On the Linkability of Complementary Information from Free Versions of People Databases

Minaxi Gupta, Yuqing (Melanie) Wu, Swapnil S. Joshi, Aparna Tiwari, Ashish Nair, Ezhilan Ilangovan
School of Informatics and Computing, Indiana University
{minaxi, yuqwu, joshis, tiwaria, asnair, ezhiilan}@cs.indiana.edu

ABSTRACT

The privacy of hundreds of millions of people today could be compromised due to people databases which claim to store many personal details about individuals, often without their knowledge. While the paid versions of these databases may be prohibitively expensive for data mining on a mass scale, in this paper, we show that even the limited information provided by the unpaid versions of these databases can be effectively exploited for its complementarity and poses a significant privacy threat since an adversary can mine this information on a mass scale free of cost and then use it to his/her advantage, hurting the privacy of individuals.

1. INTRODUCTION

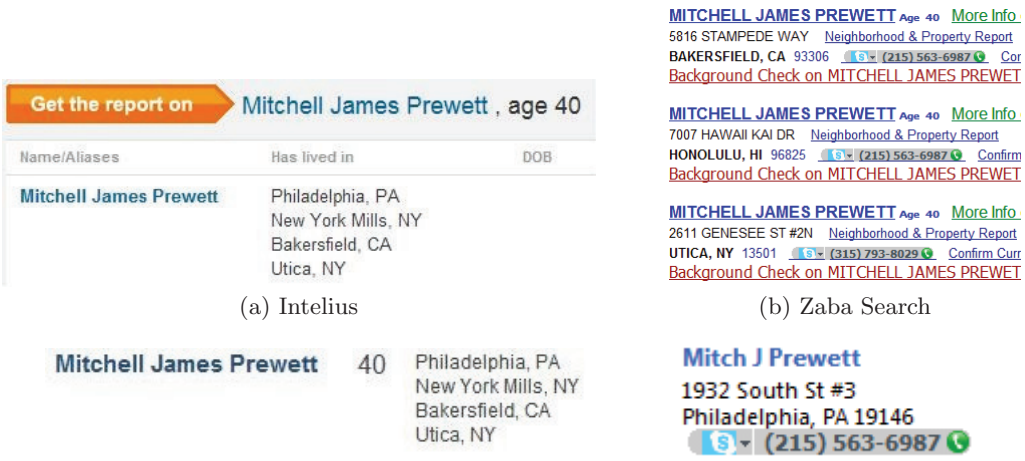
The easy availability of public records of individuals from city and state government offices, combined with an increased Web presence of Internet users, often in the form of Facebook and LinkedIn profiles, has led to the proliferation of so-called "people databases", such as Intelius [4], that contain a wide range of information about individuals, including their postal and email addresses, phone numbers, age, education, employment, property information, criminal records and relatives. Hundreds of such people databases exist, with most claiming to have information for tens to hundreds of thousands of individuals and some even boasting billions of records. Each of these databases pose privacy risks. While a few of them are free, most charge several tens of dollars in fee for searching records for a single individual, which can somewhat alleviate the privacy risk in that it makes en-mass searches prohibitively expensive. However, even the paid databases entice clients with limited free information, such age and location. In fact, unpaid versions of different people databases often contain complementary information, which if properly correlated, can prove to be a significant privacy risk by itself since anyone can collect this information on mass scale without having to spend any money and then use it to their advantage, hurting the privacy of individuals.

In this preliminary work, we examine the linkability of complementary information contained in the unpaid versions of people databases. Toward this goal, we use the concept of entity linkage or record linkage used by the database community to identify the same real-world entities referred to in different ways in multiple records. Specifically, we search for first and last name combinations of individuals in four popular people databases and apply a deterministic threshold-based algorithm to match records belonging to the same individual to estimate the privacy risk posed by ability to search these databases freely. Our algorithm uses the following four freely available attributes: full name, including the middle name or initial; postal address, including street address, city, state and zip code; age; and phone number.

