoopin [6]	25,647
Facebook and Xing [24]	158
Twitter, Flicker [19]	27,000
32 Social Network sites [14]	100,179
Foursquare, Twitter and Instagram [22]	2,579
Twitter, YouTube and Flicker [13]	41,336
Twitter and BlogCatlog [25]	3,000
Facebook and MAG [18]	1,154

TABLE II
RESULTS OF DATA COLLECTION METHODS IMPLEMENTED IN THIS WORK. DATA COLLECTION IN EACH OF THEM IS CONTINUING AND NUMBERS ARE INCREASING BY THE DAY.

Data Collection Method	Linked Identities Collected So Far
Advanced Search Operator (ASO)	9,695
Social Aggregator (SA)	53,692
Cross-Platform Sharing (CPS)	104,233
Self Disclosure (SD)	40,000
Friend Finder Feature (FFF)	500
Total Linked Identities	208,120

identities; however, their coverage in terms of the number of OSNs reached is less than our dataset. Additionally, datasets of prior works as mentioned in Table I are not publicly available while we release our dataset, more details at http://precog.iiitd.edu.in/resources.html . Table II summarizes the number of linked identities collected using each of the data collection methods. Among all the five methods, Cross-Platform Sharing (CPS) method yielded the maximum number of linked identities (104,233) keeping Twitter as the target network and Zomato, Facebook, and Instagram being the source network from where the post was shared on to Twitter. Social Aggregator (SA) method using *about.me* gave 53,692 linked identities taking into account all three approaches followed in it, namely discovery feature, which contributed 15,973, a standard dataset that added 15,620 and search engine based which yielded 22,099. Self Disclosure (SD) method, which extracted identities by parsing *bio* field of Twitter gave 40,000 linked identities. We collected 9,695 identities using Advanced Search Operator (ASO) queries on *google*. Lastly, Friend-Finder Feature (FFF) gave 500 linked identities.

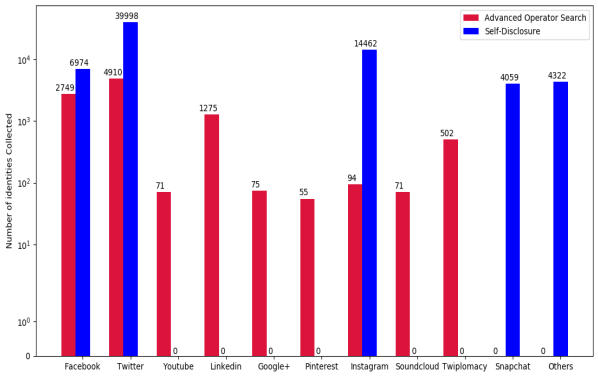

Fig. 8. Distribution of coverage of OSNs on which linked identities got collected using Advanced Search Operator (ASO) and Self Disclosure (SD) methods. Values on Y-axis are on log-scale to the base 10.

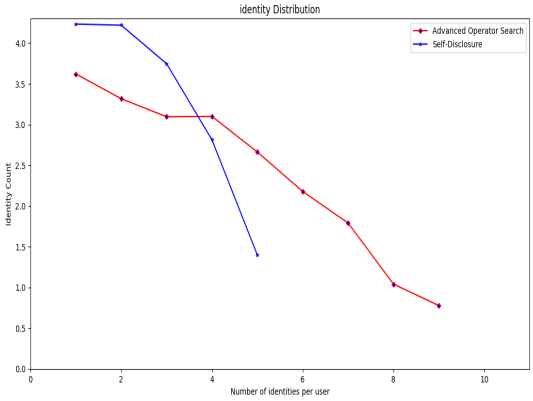

Fig. 9. Distribution of per-user identity count using all methods except social aggregator (SA) method.

A. Quantitative Evaluation

For quantitative evaluation, we evaluate data collection methods based on two metrics explained below.

1) Social Network Coverage: The data collection method is intended to collect linked identities across as many OSNs as possible. Social network coverage refers to the number of OSNs on which the given data collection method was able to collect linked identities. As depicted in Figure 8 Advanced Search Operator (ASO) method covers nine social networks Facebook, Twitter, Youtube, Linkedin, Google+, Pinterest, Instagram, Soundcloud, and Twiplomacy, during coverage of Self Disclosure (SD) method is across four social networks Twitter, Facebook, Instagram, and Snapchat (others comprises of LinkedIn and their blog/ websites). Further, in terms of OSNs coverage, Social Aggregator (SA) method performs the best. As depicted in Figure 10, a total of 43 OSNs got covered using this method. Among the three approaches employed in the SA method, the one that leverages search engine (*duckduckgo*) is giving the best results.

