Malin B. Re-identification of familial database records. To appear in the Proceedings of the 2006 American Medical Informatics Annual Fall Symposium. Washington, DC. 2006.

Re-identification of Familial Database Records

Bradley Malin Ph.D.
Department of Biomedical Informatics, School of Medicine
Vanderbilt University, Nashville, Tennessee

Many genome-based research projects include familial relationships, such as pedigrees, with genomic data records. To protect anonymity when sharing family information, data holders remove, or encode, explicit identifiers (e.g. personal name). In this paper, however, we introduce IdentiFamily, a software program that can link de-identified family relations to named people. The program extracts genealogical knowledge from publicly available records and ascertains the re-identification risk for specific family relations. We find robust genealogies on current populations can be extracted from online sources, such as newspaper obituaries and death records. We evaluate IdentiFamily on real world data for a state's capital city and demonstrate unique identifiability for approximately 70% of the population. IdentiFamily provides organizations with a tool to evaluate the anonymity of pedigrees prior to disclosure and design formal privacy protection techniques.

INTRODUCTION

Medical genetics and bioinformatics are coalescing to support advances in genotype-phenotype relations, gene discovery, and therapeutics development.[1] Until recently, biomedical researchers conducted highly specialized gene hunting expeditions to discover and map the genomic regions that influence clinically-observable disorders. In such ventures, researchers analyzed the clinical records and genomic sequences of large families. The integration of clinical-genomic analyses proved to be quite successful and has facilitated the discovery of many disease-related genes. For example, through these models, the Huntington's Disease gene was mapped in a Venezuelan family. [2]

In today's society, decreasing costs in DNA sequencing technology, digital storage, and data processing enable biomedical researchers to conduct exploratory expeditions with statistical sophistication rivaling their specialized predecessors. As a result, research agendas have broadened and now investigate how genomic variation influences complex traits, such as pharmacokinetic response and clinical severity.

The ability to capture and study mass quantities of detailed biomedical information has made personalized genomics research a reality. Health records systems integrate genomic data and familial information into databases of patient-specific health information.[3] Many organizations want to share or license their collections for various endeavors, from public use to for-profit projects. However, to disclose personal health information, the identities of the individuals to whom the genomic and clinical records correspond must be protected. This is necessary to maintain healthy doctor-patient relationships, as well as satisfy legal requirements, such as the HIPAA Privacy Rule. [4]

A number of privacy protection technologies have been proposed for genomic records. Many systems use de-identification strategies that rely on recoding and encryption of explicit identifiers (e.g. personal name, social security number, etc.). Yet, when family relations are disclosed, such as in the form of a pedigree, the potential exists for a serious privacy violation. This is because family relations reside in various formats external to the de-identified data.

In this paper, we introduce *IdentiFamily*, a software program that automates the re-identification of family relations using publicly available data. *IdentiFamily* illustrates that family relations in shared genomic data pose a privacy threat, but also serves as a tool by which organizations can measure re-identification risk and develop appropriate protection solutions.

BACKGROUND

Biomedical Research Models

Many biomedical research organizations stress the importance of historical and genealogical repositories in genome-based studies.[5-6] They are powerful models for discovering patterns of inheritance, as well as patients that are potentially useful for research projects. Often researchers are removed from the corresponding individuals, so research is performed on de-identified data. For example, in the deCode Genetics Inc. model of data sharing and research subject discovery, encrypted lists of patients and pedigrees are provided to researchers.[6] Their model hashes names into pseudonyms, but alternative models for obscuring patient names are in practice, such as reversible encryption[7-8], one-way hashing, and random value assignment[9-10]. Regardless there remains an important constant: pseudonyms do not obscure the genealogical information that is associated with the records.

Re-identification Research

The manner by which health information is re-identified depends on the type of data available.

Previous research has illustrated a number of data re-identification models for health data. Re-identification occurs when two conditions are satisfied: 1) uniqueness and 2) linkage. The first condition is satisfied when unique values exist in the shared medical records. The second condition is satisfied when attributes in the medical record can be used to link to identified information external to the medical record.

For example, the combination of demographic attributes {*Birth Date*, *Sex*, 5-*Digit Zip Code*}, found in shared de-identified clinical information such as hospital discharge databases, provide a linkage route to publicly available voter registration lists.[11] Experimental analysis showed around 87% of the United State population is uniquely characterized by combinations of values for these attributes. With respect to genomic data, knowledgebases and machine learning methods can relate de-identified genomic sequences to identified clinical records.[12] Also, an individual's health provider-visits, or trail, can be used to link an individual's data.[13] Though DNA re-identification requires uniqueness in genomic data, this is relatively easy to satisfy.[14] It is estimated 100 single nucleotide polymorphisms, features often in DNA research studies, can uniquely represent an individual.

In an earlier paper, we evaluated existing genomic data privacy protection technologies and discovered they were susceptible to various re-identification attacks.[15] We mentioned familial information can be used to compromise privacy, but only at a rudimentary level. In this paper, we detail the attack and present software to automate the process, and confirm its significance with experimental evidence.