Searching for over 850 combinations of first and last names belonging to real-world individuals, we find that almost half of them are indexed by at least one of the people databases we searched. This by itself poses a privacy concern since many of these individuals are likely to be unaware that anyone can pay to obtain detailed information about critical aspects of their life. Further, our algorithm identified that 1,316 individuals with these names were present in at least two of the four databases. Finally, for 30% of the individuals present in at least two databases, our algorithm revealed full postal addresses and phone numbers or additional cities they have lived in beyond what either database could provide. Clearly, this information is accessible to anyone who wishes to mine it and poses various kinds of privacy risks. As an example, it can be used to send targeted advertisements through the postal service based on the cities a person has lived in and his/her age profile. As another example, the phone numbers thus mined can be exploited to send targeted spam based on a person's location and age group.

Record linkage, similar in nature to ours, was the subject of the survey paper by Winkler [12]. However, the focus there was on census data which did not have issues of incompleteness or irregularities we had to deal with. Privacy issues in online social networks (OSNs), such as Facebook, have been studied along multiple dimensions. For example, Perito et al. in [7] investigated the issue of correlating user names across various OSNs and found that most user names contain identifying information which can be easily correlated across OSNs. The privacy risks of such a correlation are similar to those in our work. Works by Wondracek et al. [13] and Krishnamurthy et al. [5] looked at de-anonymization of users through group membership and leakage of personally identifiable information by OSNs to third party servers respectively.

To the best of our knowledge, the issue of privacy risks of people databases has remained a relatively under-studied topic. As of now, the mitigation of privacy risks posed by these databases is limited to commercial offerings, such as `reputation.com`, that claim to remove personal data from people databases for a charge [9]. We hope that our preliminary work will serve to expose the threats posed by these



(a) Intelius (b) Zaba Search

(c) BeenVerified (d) Switchboard

Figure 1: Example records in the four databases used in this paper

databases and in turn would lead to effective strategies individuals or database operators could use to mitigate privacy risks.

2. METHODOLOGY

Here, we describe how we choose the people databases to investigate, the algorithm we use for record linkage and our experimental setup.

2.1 People Databases

As mentioned before, there are hundreds of people databases. Some specialize in specific types of information, such as criminal records. Many use a small number of popular databases as their back ends. We selected four databases for our study based on the complementarity of free information they provided and avoided ones that used the same back ends. The first of our databases is Intelius [4], which is the largest of the pack, boasting two billion records. Many smaller people databases use it as their back end. For each record, its unpaid version provides full name, including the middle name, age, and all cities and states a person has lived in. Figure 1(a) shows an example of a record in Intelius. The second database we selected is BeenVerified [1]. Its unpaid information is similar to Intelius but we selected it to expand on the matches for people in our test set since it uses public records of corporations as its back end. An example of a record in BeenVerified is shown in Figure 1(c). The third database in our list is Zaba Search [14], which was chosen because for each record, it provides full address and phone number in addition to the middle name and age. It complements Intelius and BeenVerified in street address and phone number but does not provide all cities the searched individual may have lived in. Also, it uses Intelius as its back end, making it possible to correlate records with Intelius. Figure 1(b) shown an example of a record in Zaba Search. The final database we consider is Switchboard [11], which provides information similar to Zaba Search but uses a phone directory as its back end. An example of a record in Switchboard is shown in Figure 1(d). Note that correlating the various records for "Mitchell James Prewett" in Figure 1 across these four databases helps learn that this person has lived in five different cities. Also, postal addresses for four of these cities and two phone numbers can be learned.

2.2 Record Linkage Algorithm

Our algorithm exploits the presence of overlapping attributes present in various people databases to identify records that represent the same individual. It then takes advantage of the complementary nature of information to enrich records. In this preliminary study, we focus on four attributes: name, postal address, age and phone number. For each pair of records, we first compute a similarity score for each attribute. We then take a simple weighted average of scores across attributes to compute an overall similarity score for this pair. A match is declared when the overall similarity score exceeds a tunable threshold parameter. Subsequently, complementary information from these records is merged. We present the details next.

