A Data-Reachability Model for Elucidating Privacy and Security Risks Related to the Use of Online Social Networks

Sadie Creese, Michael Goldsmith, Jason R.C. Nurse and Elizabeth Phillips
Cyber Security Centre,
Department of Computer Science,
University of Oxford, Oxford, UK
{*sadie.creese, michael.goldsmith, jason.nurse, elizabeth.phillips*}*@cs.ox.ac.uk*

***Abstract*—Privacy and security within Online Social Networks (OSNs) has become a major concern over recent years. As individuals continue to actively use and engage with these mediums, one of the key questions that arises pertains to what unknown risks users face as a result of unchecked publishing and sharing of content and information in this space. There are numerous tools and methods under development that claim to facilitate the extraction of specific classes of personal data from online sources, either directly or through correlation across a range of inputs. In this paper we present a model which specifically aims to understand the potential risks faced should all of these tools and methods be accessible to a malicious entity. The model enables easy and direct capture of the data extraction methods through the encoding of a data-reachability matrix for which each row represents an inference or data-derivation step. Specifically, the model elucidates potential linkages between data typically exposed on social-media and networking sites, and other potentially sensitive data which may prove to be damaging in the hands of malicious parties, i.e., fraudsters, stalkers and other online and offline criminals. In essence, we view this work as a key method by which we might make cyber risk more tangible to users of OSNs.**

***Keywords*-Social-network risks, online social networks, data-reachability model, privacy, security, information leakage**

I. Introduction

Online social media and networks (OSNs) are one of the prime uses of the Internet today [1]. A quick look at current statistics highlights the staggering number of active OSN users worldwide (e.g., Facebook (http://www.facebook.com) boasts in excess of 800 million), the extent to which they are contributing (Twitter processes in 230 million tweets a day [2]), and the substantial amount of time being invested in this medium (Nielsen [1] reports that it accounts for nearly a quarter of the time Americans spend online). As individuals engage with these digital environments, this results in both conscious information sharing and publishing, and also the creation of persistent data which the user may be unaware of, perhaps as metadata or as old data thought to be removed or put out-of-reach. Examples of this information include personal content posts, photos, family and friend connections, group memberships, locations visited and events participated in. Although the attention on privacy and security within OSNs has increased in recent years (as apparent in numerous articles [3–6]), the risks associated with this medium are still arguably intangible to many [7]; it certainly is not as intuitive as locking one's front door in the physical space. This, coupled with an appetite for digital technology which shows no sign of abating, means that by engaging with OSNs, users may be exposing their identities, lives, the groups they associate with and even their employers to a range of unknown and potentially serious privacy and security risks.

In this paper we expand the discussion above and propose a novel data-reachability model for assessing the privacy and security risks in the use of OSNs. This model focuses on elucidating how heterogeneous data extraction techniques might combine into a single capability, providing linkages between data typically exposed within social-media and networking sites, and other data points in the online and offline spaces. A core motivational question is therefore, if one has certain information about an individual (or persona) in the OSN domain, and given unrestricted access to the various tools and methods being discussed in the open, what additional information can be reasonably derived and with what *ease* and *accuracy*?

This research and the resulting model are significant particularly because they assist OSN users and interested stakeholders in understanding the range of potential correlations for data and extraction methods, and thus the potential level of risk exposure faced. It is hoped that this knowledge (to which a user-friendly interface can easily be developed) would be a key piece in the puzzle of making these risks more tangible to Internet users as they assess their profiles from the attack-centric view typically adopted by online criminals. In many ways, we regard our data-reachability model as the natural progression from related work in the security and privacy of OSNs [3, 7–11], because it explicitly traces what data can reasonably be derived from what is currently shared. These derivations are based on an amalgamation and synthesis of current research and general knowledge of the field. We are not aware of any other work that attempts to create such a comprehensive model for judging a user's consequential risk exposure.

The organisation of this paper is as follows. Section II reviews the OSN domain with special emphasis on social networks as a source of data, and the existing prospects for aggregation and mining of that data. In Section III we present an overview of our proposed Data-Reachability Model along with an explanation of how it works. This is then followed by the application of the model to a few real-world cases to exemplify its novelty and use. Finally, we conclude the paper and present future work in Section IV.

II. Reviewing the Online Social Networks Space

A. Social Networks as a Data Source

Social-networking sites are an ideal source of personal information. The types of obtainable information is essentially unlimited within the parameters of what an individual chooses to reveal about themselves and their peers. Gross and Acquisti [3], for instance, found information items commonly mentioned encompass dates of birth, addresses, phone numbers, relationship status, views and interests, and screen names on other online social network (OSN) sites. A more recent study [12] also highlights these and more attributes (e.g., hobbies, home town, education, favourites, religious views and political direction), and how openly they are shared by users in four popular networks.

