Sampling Labelled Profile Data for Identity Resolution

Matthew Edwards, Stephen Wattam, Paul Rayson and Awais Rashid
School of Computing and Communications
Lancaster University
Lancaster, United Kingdom
Email: {m.edwards7,s.wattam,p.rayson,a.rashid}@lancaster.ac.uk

***Abstract*—Identity resolution capability for social networking profiles is important for a range of purposes, from open-source intelligence applications to forming semantic web connections. Yet replication of research in this area is hampered by the lack of access to ground-truth data linking the identities of profiles from different networks. Almost all data sources previously used by researchers are no longer available, and historic datasets are both of decreasing relevance to the modern social networking landscape and ethically troublesome regarding the preservation and publication of personal data. We present and evaluate a method which provides researchers in identity resolution with easy access to a realistically-challenging labelled dataset of online profiles, drawing on four of the currently largest and most influential online social networks. We validate the comparability of samples drawn through this method and discuss the implications of this mechanism for researchers as well as potential alternatives and extensions.**

***Index Terms*—replication; privacy; identity resolution; social web mining.**

I. Introduction

Identity resolution tasks are a form of classification whereby two or more profiles of a person – often from different databases – are matched together based on the similarity of their features. The aim is to identify multiple profiles referring to the same individual, where a profile may include everything from simple biographical attributes to inferred characteristics such as writing style. The aim of identity resolution is to allow different sets of information about a person to be connected.

A number of solutions have been proposed specifically for identity resolution tasks across social networking sites (SNSs), each making use of some part of the diverse feature set available in social network profiles [1], [2]. Yet without a common frame of reference to work against, these various approaches and results are difficult to compare, which hinders identification of the best-performing methods and the direction of future research.

In many machine learning domains, research is advanced by the sharing of labelled datasets for purposes of replication, validation and incremental improvement on methodology. However, ethical constraints can prevent the dissemination of such datasets when they contain significant personal information, such as is always the case with profile data from SNSs [3]. While this profile data is nominally public information, as accessible as newspapers, it would be irresponsible to assume that personal information embedded in a public profile dataset is safe to preserve forever, and allowing members to later excise their data would pose significant obstacles to maintenance and consistency of instances of the dataset. Attempts have been made to anonymise these resources, but numerous de-anonymisation attacks have been demonstrated against such ostensibly anonymised datasets [2], [3], [4].

Rather than provide a single common dataset, we propose a sampling method which should allow researchers to independently gather *comparable* datasets. We take this approach to overcome the tension between the research need for replication and ethical handling of personal information. We propose, implement, and evaluate a sampling tool for gathering labelled connections between online instances of profiles, and also for gathering suitable negative data – real profiles which a classifier may be realistically asked to discriminate from the actual linked target. The output of this tool is a labelled dataset of profiles suitable for training and evaluating systems aimed at resolving identities across different SNSs.

Providing a tool rather than a dataset allows for comparable samples of linked profiles to be independently harvested by researchers from publicly available data on SNSs, without need for public release of actual profile data snapshots.

Our aim in this publication is to demonstrate that data collected by different researchers using this tool will be sufficiently comparable that their methods and results can be contrasted with some confidence, while at the same time they are working with data realistically reflecting the current social networking landscape. This approach also allows individuals and SNSs to determine between them what information is to be revealed to the public and does not presume upon any improper access on the part of researchers acting as part of that public.

This paper is structured as follows. In Section II, we survey historic and existing sources of ground-truth data as used in previous studies, identifying issues with these sources. In Section III, we outline the sampling method we propose along with some requirements for implementing it. We demonstrate one such implementation in Section IV, and use two large samples gathered via this implementation in Section V and VI to validate that samples drawn through this method are comparable. We conclude in Section VII by discussing our results and some outstanding issues in this area.

II. Ground-Truth Data Sources

In aid of identifying suitable methodology, we survey the data sources employed in existing literature on identity resolution across social networking sites.

Malhotra et al. [5] in 2012 made use of three separate sources: Google's Social Graph API, and two social aggregators, FriendFeed and Profilactic. Of these three sources, none are still operational. This is a recurring pattern with social aggregation services similar to FriendFeed. Many exist or have existed, marketing themselves to users on the basis of consolidated access to multiple social networks, but they commonly go out of operation or are bought up by dominant social media organisations which repurpose their assets. This is disappointing, because as Malhotra et al. and also Jain et al. [6] with their small Social Graph API dataset and Irani et al. [7] with their unnamed single aggregator site all demonstrate, these sites can be a rich source of ground truth data whilst they exist.