2) Per-user linked identity count: Number of linked identities found for a given user is referred to as per-user linked identity count. Figure 9 depicts the number of linked identities per user obtained using Advanced Search Operator (ASO) and Self Disclosure (SD) method. For per-user linked identity count less than 4, SD performs better, but subsequently ASO performs well. Also as expected, with the increase in per-user linked identity count, the number of such users, decrease. Fig 11 shows the per-user identity count distribution for Social Aggregator (SA) method. Discover feature and public dataset approaches give better results during the ongoing approach of the search engine is providing comparable results with discovery feature when per-user identity count increases beyond 5.

3) Results of Cross-Platform Sharing (CPS) Method: Quantitatively, Cross Platform Sharing (CPS) method is giving the best results, in terms of number of linked identities obtained. In the CPS method, we have used Twitter as the target network on which posts from other source networks namely Zomato, Facebook and Instagram have been shared, Table III gives the distribution of the same. Cross platforms sharing from Instagram to Twitter got the most of the linked identities.

TABLE III
RESULTS OF CROSS PLATFORM SHARING (CPS) METHOD IN WHICH WE DEPICT DISTRIBUTION OF CROSS PLATFORM SHARED POSTS FROM THREE SOURCE NETWORKS NAMELY ZOMATO, FACEBOOK AND INSTAGRAM ON TWITTER.

Source Network	Linked Identities
Zomato	6,000
Facebook	40,201
Instagram	58,032
Total Linked Identities	104,233

B. Qualitative Evaluation

We leverage metrics from ISO 9000:2015[10] Standard for quality assessment, namely data completeness, consistency, accuracy, validity, availability, and timeliness.

- **Completeness**: Completeness, in our context, can be defined as the ratio of collected linked identities of a user to the actual linked identities across all OSNs for the same user. From information retrieval perspective, this is similar to recall. Ideally, the methods should contain all linked identities but in practice, it is not possible, refer Table IV for explanations.
- **Validity**: Validity in the context of linked identity collected would mean whether the collected identity pair indeed belong to the same user in the real world. We are expected to get valid linked identities as long as the users keep their identity lists and profile descriptions correctly updated in methods namely Social Aggregators (SA) and Self Disclosure (SD), respectively. In the case of the Advanced Search Operator (ASO) method, if the

[10]International Standards Organization: https://www.iso.org/standard/45481.html