2.2.1 *Measuring similarity for each attribute*

We begin by describing how we measure the similarity score for the name attribute. The same name could show up in the following ways in different records: {John Tony Smith}, {John T. Smith}, {John NULL Smith}. They all have first and last names but the first has the full middle name, the second only the middle initial, and the last does not have the middle name. NULL values for middle names can be interpreted in multiple different ways. One might take it to mean that the individual does not have a middle name or that the database did not have the middle name available. Owing to these differences, a middle name match for the same first and last name gives rise to nine cases shown in Figure 2. The topological order among these cases dictates the similarity score for the name attribute and is shown in Figure 3. We assign a normalized score for this attribute based on this topological order.

A similar approach is applied to compare the address attribute, where the sub-fields – street address, city, state, and zip code – can have partial or NULL values. For the age attribute, the variations are exact age, age range, and NULL and we treat them similarly. Finally, the phone number could be complete, may lack area code, and may be NULL.

		Record 2		
		Full middle name	Middle initial	NULL
Record 1	Full middle name	match: case 1 not: case 2	contain: case 3 not: case 4	case 5
	Middle initial	contain: case 3 not: case 4	match: case 6 not: case 7	case 8
	NULL	case 5	case 8	case 9

Figure 2: Cases in comparing the name attribute in record pairs

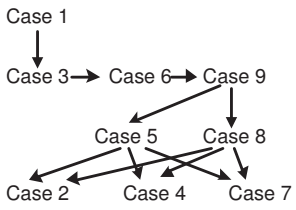


Figure 3: Topological order

We derive cases out of each of these in a manner similar to that for the name attribute and then assign a normalized score to each. The details of these attributes are omitted for brevity.

2.2.2 Computing overall similarity score

A simple approach to computing the overall similarity score would be to add the respective scores of each attribute. However, this approach ignores the common wisdom that the rarer the value of an attribute, the more confident we are in linking two records when they share the value. For example, two records for a John Smith living in Harvey, North Dakota, which has a population of 1,783, are more likely to be of the same person than those found in New York City. Similarly, the less popular a name, the higher the confidence that the records belong to the same person. While many bins could be created each for name popularity and city population and the scores of name and address attributes adjusted accordingly, we use a 3-step scale for each where either a name is unpopular, popular or in between. Similarly, a city's population is high, low or in between. In each case, we increase the score of the corresponding attribute if the name is unpopular or the city's population is low. Similarly, we decrease the score if the name is popular or if the city has a high population. For in-between name popularity and city population, we leave the score unchanged. At the end, we assign equal weights to each of the four attributes and sum up the total scores to derive the overall similarity score.

2.2.3 Enhancements

Thus far, we have assumed that each valid value of an attribute will have a fixed value (or a range, as for the age attribute). Here, we discuss the three enhancements that allow for realistic perturbations of the values of name, address and phone number attributes. The first enhancement allows for variations in first name, as in Bob vs Robert using name standardization [6]. It allows for people using different names in say, their social network profiles versus official records. We still assume that last names do not vary since most people do not change them for preference or convenience reasons. The second enhancement allows for variations in addresses, specifically, street addresses since all databases spell city names in their entirety. The people databases standardize directions in addresses, so "east", "west", "north" and "south" are represented by their first letter. The databases also uniformly shorten "street", "road", "avenue", "court", "boulevard", "parkway" etc. to their common 2-4 letter abbreviations. Further, street names with numbers are represented as such, as in "450 3rd Street" and not spelled out, as in "450 Third Street". However, we encountered extra white spaces on occasion and sometimes apartment numbers were missing in addresses from some databases. To allow for these variations, we compare street addresses between a pair of records after removing white spaces and allow for the possibility of one record's street address to be subsumed in the other. The third enhancement allows for variations in phone numbers. It allows for missing area codes and variations in which a three-digit area code and the last seven digit of a U.S. phone number can be delineated. Specifically, to account for these differences, we remove any parenthesis, hyphens and spaces before comparing phone numbers for a pair of records. We also allow for the possibility of one record's phone number to be subsumed within the other, which is useful in cases where the area code of even the next three digits of a phone number are missing. (None of the databases we use in this paper truncate attribute values, such as phone numbers or street addresses but there are some that do, such as Spokeo [10]). Finally, no accommodations were needed for age.