One of these services, Plaxo [8] (which now operates as an online address book service, with mostly private profiles), has released a tool which highlights how user annotation of links might be utilised by researchers to gather labelled profile linkage data, relying on `rel='me'` annotation within the anchor tags for links as part of a crawler. To make suitable use of this annotation, researchers would first have to gather a large random sample of profiles which contain annotated links. Though they do not explicitly state their collection method, Buccafurri et. al [9] appear to have made use of such `rel='me'` annotations and/or Friend-of-a-Friend (FOAF) data (see below) in identifying cross-links between profiles on LiveJournal, Flickr, Twitter and Youtube, a dataset which was later enriched by Bennacer et. al [10]. In this dataset of 93,169 nodes, only 462 unique cross-links are identified, suggesting such annotations are not in widespread adoption.

Golbeck and Rothstein [11] used FOAF semantic data obtained from a number of social networking sites, looking for specific shared traits in FOAF files such as chat IDs or homepages in order to identify profiles of the same person. The FOAF format – being a common format for description of profiles and their interconnections – would be theoretically ideal for gathering linked profiles, if it were widely supported by large SNSs. However this does not appear to be the case, with LiveJournal the lone popular exception amongst a largely niche set of small SNSs which support it.

Goga et al. [12] made use of the Friend Finder functionality which was formerly common on many social networks, using an existing list of 10 million email addresses to find users' accounts present on multiple social media platforms. Due to several privacy concerns raised by the feature, many social networks no longer allow email-based search for profiles, most notably Facebook [13]. Even were the functionality still available, the email addresses required in order to utilise it to gather linked profiles are typically more closely guarded than other profile information.

Narayanan et al. [2] take a somewhat different approach in their de-anonymisation study, basing their ground-truth mappings between profiles on exact matches in the username or name fields, attempting to verify such matches with a score generated from a small number of heuristics – the length and rarity of the name, and overlap in location information. As their method (topographical identification) did not rely on any of these features, this linkage method retains validity within their study, but it cannot easily be generalised as a means for other researchers to go about acquiring ground-truth mappings for identity resolution.

Based on an exploit discovered by Kaafar et. al [14], some researchers make use of the optional 'other profiles' feature of Google Buzz profiles to identify cross-links between profiles from different networks. They gather a large dataset of some 4 million profile identifiers from Buzz, a predecessor of the Google+ social network, using a graph-based crawler which collects lists of Follower/Following users from each profile. A large proportion of these profiles made use of the 'other profiles' feature, and as such this dataset has gone on to be reused in several other studies on identity resolution across SNSs [1], [15], [16]. However, Google Buzz was discontinued in 2011, and its successor Google+ does not make a profile's Circles (the Follower/Following relationship being abandoned with Buzz) easily accessible for scraping.

Based on this survey, it appears that the majority of previously employed datasets in this area of identity resolution come from sources which are no longer available for re-sampling. Those datasets which may theoretically be re-sampled in the same manner are of limited value, covering only small user populations.

III. Sampling Method

If we are to avoid making assumptions based on usernames, and cannot rely on the availability of unique identifiers persisting across SNSs (such as email addresses), then the search for ground truth data is effectively a search for instances where a user has stated a connection between two or more of their own profiles. Social aggregation services are one means by which such information may be collected. However, they appear to be an unpredictable source, not suitable for the basis of long-term research. If social aggregation services cannot be relied upon as indexes, then it may be better to examine the social networks themselves for users' revelation of connections to other networks.

This is similar to the approach used in the tool released by Plaxo [8], which examines the `rel='me'` property of links to find links which a user identifies as being another profile of theirs. This annotation does not appear to be in widespread adoption, but it may be possible to find alternative indications that a link is intended to represent another profile of the user.

Presuming for the moment one such SNS where we expect to find this ground-truth link data, which we will term the *primary study network* or *primary network*, the problems can be stated as follows:

1) Gathering a representative random sample of profiles from the *primary network*. Notably, we are not interested

in identifying the most connected users or in sampling a connected subgraph of the *primary network*, only in a random selection of profiles (or in graph terms, nodes). Previous efforts focused on crawling large graphs of SNS users through the application of breadth-first search or random walks [17] are unable to reach disconnected components of the overall graph and are usually biased towards popular nodes by early stopping.

Most desirable would be methods which can directly sample from the network, such as the ability to randomly select from assigned unique identifiers, but these indexing mechanisms are usually not publicly available. As an alternative, we suggest that the network search functionality provided by many SNSs can be used to gather unbiased samples of profiles. This functionality is provided to users to enable them to find other users based upon their name or other information. Given a random selection of search attributes (such as can be constructed based on population data such as census records), these search systems can provide a random index into the SNS's profiles.