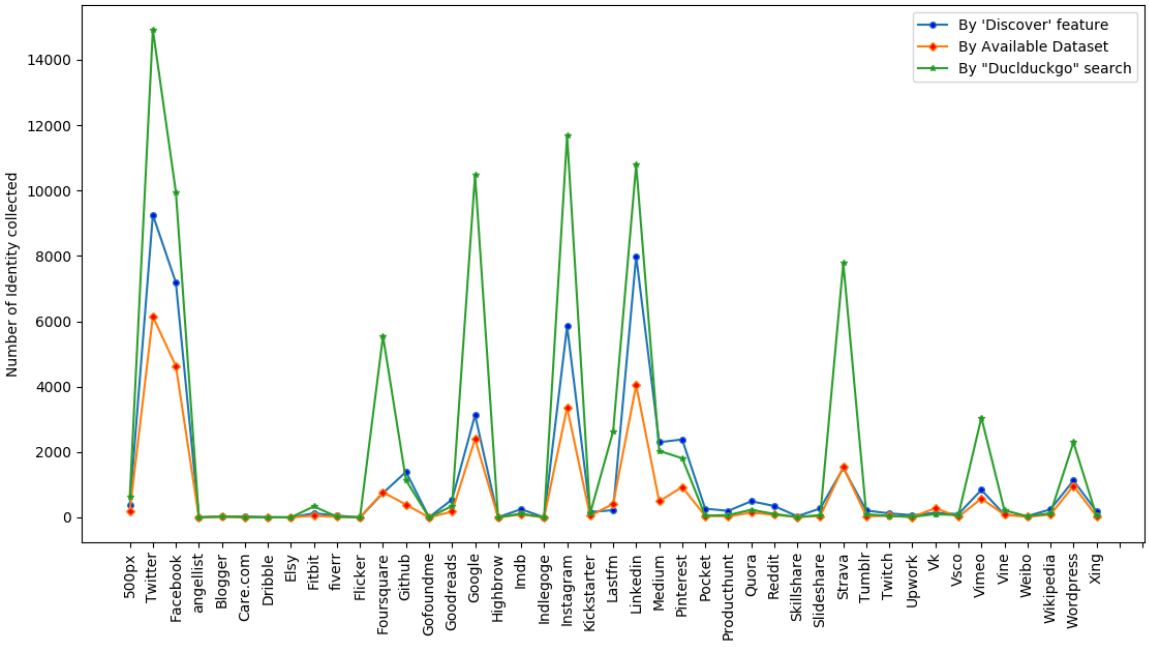

Fig. 10. Distribution of social network covered using Social Aggregator (SA) method for collection of linked identities. This method by far is the best in terms of OSN coverage with total 43 OSNs covered.

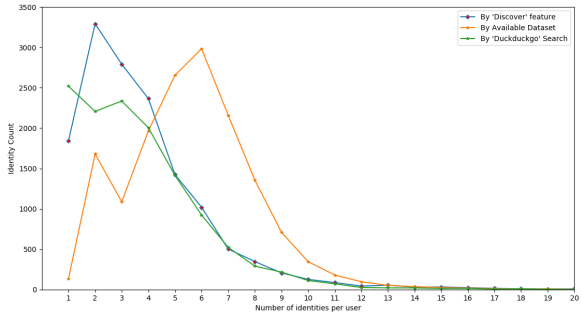

Fig. 11. Distribution of per-user linked identity count using Social Aggregator (SA) method for collection of linked identities. It may be noted that 24 users identity count more than 20 have not been plotted in this graph to keep visualization comprehensible.

TABLE IV
COMPLETENESS ANALYSIS OF DATA COLLECTION METHODS.

Method	Remarks on Completeness
ASO	Depends on number of social identities submitted by user to any server whose data is indexed by search engines
SA	Depends on number of social identities displayed by user on social aggregator sites
CPS	Depends on cross sharing activity of user
SD	Depends on amount of URLs mentioning identities on other OSNs revealed by user in his/her account description
FFF	Depends on availability of friend-finder feature on OSN and friends having account on that OSN

indexed file is quite outdated, then linked identities could be stale.

- **Consistency**: While each of the data collection methods would execute consistently, however, due to the dynamics of the OSNs, the results for each run could vary. Some OSNs provide greater re-configurability in user profiles, for instance, username can be changed in Twitter and Instagram. If a given data collection is relying upon username, then results would vary over time.
- **Accuracy**: All the methods rely upon user-contributed information. In the case of Advanced Search Operator (ASO), it is the data entered into servers which are indexed by search engines whereas in Social Aggregator (SA) and Self Disclosure (SD) methods is directly dependent on the information provided by the user. As long as user-supplied information is accurate, the data collection methods are guaranteed to return true positive linked identities.
- **Timeliness**: In the context of our problem, timeliness would mean whether linked identities for a given input user can be provided by the data collection method whenever requested. Out of the five methods, methods namely Self Disclosure (SD), Cross Platform Sharing (CPS) and Friend Finder Feature (FFF) method could be employed in such a situation.

V. CONCLUSION

Users are joining multiple on-line social networks for different purposes. An essential first step in the social profiling of

users is to *link* their identities across OSNs. Besides profiling, other important applications include recommendations and link prediction. In this work, we explained five data collection methods and compared them both qualitatively and quantitatively. Based on our experience of collecting linked identities across multiple social networks, we list down few suggestions for prospective researchers. Social Aggregator (SA) method is useful in the scenario when we want to study user behavior across a large number of OSNs. Self Disclosure (SD) method would yield a good coverage of OSNs but in a limited manner. On the contrary, if one has to target only a specific pair of OSN, then Cross Platform Sharing (CPS) method would be a good option. Advanced Search Operator (ASO) method would be useful if only popular social networks (like Facebook, LinkedIn, Twitter, etc) are to be targeted. Friend Finder Feature (FFF) is applicable only when a large pool of emails are available. FFF would also be useful in the scenario when one has to investigate an unexplored social network.