2.3 Experimental Setup

2.3.1 Names to test

We began by assembling a set of first and last name combinations to test. We used the 2010 census data set [2] which provides first names and last names in separate files, with separate files for males and females. Each file is sorted by popularity of that first or last name respectively. We pick 25 popular and relatively unpopular first and last names each from the top and bottom of each file, dividing the first names roughly equally among male and female names. Using these, we first derive all combinations of 25 less popular first names and 25 less popular last names to generate 625 relatively unpopular names. Since a name could be unpopular only because either the first or the last name is unpopular, we combine 25 popular first names with each of the 25 less popular last names to derive another set of 625 relatively unpopular names. Finally, we also derive the dual to this by combining 25 less popular first names with each of the 25 popular last names. Combined, these three lists give us 1,875 names to test. We ignore the combination where both the first and last name are popular to avoid dealing with dense data sets in this preliminary work. This data set will be explored in our future work.

2.3.2 Parameters for the record linkage algorithm

Next, we search for each name on our three lists in each of the four people databases mentioned in Section 2.1 and then apply the algorithm from Section 2.2 on each pair of records. Note that in general, people databases allow searching for other attributes as well though searches on names are the most popular. The similarity scores for each of the four attributes are normalized between 1 and -1, with steps guided

by the number of distinct cases. Specifically, upon some experimentation, we decided to set the similarity score for the name attribute shown in Figure 2 in case 1 to 1, in case 3 to 0.5, in case 6 to 0.25 and in case 9 to 0. Cases 5 and 8 were assigned a similarity score of -0.5 and cases 2, 4 and 7 a score of -1. Cases for other attributes were similarly handled. In matching a pair of records, if an attribute had multiple values, which was the often the case for records from Intelius and BeenVerified and occasionally for Zaba Search, we did a pairwise comparison for values for each attribute and took the maximum to be the similarity score for that attribute.

As for incorporating the enhancements, a name was considered popular if people databases returned more than 5000 records, less popular if they returned less than 300 records and in between otherwise. The similarity score for the name attribute for an unpopular name was increased by 0.25 and that of a popular name was decreased by 0.25. The similarity score for in-between name popularity was left untouched. Similarly, a city was considered densely populated if its population was more than 500K, sparsely populated if it had less than 10K residents and in between otherwise. (We use [3] to find population of all cities in our records.) Similar to the name attribute, the similarity score for the address attribute was increased by 0.25 for sparsely populated cities, decreased by 0.25 for densely populated cities and left untouched for all other cities. Finally, based on initial experiments, we set the threshold for overall similarity score to be 1.25 to declared a match among a given pair of records.

3. RESULTS

We experimented with the output of our algorithm on three different data sets, each containing 625 first and last name combinations, as described in Section 2.3. In this Section, we describe the results for each.

Data Set 1: Less Popular First Name, Less Popular Last Name: Searching for the 625 first and last combinations in this data set in each of the four people databases, Intelius, BeenVerified, Zaba Search and Switchboard, yielded 7,474 records. Often databases output records for similar names as well. So, searching for "James Werner" may also yield "James Warner". Since we allowed for variations only on first names (and that too only the commonly accepted shortened versions) and not last names, we ignored records that contained other differences in names. This filtering yielded 370 records for 72 names that we put through our record linkage algorithm described in Section 2.2. (Incidentally, 79 of these records were from Intelius, 122 from BeenVerified, 155 from Zaba Search, where no attempts to match records belonging to the same person were made, and only 14 from Switchboard.) This finding by itself is a privacy risk since it says that *11% of names derived from randomly assembled, relatively unpopular first and last name combinations were indexed by at least one of the people databases.*