2) Identifying in randomly selected profiles those linked profiles which belong to networks of interest. While links act as identifiers for a profile, extracting the profile content is an involved process highly dependent on the network being targeted. As such, it is prudent to focus on a few such networks – *secondary study networks* – and discard links to other networks.
3) Gathering plausible negative examples for a 'realistically challenging' dataset. A sample consisting of only those profiles which are known to be matched would be of little use for training and evaluating a classifier. As well as positive examples of profiles which should be matched, an appropriate sample should be made of those profiles which are not matched in other networks, for both primary and secondary study networks.

   For any profile, it would be possible to use other profiles in the same network as negative examples, but these profiles would make for a poor candidate set, being mostly easily distinguishable from the true results. Instead, we opt for a candidate set which more reasonably reflects real disambiguation tasks with public social network data – search results in the *secondary study network*, with the query being constructed based on attributes of the *primary network* profile from which a link was found. We believe such a dataset better reflects a core issue of identity resolution: given a particular individual profile, how do we find out which of many profiles with the same name are the ones to be connected?

Note that users voluntarily complete these fields in their profiles, and so as with previously discussed datasets, the datasets we aim to generate may not be valid for *adversarial* profile linkage tasks, where the emphasis is on detecting a link between a user who is attempting to mask any connection between their two profiles. Nor should the sampling method be taken to enumerate all matching profiles in the *primary network*, or any similar property which assumes an exhaustive exploration of any of the study networks. The dataset should remain relevant for purposes such as estimating the privacy impact of revealing certain profile attributes, testing existing identity resolution methods and comparing behaviours between the same individuals on different social networks.

Considering these issues, the requirements for our method are:

1) A *primary study network* in which users provide links which can be understood as statements that the link refers to another profile of theirs. This network must have a network search system which can be used for random sampling of profiles.
2) A set of *secondary study networks* which are linked to from the *primary network*. These networks must have an index suitable for selecting negative examples.

IV. Implementation

One of the most promising data sources as a *primary network* for implementing this sampling of ground-truth data would appear to be Google+. As previously mentioned, Google+ provides an "other profiles" field on a person's profile page where users can provide links to their profiles elsewhere on the web. This field is accessible via the Google+ API and so it is possible to automatically examine the Google+ network to find profiles which link to other profiles of the same person.

There are other reasons to favour the selection of Google+: while it is difficult to predict the shifting landscape of SNSs, Google as an organisation seems unlikely to disappear in the short-term, and it seems reasonably likely to maintain the Google+ service or an equivalent network for the next few years. At the same time, a number of influential studies referenced above have historically made use of a dataset drawn from Google profiles.

The *primary network* must also have a search system which can be used to perform random sampling from the network. For this, we draw upon the approach of Gonzalez et. al. [18], whereby a random sample of names from a large list of uncommon surnames are used as input into Google+'s profile search API, and those result sets numbering less than Google+'s cap on responses are taken as an unbiased sample of profiles. The aim of using uncommon surnames is to increase the likelihood of retrieving result sets numbering less than the results cap. Because the Google+ search API limits the number of returned profiles to a maximum of 300 per query[1], and these results are ordered by popularity, a sample which includes all search results would be biased towards more popular users. Therefore, we accept only those profiles returned by queries which have fewer than 300 results in total.

In detail, our method proceeds as follows.

1) Initial search terms are randomly selected from a list of 128,000 uncommon US surnames. Following Gonzalez et al. [18] this list was drawn from those surnames which

[1]At the time of publication for Gonzalez et al.[18], this limit was 1000

occurred more than 100 times and less than 1000 times in the US Census 2000[2].

2) The Google+ search API is queried for these terms. Those result sets with < 300 items are taken as unbiased.
3) The search phase completed, all publicly available data on the accepted profiles is downloaded via the Google+ API. Two formats are used to store the data – one which records the exact queries and the raw responses, and another which standardises the data into a Profile object.
4) The 'other profiles' sections of the Google+ profiles gathered are examined to establish the ground-truth true links. Where a link is made to one of our *secondary networks*, that link is queued for download and a record is made of the connection between the two profiles.
5) The full name attributes of the Google+ profiles are then gathered to create a second set of search terms.
6) This second search term list is then entered into the search functionality for each of the *secondary networks*, and the resulting profiles are queued for later download. These results form the realistic candidate set for attempted identity resolution from the seed profile.
7) The profiles indicated by the true links and the candidate sets from the name-based searches are then downloaded from their respective networks' APIs, and stored in the same manner as the Google+ profiles.