Speaking to the ease with which these and other personal information items can be attained from social sites, there are typically three factors to be considered. These are a user's privacy settings (which countless studies including [4, 13] have shown are much lower than they should be), the strength of the privacy controls of the site (these are occasionally deficient as OSN providers grapple with the difficult task of balancing privacy and sociability), and the intrinsic ease of extracting such information (which will be greater for semantically tagged information for example, than for plain-text). There is also a growing case for a fourth factor that acknowledges the fact that involuntary information leakage may occur through one's friends [10, 14]; in some networks, however, this is still subject to the user's privacy settings in so far as is practically possible. In this paper we concern ourselves with understanding the risks associated with techniques which might be utilised by an attacker external to a particular social-network environment or application. Clearly, it is possible for the service and application providers to aggregate personal information on their users for various purposes either in an authorised or unauthorised manner [15, 16].

Further to the general discussion of what personal information is shared online, it is interesting to note that the amount of information revealed by an individual using social networking has been shown to vary considerably between OSN sites. For instance, Facebook users reveal their friends 81.98% of the time whereas members of Xing (http://www.xing.com) only reveal this information 47.25% of the time [17]. There is also a similar story in terms of the number of friends that users register across social-networking sites; users of Twitter have on average 65 friends, whereas on Facebook they average 142. Labitzke et al. [12] present findings to support this general variation as they assessed several OSNs including Facebook, StudiVZ (http://www.studivz.net/) and MySpace (http://www.myspace.com/).

A likely reason for some of the divergence in the type of information revealed is simply that certain information is more relevant to certain sites. For example, on more social sites such as Twitter and Facebook users may focus on thoughts and activities in the 'here and now', whilst more factual and static information may be placed on a LinkedIn (http://www.linkedin.com) profile which potential employers or business associates may peruse. Regardless of its fragmented location, as the personal information is online, it is still plausible to aggregate it and thus use these social and professional networks as sources of data.

A noteworthy and challenging reality from a data source perspective is that although individuals often utilise multiple social-networking sites, the semantic value of the information revealed across the sites is also subject to variation. Balduzzi et al. [17] probed eight different social-networking sites and after identifying the same user across different sites, it was detected that the individual's name was the most common factor to vary from site-to-site, with 72.65% of the sample having different names on two or more sites and 17.66% having four or more different names. Out of another sample, it was discovered that 34.49% gave different ages across sites. Sexual preference, however, was more accurate with only 7.63% out of the sample who revealed their sexual preference on more than one site giving more than one value.

Thus, despite research in [18] showing that the overall presentation of an individual online is the same as in real life, one can see that minor discrepancies are still apparent. The discrepancy in name across sites may be explained by the use of usernames or abbreviations (e.g., Joseph Denver on one site and Jo Denver on another), whereas the difference in sexual preference may be explained by the individual being unwilling to reveal their true sexual preference to their peers or their employer for fear of discrimination, but may be willing to reveal it in another social circle. Similar realities hold for the other attributes commonly portrayed on social sites. In general, such discrepancies may be the result of conscious or subconscious attempts to protect privacy by users. Our reachability framework provides a mechanism with which to begin questioning the degree to which it is possible to circumvent such privacy strategies.

B. Aggregating and Mining OSNs

We now briefly consider some of the existing prospects for aggregation and mining of data typically available from OSNs. This is intended to give insight into how the social data may be extracted and then be used. As the focus of

this paper is more towards the security and privacy risks, we present the more attack-oriented research within this domain.

In [8], the authors introduce and explain several methods for extracting personal data (e.g., preferences, interests, location) and social graphs (i.e., users and their friendship associations) from social networks, ranging from crawling public profiles to creating false profiles. Social graphs, they note, can prove very useful in detecting communities with common interests and even in inferring private information from an individual's friend. Bilge et al. [19] also engage in extraction research, this time with the aim of investigating the ease at which attackers can launch automated crawling and identity-theft attacks in the social-networking space, with the aim of gathering personal information and friendship associations. Their work and evaluation results show that through the use of two automated attacks, profile cloning and cross-site profile cloning, gathering this information is both possible and feasible, though highly illegal. Although the authors do not give access to their prototype attack system, they outline general guidance on its architecture, functions (including a CAPTCHA breaker) and research developments that it is based on. Similar to [8] therefore, this research provides a possible platform for aggregation of social behaviour data on an individual.