The merging process of our algorithm yielded 40 distinct real-world entities, implying that 40 people were found to be listed under at least two people databases. We verified all these results by hand and also by considering Pipl [8], which is an aggregator site that appears to do merging of a nature similar to our work. In all but 11 of these 40 cases, merging records across databases did not yield any new information. Most of these 29 cases were the type where a name was found both in Intelius and BeenVerified and the record in one database was a strict subset of the other. Since Intelius and BeenVerified provide values for identical attributes, no new information was revealed for these 29 individuals, in turn implying that the merger did not increase the privacy risk for these individuals beyond being found in at least one people database.

Our record linkage algorithm revealed significant complementary information about 11 names (see Table 1). These 11 cases either revealed one or more exact addresses a person had stayed in and their phone numbers (often through a merger of Intelius and Zaba Search records), or additional cities a person had lived in (often through a merger of Intelius and BeenVerified records). We describe three of the most interesting cases here. The first was the case of a 40-year old male where an Intelius record (containing his full middle name) indicated that he had lived in 7 different cities in the U.S. Another 11 Zaba Search records containing the same full name gave the exact street addresses (sometimes multiple addresses in a city) for all 7 cities, along with apartment numbers, zip codes and phone numbers. Incidentally, all addresses had the same phone number. The second case was that of a 44-year old male where a BeenVerified record indicated that he had lived in three cities in the U.S. state of Arizona. Another 8 Zaba Search records revealed his exact addresses, including street addresses and zip codes. Further, three of the Zaba Search records also contained three unique phone numbers, with one missing the area code. Finally, the third case was that of a 57 year old male for which Intelius gave one record containing three U.S. cities. Four records in Zaba Search for this individual revealed exact addresses in two of the three cities, along with a phone number.

Data set	Names	Records found	Total linked	Interesting linkages
Data set 1	625	370 (72 names)	40	11
Data set 2	137	5,096 (143 names)	677	216
Data set 3	97	5,492 (184 names)	599	170

Table 1: Overview of searches in people databases and output of the record linkage algorithm

Data Set 2: Popular First Name, Less Popular Last Name: Due to the popularity of first names in this data set, we found many matches for the 625 names in this data set. Since the naive implementation of our algorithm did a pairwise comparison for each record, the running time for this data set appeared prohibitive. Thus, we decided to trim names which yielded more than 60 records in all four databases combined. This left us with 137 first and last name combinations. After removing records containing undesirable variations in either first or last names, we were left with 5,096 records for 143 of these names. As in the case of the first data set, this highlights the privacy risk since this is a large number of randomly assembled first and last name combinations to be indexed by at least one of the four databases we considered.

Subjecting these 5,096 records to our record linkage algorithm yielded 677 matches, implying that 677 real-world entities were found to be listed under at least two databases (see Table 1). For 216 of these, sufficient complementary in-

formation was available to learn their exact postal addresses, phone numbers or at least one other city they had lived in. Specifically, Intelius or BeenVerified records were linked to one or more Zaba Search records in 65 of these cases, implying that one or more street address and phone numbers were learned through correlation. The interesting cases for this data set were similar to those discussed for data set 1, so we mention only one. For a 65-year old woman, our algorithm found a record each in Intelius and BeenVerified. Both records pointed to the same one city in the U.S. state of Oregon this individual had lived in. Zaba Search yielded 13 records, 12 of which pointed to different street addresses and two different zip codes she had lived in. Additionally, we learned that she had held exactly one phone number at all of these addresses. Clearly, this information poses immense privacy risk for this individual since not only can an adversary learn about her present and past postal addresses and her phone number but the linkage reveals other information about her, that she has never lived anywhere else but one city and has kept only one phone number for a very long time.