There are a few implications of this method which should be borne in mind. Firstly, surnames of profiles will be unusually distinctive as compared with a population average, though the procedure for selection of negative results given above should mitigate this impact. Secondly, these names are those which are uncommon in the United States. As previously addressed by Gonzalez et al, the diverse immigrant history of the United States combined with the US bias in Google+ membership would mitigate the US-centric aspect of this concern, but there are possible correlates of low-incidence surnames with recent immigration and thus socio-economic status and perhaps in turn lower digital literacy. Next, it should be noted that the sample mechanism used has only 128,000 different search possibilities, with a proportional chance of collision, and also a maximum theoretical result size of 38,272,000 Google+ profiles (though in practice there are likely to be far fewer than this). Finally, there will be at most 299 Google+ profiles with the same name, so for a method attempting specifically to discriminate between such profiles, its capability cannot be demonstrated as greater than this limit.

A. Secondary Study Networks

We are interested in selecting only nodes from a specific set of SNSs we term the *secondary study networks*. Extracting structured information from profile pages involves API queries for the content of profile pages identified by URLs, analysis of which must be specific to the social network in question. Additionally, name-based search functionality must be implemented for each social network being sampled, in order to furnish negative examples. We are therefore interested in finding the right social networking sites to form the initial set of *study networks* from which to draw our samples.

[2]Data on surnames occurring less than 100 times was not available

Network	Links Counted	Percent of Linked
youtube.com	322	22.5%
picasaweb.google.com	214	14.9%
facebook.com	195	13.7%
twitter.com	186	13.0%
linkedin.com	65	4.5%
blogspot.com	59	4.1%
google.com/reader/	26	1.8%
profile.live.com	24	1.7%
flickr.com	22	1.5%
yahoo.com	16	1.1%
instagram.com	15	1.1%
blogger.com	12	0.8%
tumblr.com	11	0.8%
soundcloud.com	10	0.7%

TABLE I: Most commonly linked profile networks.

A number of constraints exist, including that the network in question must make profiles public (to members of the network, if not the wider internet) and allow for name-based search. The main deciding factor for including a network will be whether a significant number of Google+ profiles link to profiles in the network, as this furnishes researchers with a greater number of positive examples to analyse, and focuses efforts on linkage tasks likely to be of more value in application scenarios. Our method could easily be extended to include less-frequently-linked networks, though researchers may need to select larger initial samples from Google+ to get representative sets of linked profiles. Using our proposed sampling procedure for Google+ profiles and examining those profiles with links to other networks, we counted the number of links to other networks.

As shown in Table I, the most common networks which were not other services owned by Google (which we might expect to be overrepresented, and are increasingly integrated into Google+) were Facebook, Twitter and LinkedIn. These top three networks would appear to be mostly suitable as *secondary networks*, with some minor caveats regarding their accessibility: for example, LinkedIn does not offer a global name-based search feature within its ordinary public API, but this functionality can be obtained through web-scraping calls.

V. Evaluation

Our primary evaluation of the sampling method is to compare the distribution of certain node attributes in different samples gathered by our implementation. The node attributes that are the simplest to compare in this manner are numerical, so we examine the distribution of certain numeric properties of nodes – such as counts of followers and posts – in different samples gathered from the Google+ and Twitter profile networks via our initial sampling method.

Using the methodology described above, we gathered two large independent samples from both the Google+ and Twitter networks. The two samples of the Google+ network had respective sizes of 4,986 and 11,719 nodes, while the samples

of the Twitter network had 8,259 and 17,862 nodes. These samples (henceforth Datasets 1 & 2) were gathered over Oct-Nov 2015 and Dec-Jan 2016 respectively. A number of numeric properties were recorded reflecting attributes of interest to identity resolution research.

We could attempt to demonstrate a lack of statistically significant differences between these samples by aiming to *fail* a statistical test such as the two-sample Kolmogorov-Smirnov test. However, the large sample sizes mean that a direct test for statistically significant differences between the two samples would likely be overpowered for the usual critical values, with a high chance of committing a Type I error and finding a false difference between the groups.

Rather than focusing on statistical significance, we can test whether there are important differences between the samples by comparing the effect sizes between the two samples. Table II shows comparisons between counts of attributes for each node. Cohen's *d* is the typical measure of effect size, but its calculation relies on assumptions of normality which are violated in social network data, which tends to follow power-law distributions. We instead use a nonparametric measure of effect size known as Cliff's δ which has been recommended for such situations [19].