Yet another attack on an individual's identity online can be found in [20]. Here, researchers explicitly introduce the concept of an online social footprint which is an aggregation of a user's profile and information from various OSN sites. Their attack attempts to reconstruct an individual's social footprint by using a pseudonym (which is guessed, likely based on an e-mail address or some other source) or the individual's name. This work builds on the reality that users share different amounts and types of information across different sites ([17]). Their study finds that over 40% of an individual's footprint can be reconstructed if a popular pseudonym (i.e., an alias that is used across a majority of sites to which the user subscribes) is discovered and 10–35% can be reconstructed based on the person's name. This is of special interest to our work because reconstructed footprints via linked profiles/aliases provide a larger set of data which may in turn lead to more accurate inferences in our model (as will be discussed).

Perito et al. [21] propose a family of techniques that use usernames (which they argue are easy to collect as sites make them publicly available) to link online user identities. The first set of techniques draws on language model theory and Markov-Chain methods to estimate the uniqueness of a username, then, use this to determine whether to link profiles that possess the same username. These techniques are then extended to accommodate profiles that are linked but have different usernames. Therefore, given two usernames, their method determines how likely it is that these refer to the same individual. The evaluation results of both techniques are also quite favourable, with displays of high levels of precision in predicting linked aliases. In many ways, this approach fits perfectly with our work on alias linkage towards social-footprint reconstruction and resulting risk exposure.

Finally, even independent individuals have begun their own campaigns at raising awareness of the amount of information (inadvertently) shared online. This can be seen in sites such as PleaseRobMe.com and ICanStalkU.com which aim to educate individuals on how simple it is for anyone to monitor and track them based on tweets and Foursquare (http://www.foursquare.com) check-ins. This is a perfect example of how information shared or inadvertently leaked in the online world can have real life ramifications.

We do not have space in a paper of this size to detail all of the existing capability and research relating to data aggregation and mining pertinent to the OSN environment. Our view is that there is nothing particularly hard about the technology required to aggregate and mine social-network data as compared to any other data. In general it is easier to mine structured data than to extract value from free-text. There is significant focus within the research communities (e.g. [22]) on developing methods to handle free-text in environments for which a structure might be determined. Arguably social networks could provide such an environment, and the embracing of Web 2.0 is likely to encourage users to provide partial structures through more detailed 'tagging' of data. Therefore, for the purposes of our analysis we consider the aggregation/mining of social-network sites to be possible. The next section introduces our model.

III. The Data-Reachability Model

A. The Matrix and How it works

The Data-Reachability Model is encoded in a matrix, with associated analytics, and captures the ability to determine personal data using a range of published methods. Specifically, we identify a range of personal attributes of an individual which might be sourced from their online presence, ranging from usernames and email addresses to friends, other social relationships and profile photos. These attributes, which we refer to as *data points*, form the currency of the reachability task at hand: some points are readily available, certain combinations are potentially of high value to an attacker; a dangerous exposure thus arises if there are ways of acquiring the latter given the former.

For any claimed data derivation or inference method among those surveyed, we note the data points reached from a given initial set (with perhaps limited probability of success or confidence in the results), creating a row within the matrix, an excerpt of which is depicted in Figure 1.

One is thus able to use the model to document inference rules which when combined with a set of initial data points can result in the establishing of others. For any particular set of initial input data points, continuing the derivation process until no new information can be gleaned is tantamount to calculating the transitive closure of the derivability relation;

considering the inverse of this transitive closure allows us to calculate the minimal sets of data points necessary to derive a given data point or set of data points. Of course, the longer the chain of derivations required to reach a target set of data points from an input set, in general the greater the effort and probably the less the confidence in the result (although this latter point is a topic for further consideration).