Data Set 3: Less Popular First Name, Popular Last Name: This is the dual of data set 2, with similar results. Pruning name combinations that yielded more than 60 records across all four databases combined to speed up the running time of our algorithm, we were left with 97 of the 625 names we started with. These 97 names corresponded to 5,492 records (for 184 names) from the four people databases. Subjecting them to the record linkage algorithm produced 599 matches, implying that 599 real-world entities were found to be listed under at least two people databases (see Table 1). For 170 of these individuals, our algorithm helped learn additional information, including full postal addresses and phone numbers or other cities they had lived in. 74 of these correlations were particularly interesting, in that full postal addresses and phone numbers were learned through record linkage. The interesting cases here are similar to those for previous data sets so we omit them for brevity sake.

4. DISCUSSION

Using a simple record linkage algorithm we developed to merge records across different people databases, this preliminary work exposed the linkability of complementary information available freely from various people databases. Advertisers looking to send targeted advertisements can use postal addresses or phone numbers derived through record linkage to send advertisements based on a person's age and residence history. They can also buy additional information about the smaller set of correlated individuals, such as email addresses, education, occupation, criminal records, property information and information about relatives to expand the scope of what is possible through the unpaid information. Further, spammers can avail these options to send spam through email, phone or postal addresses. We hope that our work will bring attention to these issues and in turn spur research in methods individuals and database operators can use to mitigate privacy concerns arising out of linkability.

This preliminary work has opened multiple avenues of future exploration which we plan to undertake. First, we did not test our algorithm on names where both first and last names are popular. That name popularity will play an interesting role in the ability of our algorithm to generate correct linkages is evident from the data sets 2 and 3, which only had popular first or last names but not both and generated more records per name compared to the case when both the first and last name were relatively unpopular (see Table 1). Second, we only tested our record linkage algorithm on four databases. While we chose them upon analyzing several tens of existing people databases, more options exist and may reveal other databases that either use a different back end or provide complementary free information. An exploration of such databases can lead to expanding the scope of linkability our work explored by expanding on the four freely available attributes. A third avenue of exploration is searching people databases on attributes other than {first name, last name} combination. Examples of such attributes include phone numbers, email addresses, occupation etc., which many people databases already support. Doing so may provide new insights into record linkage. Further, our current implementation will miss cases where maiden or current last names for women are searched but the database may contain both current and maiden name or either. Finally, our implementation of the record linkage algorithm was naive and hence slow. A practical application of the privacy threats caused by its execution can only be best explored upon investing effort into making it fast and efficient.

5. REFERENCES

[1] Beenverified: A people search engine. `http://www.beenverified.com`.

[2] Names from the U.S. Census Bureau (1990). `http://names.mongabay.com/most_common_surnames.htm`.

[3] City population (U.S.A.). `http://www.citypopulation.de/USA.html`.

[4] Intelius: A people search engine. `http://www.intelius.com`.

[5] Krishnamurthy, B., and Wills, C. On the leakage of personally identifiable information via online social networks. In *ACM/USENIX Workshop on Online Social Networks (WOSN)* (2009).

[6] Variations of English names. `http://www.behindthename.com/names/usage/english`.

[7] Perito, D., Castelluccia, C., Kaafar, M., and Manils, P. How unique and traceable are usernames? In *Privacy Enhancing Technologies (PETS) Symposium* (2011).

[8] Pipl: The most comprehensive people search on the Web. `http://www.pipl.com`.

[9] Tools to control your online reputation. `http://www.reputation.com/`.

[10] Spokeo: A people search engine. `http://www.spokeo.com`.

[11] Swichboard: A people search engine. `http://www.people.switchboard.com`.

[12] Winkler, W. Matching and record linkage, 1995. `https://www.census.gov/srd/papers/pdf/rr93-8.pdf`.

[13] Wondracek, G., Holz, T., Kirda, E., and Kruegel, C. A practical attack to de-anonymize social network users. In *IEEE Symposium on Security and Privacy (SP)* (2010).

[14] Zaba Search: A people search engine. `http://www.zabasearch.com`.