Other properties of the two samples may impact their comparability for research purposes. We can more directly examine this by reference to the Kullback-Leibler divergence, also known as *information gain* when using one sample in order to approximate the other. This measure directly relates to our intended use of the sampling mechanism – as a means for researchers to compare results obtained on one sample with existing results obtained on a similarly collected sample. Table II reports the Kullback-Leibler (KL) divergence between the two samples, with measures discretised into 15 bins for computation[3]. As the KL divergence is non-symmetrical between distributions, the figures reported are the average of both directions of the measure.

Property	G+ $\lvert\delta\rvert$	Tw $\lvert\delta\rvert$	G+ KL	Tw KL
Age*	<0.01	–	0.01	–
NumFollowers	0.05	0.04	0.01	0.00
NumFollowing	–	0.03	–	0.00
NumInteracted	0.05	0.04	0.05	0.01
NumLocations	<0.01	0.02	0.01	0.00
NumTexts	0.03	0.05	0.00	0.01
NumDescribes	0.04	0.02	0.01	0.02
NumLinks	0.04	0.05	0.02	0.04
NumPics	<0.01	0.02	0.00	0.00
NumTimes	0.04	0.05	0.13	0.01

TABLE II: Nonparametric effect sizes and average KL-divergence for comparison of samples from the Google+ and Twitter networks. *Age where available.

The KL results show that very little divergence is present between the two samples, or, alternately, that very little information is lost when using one to approximate the other. Similarly, The average of all δ for Google+ comparisons is <0.03 and for Twitter is <0.04, indicating a very low practical difference overall between properties in the two samples. The large sample size makes us confident that we are not failing to detect larger effects.

[3]Based on Sturge's formula, $k = \lceil log_2 n + 1 \rceil$

VI. Application of Existing Identity Resolution Approach

As a secondary evaluation of our proposed approach, we apply an existing identity resolution method to both of our datasets. This serves to illustrate a possible use of these samples and further validates the comparability of results drawn from different samples. The aim here is not to provide a novel and competitive classifier, but to demonstrate the viability of our suggested replication method.

Following Goga et al. [1], we investigated the three features they used for identity resolution: the *name*, *location* and *profile image* of each pair of profiles.

A. Username

Usernames have often been considered a useful feature in identity resolution. Perito et al. [15] provide a full treatment of this topic. However, facets of our sampling method make names unlikely to be effective features: the 'display name' feature was used to generate the negative examples, so all comparisons are between profiles with highly similar names.

The effect is that names are not highly discriminative features in the comparisons made in our datasets, as shown in Figure 1a. In fact, the average Levenshtein distance between matched pairs of profiles was actually greater than the distance between unmatched pairs (5.82 and 4.01 for matched vs 2.75 and 3.24 for unmatched). This is the reverse of the normally expected direction in broader comparisons.

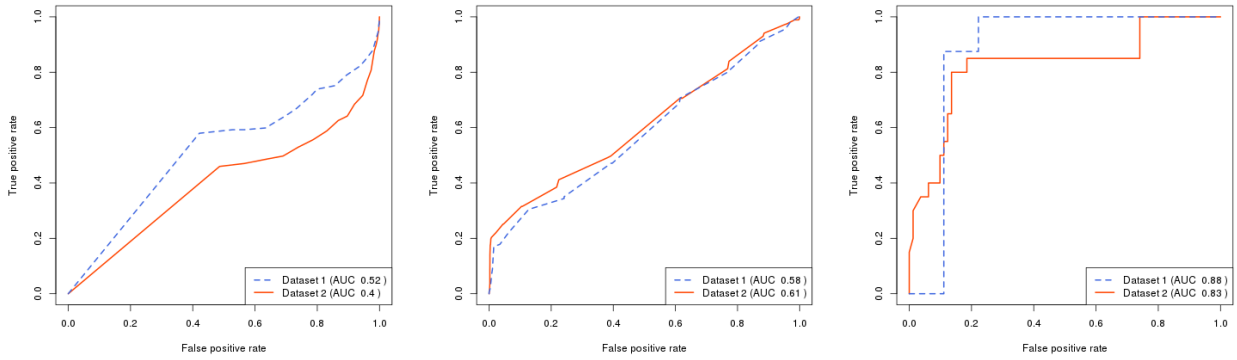
B. Image

We use a *perceptual hashing* technique to identify the key features of all profile images, and then calculate the Hamming distance between these two hashes [20], to test for superficial adjustments to the same avatar image. This feature showed some small but consistent discrimination, with the average Hamming distance between matched pairs being 27.72 and 27.02 for Datasets 1 and 2 respectively, and 31.66 and 31.94 between unmatched pairs. Just as Goga et al. discovered, simple threshold-based classification using this image feature has poor *recall*, but high *precision* – not many users do use the same profile image, but when they do they are very likely to be the same person. As Figure 1b shows, this means this type of image similarity performs poorly as a classifier by itself.