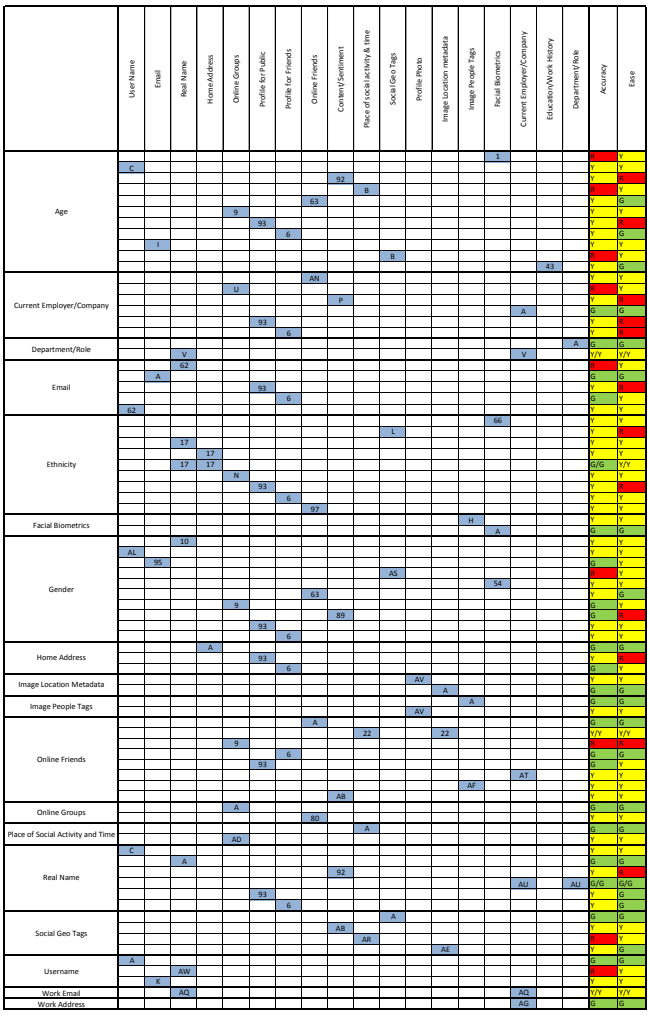

Figure 1. The Data-Reachability Matrix

In terms of the specific workings of the model, the matrix should be read from left to right with the left-most item defining the target information and the headings at the top specifying which data points may be combinable in order to derive the target. Filled out cells in a row mark a combination from which we believe or have evidence that such an inference may be possible. In essence, each row captures how a conjunction of data points can be combined to yield the data point in question; in general, there are several such rules for deriving each data point.

We have also used a coding system consisting of colours, numbers and letters. In the accuracy column, the colour chosen for an intersection represents the degree of accuracy with which a data point may be used to infer a piece of target information; green (G) is in excess of 70% accuracy, yellow (Y) is 35–70% and red (R) is less than 35%. Within the ease column the colour defines how easy it is to get a data point in cases where actual data extraction is necessary or to move from a data point to a target; green (G) is high ease, yellow (Y) is moderate ease and red (R) is difficult. In general, the accuracy ratings we assign according to references' claims, and rely on the scientific process to refine these claims over time, i.e., as the community applies the methods then the body of evidence for the claimed accuracy builds, and the matrix values are updated accordingly. The ease ratings are assessed according to a cost function. At the time of writing we consider three levels of ease: little or no skill required as method fully automated; medium skill required to apply the method (long hand or using tool support) which can be learnt; high skill required involving specialist experience (whether tools are provided or not). However, future work will look to expand and enhance this cost function further.

Numbers within cells represent inferences for which we could find published evidence. Conversely, letters within cells reflect inferences which we have postulated based on general knowledge of the field and several rounds of brainstorming. The numbers and letters shown in Figure 1 are part of a much bigger justification catalogue (much too large to reproduce in its entirety here) and as such, they do not specifically coincide with items this paper, i.e., numbers do not reflect references in this article's bibliography.

To give an example of the type of inferences made, we now consider a few of them from the model excerpt in Figure 1. In matrix point *#63* (which pertains to reference [9]), it was discovered that respectable levels of accuracy could be achieved when trying to derive characteristics including Age and Gender, based on a user's Friends information. We judge the ease with which this can be done as generally high because it only relies on friends (or a subset thereof) sharing their own ages, and having access to such information. Other research in [11] further supports this type of deduction based on Friends, and for Age in particular *#43* (i.e. [23]) also using Education History. For matrix point *#17*, which refers to ethnicity research by Fiscella and Fremont [24], we see that using geocoding (Home Address) to determine ethnicity is generally possible but may be less accurate for females than males. Furthermore, it was difficult to distinguish between Hispanic and Asians/Pacific Islanders, while only analysing surname (Real Name) did not give accurate inferences for African Americans. By using geocoding and surname together however, 80% of individuals could be correctly identified and there was a 90% accuracy for finding negatives. Moderate ease in using these names is supposed given that some surnames are culturally neutral or may come from long lines of ancestry. Point *#17* is a prime example of how combining data points can be useful

to achieve higher accuracy in deriving target information.