There are a few limitations to our work. For Social Aggregator (SA) method, we have investigated $about.me$, it would be interesting to extend it over other platforms like Google+. Similarly in Advanced Search Operator (ASO) method, we may go beyond $google$ search engine and explore another search engines like bing, duckduckgo, etc. In Cross Platform Sharing (CSP) method, we have taken Twitter as the target social network, which can be extended to include other OSNs as well. Similarly, only Twitter's bio field is being parsed in Self Disclosure (SD) method.

Finally, for ethical reasons, all data collection methods in this paper operate on public data only and rely upon the fact that the user has shared this data explicitly at some point in time. However, users may not be aware of the implications of public availability of their data. For users who are privacy concerned and would not want their identities to be linked, it is highly recommended that they should not cross-post content across OSNs, not provide details of other OSNs on their social media profile pages, use a specific email (not known to their friends) for registering at OSN and not register on websites whose *robots.txt* allows crawling. However, regardless, this work is a step towards a tool [4] that can help users understand and control the amount of their data that is available on OSNs so that they could safeguard themselves from online profiling.

REFERENCES

[1] B. Krulwich, "Lifestyle finder: Intelligent user profiling using large-scale demographic data," *AI magazine*, vol. 18, no. 2, pp. 37–37, 1997.

[2] W.-S. Yang, J.-B. Dia, H.-C. Cheng, and H.-T. Lin, "Mining social networks for targeted advertising," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, vol. 6. IEEE, 2006, pp. 137a–137a.

[3] M. Huber, M. Mulazzani, M. Leithner, S. Schrittwieser, G. Wondracek, and E. Weippl, "Social snapshots: Digital forensics for online social networks," in *Proceedings of the 27th annual computer security applications conference*. ACM, 2011, pp. 113–122.

[4] R. Kaushal, S. Chandok, P. Jain, P. Dewan, N. Gupta, and P. Kumaraguru, "Nudging nemo: Helping users control linkability across social networks," in *International Conference on Social Informatics*. Springer, 2017, pp. 477–490.

[5] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "Hydra: Large-scale social identity linkage via heterogeneous behavior modeling," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, 2014, pp. 51–62.

[6] X. Mu, F. Zhu, E.-P. Lim, J. Xiao, J. Wang, and Z.-H. Zhou, "User identity linkage by latent user space modelling," 2016.

[7] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *Proceedings of the VLDB Endowment*, vol. 7, no. 5, pp. 377–388, 2014.

[8] Y. Shen and H. Jin, "Controllable information sharing for user accounts linkage across multiple online social networks," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 381–390.

[9] H. Zhang, M.-Y. Kan, Y. Liu, and S. Ma, "Online social network profile linkage," in *Asia Information Retrieval Symposium*. Springer, 2014, pp. 197–208.

[10] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2013, pp. 179–188.

[11] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, "Discovering links among social networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 467–482.

[12] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2011, pp. 1–17.

[13] A. Malhotra, L. Totti, W. Meira Jr, P. Kumaraguru, and V. Almeida, "Studying user footprints in different online social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, 2012, pp. 1065–1070.

[14] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 41–49.

[15] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 447–458.

[16] O. Goga, D. Perito, H. Lei, R. Teixeira, and R. Sommer, "Large-scale correlation of accounts across social networks," *University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002*, 2013.

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

[18] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach." in *IJCAI*, 2016, pp. 1823–1829.

[19] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Entity matching in online social networks," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 339–344.

[20] S. Labitzke, I. Taranu, and H. Hartenstein, "What your friends tell others about you: Low cost linkability of social network profiles," in *Proc. 5th International ACM Workshop on Social Network Mining and Analysis, San Diego, CA, USA*, 2011, pp. 1065–1070.

[21] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name?: an unsupervised approach to link users across communities," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 495–504.

[22] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 707–719.

[23] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," in *Proceedings of the eleventh international workshop on Web information and data management*. ACM, 2009, pp. 67–75.

[24] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *2009 30th IEEE Symposium on Security and Privacy*, May 2009, pp. 173–187.

[25] Y. Nie, Y. Jia, S. Li, X. Zhu, A. Li, and B. Zhou, "Identifying users across social networks based on dynamic core interests," *Neurocomputing*, vol. 210, pp. 107–115, 2016.