C. Location

Location data such as geolocated status updates or persistent 'hometown' or 'location' fields can be a good feature when it is available. However, location data is quite rare in our dataset, and this rarity is compounded by location comparisons only being possible where both profiles have location data: only 72 of 9558 comparisons in Dataset 2 and 17 of 1309 comparisons in Dataset 1 could use geodesic distance as a feature.

As Figure 1c shows, however, within this small subset, location distance was highly predictive.



(a) Name-based classification (b) Image-based classification (c) Location-based classification

Fig. 1: ROC plots for individual feature classifiers

D. Combined

We investigated the identifiability of these features jointly as part of a binary logistic regression classifier combining all three features, using a ten-fold cross-validation approach.

An important issue for classification tasks such as this is the handling of missing data. The majority of comparisons lack a location distance component, so how we handle this has a significant impact on model performance. Naively omitting records with missing data produces good-looking performance, as shown in Figure 2a, but tells us little about performance for the majority of cases. Imputing missing data with feature averages produces a more muted performance across more examples, shown in Figure 2b.

Performance in general was quite poor where location information was not available, unlike the findings of Goga et al. [1]. We can attribute this largely to the differences in the discriminative ability of the *username* feature, as this has poor performance within our dataset due to the manner in which negative examples are gathered and comparisons are made.

Our aim was not to provide a competitive identity resolution approach, but to demonstrate the comparability of results obtained through different samples via our methodology. We can see from the ROC plots that this is validated, with curves following the same trajectories with only minor deviations. Dataset 2 does tend to produce marginally better performance, but this is due to training benefiting from a larger sample size. Randomly subsampling 1000 data points from both samples produces a much closer match, as illustrated in Figure 2c.

VII. Discussion

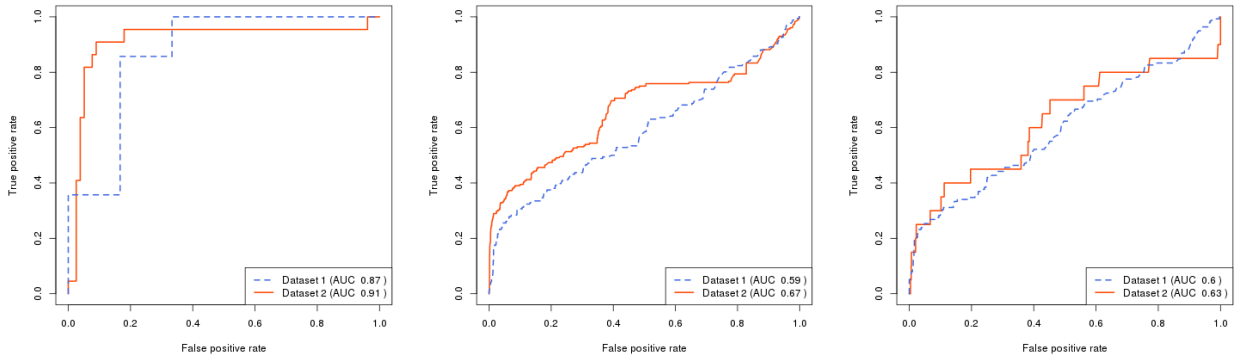
A. Implications for identity resolution research

We have presented a sampling mechanism for gathering ground-truth links between profile networks and appropriate negative examples, in proportion to their appearance in real-world data. Our evaluations confirm that samples being drawn in this manner are sufficiently comparable that methods developed against one sample should transfer to other samples drawn in the same manner with minimal impact – based on this initial analysis we should expect even small effect sizes to be replicated between experiments performed on different samples. We can also expect that ROC curves from a method trialled on one dataset to closely track those from another.

A common reference point for experimentation is necessary for researchers to compare their methodologies, and sampling mechanisms which reflect their population are necessary for properly grounding results. Both comparison and reference to the true population are necessary for advancing the state of the art. It is our hope that our sampling method will be used by researchers in identity resolution as a basis for reproducing each others' results and comparing identity resolution systems which make use of the heterogeneous data available in SNS profiles, something which has been hindered by the difficulties in obtaining and sharing such privacy-sensitive data.