Two example justifications drawing on general knowledge include *U* and *AB*. For matrix point *U*, we conjecture that it may be possible to determine an individual's current employer/company from the Online Groups they are members of. For instance, if an individual is a member of the Multinational Corp. Sales Team then there is a high likelihood that they work for Multinational Corp. This method is relatively easy to utilise (as it only relies on viewing group members) but the accuracy of the information is low. This lower accuracy is attributed to the reality that unless it is a strictly official group, group members may be simply stakeholders who are past employees, or who like the company's products or follow them as a source of company news. Lastly, for point *AB*, it can be noted that even if a OSN user has made their Online Friends list private, it may be possible to infer friendships and Place of Social Activity from their online content (posts, recent activity, and so on). Take the example where a user updates their status to "Really enjoyed seeing Take That last night with Rosie Evans and Emily Thomas". This type of post would reveal that the individual is likely friends with Rosie and Emily and that they were at a Take That concert. This has a moderate level of accuracy as the information published is usually truthful but may be ambiguous, and medium ease because automated extraction using Natural Language Processing or Named-Entity Recognition can be somewhat challenging.

This completes our brief description of the matrix. It should be noted that the value of the model is in the approach, not specifically the values with which the matrix is populated. Indeed, different users of the approach might configure the matrix to represent their own unique perspectives on what they believe to be achievable for a specific data subject, given a particular threat capability and motive.

B. Applying the Model to a Scenario

In this section we discuss the model's application to a common scenario to demonstrate its ability to infer data and thus highlight areas where there might be unknown risks via information leakage, to a user's privacy and security. As Figure 1 captures only a subset of our model, occasionally we introduce other data points from the complete model which may also be derived. In these cases, and when not obvious, justifications are given to shed light on the inferences made. These will be kept brief however, so as to not detract from the overall model reachability discussions.

The scenario to be discussed features a user of online social-network site that has their Real Name, Online Friends information and Profile Photo of themselves publicly available. This is typically the default setting in numerous social sites, and arguably one that may seem relatively benign. Considering this initial set of data points therefore, the type of questions we attempt to answer using the model are, if someone (e.g., a malicious party) was to access this information: (i) what other information and data attributes could possibly be inferred? (ii) what derivations could be fairly made with medium-to-high accuracy? (iii) how far could an attacker get given that they were only interested in easy to reach information with at least moderate accuracy? These are all questions that will be on the minds of attackers when engaging in reconnaissance in preparation for identity theft and other personalised attacks, and as such also ought to be on the minds of OSN users as areas of potential increased risk when they share data online.

Using the model, one can quickly answer question (i) and infer the following data points (disregarding ease or accuracy). In Round 1, generally derivable points include: Age ([9]); Gender ([9, 25]); Ethnicity ([24]); Username (using variations of an individual's Name–similar to [20]); personal Email (again, using derivatives based on Real Name, in combination with popular email service domains, e.g., Hotmail, Gmail, AOL, Yahoo! [13,17]); Online Groups ([26] use Online Friends and their group memberships to predict this); Offline Friends (it is certainly not uncommon that online friends have offline relationships as well, e.g., [27]); Employer (if numerous of one's friends work for a particular company, although it may be somewhat tenuous, there is an increased likelihood one works there as well); Image Location Metadata (metadata is a rich source of information [28] and can be extracted from a Profile Photo given it was taken with a GPS-enabled camera or smartphone); and Image People Tags (from a Profile Photo, one may be able to identify/tag individuals—in some cases this may already be done thanks to OSN people tagging features).

Round 2 would lead to the following points (inferring based on Round 1 inferences and initial data): Facial biometrics data (through analysis of Image People Tags or identified in Profile Photos–[3] considers this from the perspective of identifiability); Place of Social Activity and Time (if an individual is a member of an Online Group then it may be possible to determine some of the individual's social activity based on the group's events); Social Geo Tags which show physical location (this can be collected from Image Location Metadata [28] which in modern cameras and smartphone devices is likely to include embedded GPS data); Username (this has already been derived from Real Name but given a scenario where one starts from an Email, the local-part could give Username [21] or potentially found using the search-by-email functionality in OSN sites); Work Email address (in numerous companies, Real Name plus Employer's domain name—which is easily searchable knowing Company name—can be combined in the format `FirstName.LastName@companydomain`, `FirstInitial.LastName@companydomain`, and so on, to derive this address which can then be verified to some extent by sending a test email; big companies such as IBM even provide searchable employee directories online, i.e., http://www.ibm.com/contact/employees/servlets/lookup, that

give this data and contact phone numbers); Department or Role (this can be established provided that the Company maintains public-facing home pages for its employees, or alternatively one may search using Real Name and Company name on sites such as LinkedIn); and Work Address (again, easy to lookup given Company name).