The implementation we present focuses on the Google+ profile network as its *primary network*. However, our method is not restricted to application on just this network. Any SNS which provides a similar field to the 'other profiles' field within Google+ and makes this field publicly accessible would prove a suitable replacement. Indeed, recent work in identity resolution has started to recognise the identification use of URLs included in Twitter profiles [21]. While this field is less well-designated than the 'other profiles' field on Google+, and its utility as a source of ground truth must be investigated, it would provisionally appear to be a candidate replacement for the Google+ 'other profiles' attribute which would allow samples to be drawn with Twitter as the *primary network*.

Similarly, the implementation we present suggests that blocking – the generation of candidate record pairs for identity resolution – be based on the name of one or more profiles, as this is the search mechanism used for collecting negative examples. This is not necessarily problematic, as name fields are often used as blocking keys, but we note that alternative search systems can be used for finding candidate profiles, including searches based on content and network properties as described by Jain et al. [6]. Generally speaking, any property which can be used to generate negative examples from search



(a) Omitting records with missing features (b) Missing features replaced with means (c) Subsampling 1000 of each dataset

Fig. 2: ROC plots for combined classifiers

of *secondary study networks* can also be used for blocking.

This may be particularly important when considering the performance of classifiers which include profile name similarity as a key feature, as sampling negative results based on name necessarily reduces its utility as a distinguishing feature. Such a task, however, realistically reflects real-world challenges in disambiguating users with the same or similar names.

Finally, we note that while our approach is particularly tailored to research for identifying links between profiles on SNSs, the generation of accurate ground-truth data is a recognised problem for identity resolution in general [22], and it is possible that this sampling approach could be informative for researchers working within similar constraints, such as in bibliographical or medical record linkage.

B. Limitations of the tool

We have realised our implementation in a Python tool capable of sampling ground truth data from the *primary* and *secondary* networks given in this paper. The potential for one or more SNSs to alter or close their public API is a partial threat to continued functionality of our sampling tool. While the tool has been designed in a modular manner, so that secondary study network APIs which no longer work need not impair the general operation of the tool, it is likely that maintenance will be necessary to keep these modules functional. Policy changes on the part of the SNS may similarly affect the data this tool is able to provide to researchers. We note also potential improvements in the speed and reliability of the tool which could be achieved through sustained development.

In this work we have concentrated on development of a sampling tool which uses the APIs provided by the SNSs, using only the access rights granted to any app developer. This is ethically necessary: our position as members of the public ensures we do not gain improper access to the profile content of users by e.g. befriending them, or paying for profile information as an advertiser. Authentication with the SNS means their release of the data being sampled is tracked and recorded. However, use of the APIs for these services can be limiting – in some cases, content which a member of the public may view on the web is not available within the API.

A possible solution to these limits would be to apply web-scraping technology to enrich profile data. This would bypass many hurdles with API limitations. However, this is not a straightforward proposition: modern SNSs make extensive use of asynchronously-loaded content, with little profile information accessible at the initial page load. Scraping technology has advanced in step, but a scraper intent on accessing large numbers of profiles may also have to contend with accounts and IP addresses being blacklisted, necessitating greater infrastructural requirements – such as a cooperating network of machines – for any sampling tool, which would hinder replication. Overcoming these issues may require centralisation of the sampling tool as a service for researchers, which re-opens questions about sharing profile data.

C. Privacy & Ethics

The issue underlying the design of this sampling mechanism can be described as an ethical tension. It is easy for scientists to identify that making their results replicable is ethically necessary, this having long been a guiding principle of science. A direct approach to satisfying this replication requirement would be to release all the data used in an experiment, and in most areas this is still appropriate. At the same time, however, there is an increasing recognition of the paramount ethical obligations to protect the privacy of data subjects [23]. Even where, as in psychology or the social sciences, waivers can be be gathered to permit the release of some personal information, only relevant data is collected and communicated, to reduce the risk of a subject being identified. In large-scale studies of social networks, contacting profile owners for approval would be impractical, and in the field of identity resolution in particular it is not sensible to talk of removing personally identifiable information from a data release (except perhaps as a research challenge). Researchers are presented with a difficult choice: either they never release their data, protecting their subjects but hindering the development of their field,

or else release it, and risk harm to their many subjects and perhaps also personal legal consequences.

Our contribution here has been to identify a means for researchers in identity resolution – and related fields – to fulfil their ethical duties to their profession and colleagues without revealing the personal information of their subjects, drawing upon the reachability of a common population for sampling purposes. However, our approach cannot be said to entirely remove the underlying tension. For one, researchers must remain cautious about how they store and present data from these samples. For another, the scraping countermeasures discussed above require a careful response: the decreasing availability of useful ground truth data about the identities of social media users may be a barrier for research in this field, but it could also be more positively viewed as an indication that social networking sites are becoming more protective of their users' privacy.

VIII. Conclusions and Future Work

We have identified a troublesome area for research into identity resolution, and proposed and validated a solution based on rigorous sampling from a commonly accessible population. This solution allows researchers in identity resolution to replicate results and compare methodologies. Future work in this area might focus on improving this approach by finding other practical search methods to use for random sampling, such as content or network-based search. Identifying a suitable *primary network* which allows more rapid collection and iteration over profiles would also be of benefit, improving the collection rate. Further-reaching improvements might be to identify alternative mechanisms for researchers to index social networks, perhaps by negotiating with SNSs for privileged access for sampling purposes, or by funding and building a social aggregation service specifically as a research resource.

Further goals for addressing reproducibility in identity resolution research should include the organisation of competitive events to spur development and comparison of new methods – itself necessitating a standard evaluation framework such as that provided by Köpcke et al. [22] – or else focus on a means of describing and sharing the data transformations necessary to take raw profile data such as that provided by our tool and create the data format used in classification.

Our implementation of the sampling mechanism using Google+ as a *primary network* is available online at http://www.github.com/Betawolf/identity-sampler, distributed under the Creative Commons Attribution-Noncommercial Sharealike 4.0 International Public License. This work is partly supported by EPSRC grant EP/N028112/1.

References

[1] O. Goga, D. Perito, H. Lei, R. Teixeira, and R. Sommer, "Large-scale correlation of accounts across social networks," Technical report, Tech. Rep., 2013.

[2] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P 2009)*. IEEE Computer Society, May 2009, pp. 173–187.

[3] M. Zimmer, ""But the data is already public": on the ethics of research in facebook," *Ethics and information technology*, vol. 12, no. 4, pp. 313–325, 2010.

[4] X. Ding, L. Zhang, Z. Wan, and M. Gu, "A brief survey on de-anonymization attacks in online social networks," in *Computational Aspects of Social Networks (CASoN), 2010 International Conference on*. IEEE, 2010, pp. 611–615.

[5] A. Malhotra, L. Totti, W. Meira Jr., P. Kumaraguru, and V. Almeida, "Studying user footprints in different online social networks," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ser. ASONAM '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1065–1070. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2012.184

[6] P. Jain, P. Kumaraguru, and A. Joshi, "@ i seek'fb. me': identifying users across multiple online social networks," in *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, pp. 1259–1268.

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

[8] Plaxo, "Building an open social graph," accessed: 2016-01-30. [Online]. Available: http://www.plaxo.com/info/opensocialgraph

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

[10] N. Bennacer, C. N. Jipmo, A. Penta, and G. Quercini, "Matching user profiles across social networks," in *Advanced Information Systems Engineering*. Springer, 2014, pp. 424–438.

[11] J. Golbeck and M. Rothstein, "Linking social networks on the web with FOAF: A semantic web case study." in *AAAI*, vol. 8, 2008, pp. 1138–1143.

[12] 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*, ser. WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 447–458. [Online]. Available: http://dl.acm.org/citation.cfm?id=2488388.2488428

[13] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel, "Abusing social networks for automated user profiling," in *Recent Advances in Intrusion Detection*. Springer, 2010, pp. 422–441.

[14] M. A. Kaafar and P. Manils, "Why spammers should thank google?" in *Proceedings of the 3rd Workshop on Social Network Systems*. ACM, 2010, p. 4.

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

[16] T. Chen, M. A. Kaafar, A. Friedman, and R. Boreli, "Is more always merrier?: A deep dive into online social footprints," in *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, ser. WOSN '12. New York, NY, USA: ACM, 2012, pp. 67–72. [Online]. Available: http://doi.acm.org/10.1145/2342549.2342565

[17] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, "Crawling social internetworking systems," in *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, 2012, pp. 506–510.

[18] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas, "Google+ or google-?: dissecting the evolution of the new osn in its first year," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 483–494.

[19] K. Meissel, "A practical guide to using cliffs delta as a measure of effect size where parametric equivalents are inappropriate," in *ACSPRI Social Science Methodology Conference*, 2010.

[20] E. Klinger and D. Starkweather, "pHash –the open source perceptual hash library," accessed 2016-05-19. [Online]. Available: http://www.phash.org/apps/

[21] P. Jain, "Automated methods for identity resolution across heterogeneous social platforms," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 2015, pp. 307–310.

[22] H. Köpcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 484–493, 2010.

[23] A. Markham and E. Buchanan, "Ethical decision-making and internet research," *Recommendations from the AoIR Ethics Working Committee (Version 2.0)*, 2012.