The inferences from Round 3 are: Links to other Social sites with potentially more complete Online Profiles if the individual uses the same or a similar Username or personal Email across sites. The Namechk (http://www.namechk.com) tool is advantageous here as it allows automated checks of a plethora of OSN and other sites to determine if a specified Username is taken. Research in [20, 21] can also be utilised to some degree at this point in linking online identities. Other Profiles may provide the same information and thus potentially validate the first site's (identity's) details, or complementary information which could allow us to build a more rounded social footprint and even apply the model again to the new data points to discover what else may be inferred. From just those three initial data points therefore, one can see how much information could possibly be determined by a resolute perpetrator. There has even been recent work [14] which suggests that simply hiding Friends lists (a key source of inferences above) is not in itself a panacea to protecting against inference attacks.

Question (ii) pertained to the derivations that could be made from Real Name, Online Friends and Profile Photo with medium-to-high accuracy. Using the matrix, the first round of inferences are: Age (moderate [9]); Gender (generally moderate [9] but slightly higher from Name [25]); Ethnicity ([24] alludes to relatively accurate); Online Groups (medium accuracy considering that albeit likely, you may not have exactly the same interests and therefore decide to join the same groups as your friends); Employer (this inference's accuracy is limited by the reality that the person may have worked for the company in the past or they work with, as opposed to for, it now); Image Location Metadata (on average, moderate accuracy seems appropriate given that even if the Photo was not taken with a GPS-enabled camera, one may still be able to extract possible location data from assessing background scenes); and Image People Tags (facial recognition software is useful in facilitating this tagging but nuances still remain that lower prediction accuracy).

The Round 2 inferences include: Facial Biometrics (moderate accuracy because studies [3] have found that individuals only display identifiable images 61% of the time and even then, hats, glasses and other items may reduce accuracy of biometric data); Social Geo Tags (metadata, particularly if GPS-based, supplies reasonably reliable location data [28]); Work Email address (high levels of accuracy are possible but this is bound by the difficulty in discovering which email address format is used by the Company); and Department or Role (high to moderate accuracy given that the Company Web site is properly maintained). There are no Round 3 inferences to be made which satisfy the accuracy criteria at this level; even if there were, we might begin to discount their ease and accuracy after such a long chain of inferences. Generally, this and the previous paragraph highlight potential inferences that can be made with good levels of accuracy.

Finally, in question (iii) where the criteria is easy to reach information with moderate accuracy, only two inference can be made, namely Age and Gender; Image Location Metadata might also be possible given that the Photo was taken with a GPS-enabled camera or smartphone. Although this is not much, it is still useful, medium quality information for an attacker, that may then be used as a springboard to other assaults. The profile cloning technique [19] for example comes to mind as the malicious party would have a Photo of the target, their Name and Friend associations, and now a rough idea of birth year (from Age), Gender, and possibly photo metadata fragments. Assuming friends accept friendship requests from the cloned profile (a situation that has been shown to occur [19]) which is also supported by a near-accurate birth year, the attacker can now potentially view additional information on the target via the 'Friends of Friends' visibility feature (most common to Facebook) or even view the target's Profile depending on privacy settings.

Although not covered in this paper due to lack of space, other intriguing questions that our model will help to answer are: if a user shares their Username and Location data (possibly via a Tweet or Foursquare check-in), what else may be inferred with medium ease? Is there any way to ascertain a friend relationship based on location and time? If so, what degree of accuracy can be achieved and with what ease? Also, at the higher level, are there particularly easy or 'game changing' progressions from specific OSN data points to offline personal information? This may encompass where a person lives or works, details on family members, the person's educational/work history and even who their boss may be. These are all significant factors where a user's (or his associates') privacy and security are put at higher risk (e.g., in terms of identity theft, fraudsters, stalking).

IV. Conclusions

In this paper, we have introduced a data-reachability model for elucidating privacy and security risks and concerns in the usage of OSNs. There are numerous tools and methods under development that claim to facilitate the extraction of specific classes of personal data from online sources either directly or through correlation across a range of inputs. The main purpose of our model is to understand the potential risks faced should all of these tools and methods be accessible to a malicious entity. The model enables easy and direct capture of current capabilities through their encoding in a reachability matrix. Specifically, the model elucidates potential linkages between data typically exposed within social-media and networking sites, and other potentially sensitive data particularly in the offline world.

Based upon the research we have surveyed, which informs the current version of the matrix, we conclude that (i) even with a trivial and common set of attributed data points being shared via online social networks it is possible to derive a much richer identity that has the potential for much greater negative privacy and security risks and consequences than might have been assumed. Further, (ii) in some cases, it is reasonably easy to make these data inferences and with, at times, notable levels of accuracy.

V. Limitations and Future Work

Considering the broad nature of the OSN field and our ambitious aim to characterise it in a useful model, limitations are to be expected. The first point of note is that in some parts of our study, particularly in the assessment and subsequent ratings of inference accuracy and ease, we generally assume the average-case scenario with regards to quality and richness of data points. There are so many potential differences in what might be shared, the level of detail, and the information present across users, devices and systems, that capturing this in a succinct way would be impractical, hence our reliance on the average-case scenario. The model does however allow us to explore sensitivity to specific data points, in other words we can identify those which are most enabling. Therefore, we should also be able to identify *protective* measures in terms of data points to guard against sharing, in order to maximise privacy. This is something that we will be looking to incorporate in the future.

The next limitation focuses on the reality that, even though our study attempts to be thorough and reflect the current state of the art, considering the amount of material in the OSN field there may be some existing inferences we are yet to discover. Also, because we draw on existing studies and research, we take their claims and conclusions with regards to accuracy and ease as definitive, as we have *not yet* conducted a practical validation. As it pertains to ease however, we do anticipate expansions of our cost function to consider factors such as level of tool support, ease of deployment, cost of access to the method and likely success. This rating would need to be monitored and reviewed, and there is even the option of using a freshness rating to reflect when it was last reviewed. More generally, as this is a rapidly progressing domain, there are likely to be changes by OSN providers and new developments by researchers[1] that add new scope for inferences – and equally, where loopholes are closed thereby invalidating rules – that together impact the models' longevity as it now stands. At the least, an ongoing monitoring effort is necessary to keep it up-to-date and ensure matrix values are still relevant, particularly for 'game changing' inferences that are often used or create profitable links between sets of data points.

As stated earlier, we perceive the primary value of our research to be in the method, as opposed to the specific contents of the matrix; different users of the model might add their own insight into accuracy and ease, as well as performing such maintenance based upon new developments or personal experience. Indeed, we imagine that it is possible to tailor an analysis for a particular threat by focusing on specific levels of ease (which would be selected in-line with the expected threat capability) and by configuring the accuracy levels of interest depending upon the use with which the threat is anticipated to put the data to (i.e. medium to low accuracy is sometimes sufficient if part of an intelligence gathering operation or performed by well-resourced and indiscriminate attackers).

Other ways in which we may further this work include the critical evaluation of the model to validate the inferences made (both from the literature and our own) and the degrees of accuracy and ease currently associated with them. The findings of this evaluation will provide useful insight for our work on privacy and security risks and the field in general, and will be fed back into the model to allow any necessary updates of inferences and ease/accuracy levels. We have already made some initial progress on this task through experimentation directed at volunteer targets using techniques and tools that have been discussed in this paper. Furthermore, as appropriate, this evaluation may lead to a refining of data points to eliminate 'false positives' where we appear to be able to infer what is actually impractical.

Another area which we will focus on during our evaluation is the issue caused by noise and errors within chained inferences. For example, if an inference can be made from A to B, then B to C, and then C to D, if there was an error earlier on, how does this propagate and affect subsequent derivations? There is also the reality that a perfectly accurate A may lead to a partially accurate B, then if that is used to infer C and later, D, accuracy and precision are likely to decrease during this chaining. If found to be a persistent issue during our evaluation, we will need to find a systematic way incorporate this deterioration of quality into our matrix and the ease/accuracy levels. This process is however substantially helped by the existence of our current matrix as now we have a better idea of where and how to concentrate such chaining analyses.

Finally, as this model defines inferences in simple matrix format, there is scope for automating the current derivation process and additional chaining of inferences. As such, we have developed a prototype tool based on the current version of the matrix to explore how the individual inferences from our complete model might be composable to derive a given data point from a set of starting points that has not hitherto been considered. The tool is also able to represent the matrix

[1]Changes by providers include those which cause concerns regarding privacy and security (Facebook Timeline which is currently being rolled out is a prime example of this [29]) and researchers assessing how additional inferences can be made from new OSN interfaces. From an attacker's perspective, new designs are ideal as users are unfamiliar with them and, based on reports to-date, are likely to expose too much at least initially.

as a directed graph with accuracy and ease shown by edge colours and weights. Next we hope to create a user-friendly, query-engine interface for OSN users where they can easily see what inferences could be made by a malicious party based on what they are currently sharing.

It is worth noting that this work is part of a broader study considering a wide range of data points and aimed at exposing potential cross-domain inferences. Additionally, our work aims at identifying critical inferences and data points to guide privacy-friendly modifications to OSNs and targeted paranoia on the part of users.

REFERENCES

[1] The Nielsen Company, "Social media report: Q3 2011," http://blog.nielsen.com/nielsenwire/social/, 2011.

[2] The Huffington Post, "Twitter finally shares key stats: 40 percent of active users are lurkers," http://www.huffingtonpost.com/2011/09/08/twitter-stats_n_954121.html, 2011.

[3] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2005, pp. 71–80.

[4] D. Rosenblum, "What anyone can know: The privacy risks of social networking sites," *IEEE Security & Privacy*, 2007.

[5] Symantec Corporation, "The risks of social networking (white paper)," 2010. [Online]. Available: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/the_risks_of_social_networking.pdf

[6] The Telegraph, "MoD issues videos warning Twitter generation that 'careless talk costs lives'," http://www.telegraph.co.uk/technology/social-media/8574696/MoD-issues-videos-warning-Twitter-that-Careless-talk-costs-lives.html, 2011.

[7] C. Rose, "The security implications of ubiquitous social media," *International Journal of Management & Information Systems (IJMIS)*, vol. 15, no. 1, 2011.

[8] J. Bonneau, J. Anderson, and G. Danezis, "Prying data out of a social network," in *International Conference on Advances in Social Network Analysis and Mining*. IEEE, 2009, pp. 249–254.

[9] J. He, W. Chu, and Z. Liu, "Inferring privacy information from social networks," *Intelligence and Security Informatics*, pp. 154–165, 2006.

[10] I.-F. Lam, K.-T. Chen, and L.-J. Chen, "Involuntary information leakage in social network services," in *3rd International Workshop on Security: Advances in Information and Computer Security*, 2008, pp. 167–183.

[11] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *3rd ACM International Conference on Web search and Data Mining*, 2010, pp. 251–260.

[12] S. Labitzke, I. Taranu, and H. Hartenstein, "What your friends tell others about you: Low cost linkability of social network profiles," in *5th SNA-KDD Workshop on Social Network Mining and Analysis*, 2011.

[13] I. Polakis, G. Kontaxis, S. Antonatos, E. Gessiou, T. Petsas, and E. Markatos, "Using social networks to harvest email addresses," in *9th annual ACM Workshop on Privacy in the Electronic Society*. ACM, 2010, pp. 11–20.

[14] C. Tang, Y. Wang, H. Xiong, T. Yang, J. Hu, Q. Shen, and Z. Chen, "Need for symmetry: Addressing privacy risks in online social networks," in *25th IEEE International Conference on Advanced Information Networking and Applications*, 2011, pp. 534–541.

[15] BBC News, "Social apps 'harvest smartphone contacts'," http://www.bbc.co.uk/news/technology-17051910, 2012.

[16] MSN Money, "Are apps exposing you to ID theft?" http://money.msn.com/saving-money-tips/post.aspx?post=27fabeeb-7da2-489d-880c-99d7f2dde500, 2012.

[17] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel, "Abusing social networks for automated user profiling," in *Recent Advances in Intrusion Detection*, ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds. Springer, 2010, vol. 6307, pp. 422–441.

[18] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling, "Facebook profiles reflect actual personality, not self-idealization," *Psychological Science*, vol. 21, no. 3, pp. 372–374, 2010.

[19] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: automated identity theft attacks on social networks," in *The 18th international conference on World wide web*. ACM, 2009, pp. 551–560.

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

[21] D. Perito, C. Castelluccia, M. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science, S. Fischer-Hubner and N. Hopper, Eds. Springer, 2011, vol. 6794, pp. 1–17.

[22] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *Journal of Information Science*, vol. 36, no. 3, pp. 306–323, 2010.

[23] R. Dey, C. Tang, K. Ross, and N. Saxena, "Estimating age privacy leakage in online social networks," in *31st Annual IEEE International Conference on Computer Communications MiniConference*, 2012.

[24] K. Fiscella and A. M. Fremont, "Use of geocoding and surname analysis to estimate race and ethnicity," *Health Services Research*, 2006.

[25] C. Tang, K. Ross, N. Saxena, and R. Chen, "What's in a name: A study of names, gender inference, and gender behavior in Facebook," *Database Systems for Adanced Applications*, pp. 344–356, 2011.

[26] F. Shah and G. Sukthankar, "Using network structure to identify groups in virtual worlds," in *5th International AAAI Conference on Weblogs and Social Media*, 2011.

[27] K. Subrahmanyam, S. Reich, N. Waechter, and G. Espinoza, "Online and offline social networks: Use of social networking sites by emerging adults," *Journal of Applied Developmental Psychology*, vol. 29, no. 6, pp. 420–433, 2008.

[28] Metadata Working Group, "Guidelines for handling image metadata," 2010. [Online]. Available: http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf

[29] Sky News, "Facebook timeline timebomb: One week to adapt," http://news.sky.com/home/technology/article/16158906, 2012.