METHODS

IdentiFamily is a software program designed to link de-identified pedigrees to named individuals in publicly available information. The program architecture is in Figure 1. In the following sections, we provide intuition into the *IdentiFamily* engine and the privacy attack. The engine is organized into four components: *i*) extract, *ii*) validate, *iii*) structure, and *iv*) link.

Step 1: Extract. First, *IdentiFamily* extracts information on family relations that contains personal names. In many instances, this information is available in publicly available genealogical databases that provide unstructured text or semi-structured documents. Genealogical records are available in many online systems that tailor to ancestry tracing, such as Ancestry.com, Infospace.com, RootsWeb.com, Geneanet.com, FamilySearch.org, and Genealogy.com. From these sites, family structures with named individuals can be extracted. Despite the quantity of information in genealogy-focused sites, the information is mainly reported on earlier generations. In contrast, genomic-based research projects focus on the current generation. Thus, the focus of *IdentiFamily* was shifted to alternative public records. Specifically, *IdentiFamily* extracts personal information from death records reported in online newspapers. For each database crawled, we designed query templates and directed web-crawling strategies to extract records.

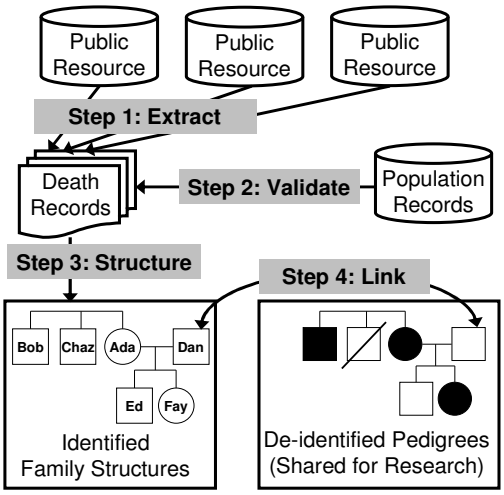

Fig 1. *IdentiFamily* architecture.

Step 2: Validate. Extracted family information that appears unique may not be so when in the context of a larger population. If extracted records do not cover a significant portion of the population to which it corresponds, then it is not unique and claims of subsequent linkage can not be substantiated. Therefore, after we extract the names of the deceased individuals, we cross-reference with information in databases that provide coverage on a population of interest. For *IdentiFamily*, we cross-reference with the Social Security Death Index (SSDI) Database.* The SSDI is administered by the Social Security Administration and provides the names and time of death residence for all people issued a social security number. As with the online newspapers, *IdentiFamily* queries and crawls the SSDI for either a specified population (e.g. city or county) or a specified name.

Step 3: Structure. To construct genealogies from death records, *IdentiFamily* uses template-based extraction.[16] This is done by filling in a template of a genealogy which we call a *family structure*, which is the skeleton of a pedigree. The family structure template is completed by searching a death record for keywords, such as brother, sister, son, daughter, nieces, grandchildren, and then backtrack for the number of each type (e.g. "a", "one", "two", "twenty-two"). We ascertain the gender of the deceased individual by

* The SSDI Database is available at http://ssdi.rootsweb.com.

looking for words, such as "he", "she", "Mrs", "Mr". We also extract precedence information, such as when particular individuals in a family structure survived or prececeded the deceased individual in death.

Step 4: Link. Once *IdentiFamily* constructs a set of family structures, the uniqueness of each is evaluated. Family structures that are sufficiently unique can be used to link named individuals to de-identified pedigrees shared for research purposes.

Family structures are useful for linkage purposes because genealogies are often rich in depth and variation. In theory, family structures can have many configurations and variations. The quantity of family structure configurations enables the re-identification attack. Consider the complexity and number of family structures that one can discern. As a base case, let us begin with the simple family structure of two parents and one child in Figure 2a. There must exist a man and woman to produce a child, so the only variable is the gender of the child. For notation, we use the variable V_i, to represent the ambiguous variable of gender for child i, which can be either male (M) or female (F).

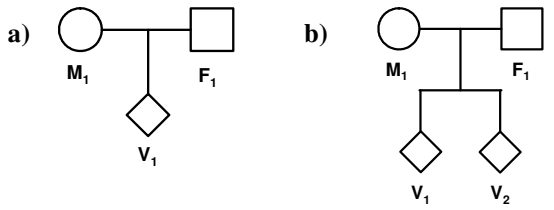

Fig. 2. Family with a) one child and b) two children.

There are 2 variants on the simple nuclear family structure: V_1=M and V_1=F. Similarly, in Figure 2b we depict the case of two children. Here, there are three family structures: V_1=M, V_2=M; V_1=M, V_2=F; and V_1=F, V_2=F. It generalizes that for n variables in the same generation there are n+1 sibling structures.

As we illustrate in our experiments below, information on two generations (parental and child) can be readily extracted from online repositories. As a result, we can find not only how many children the parents have, but also how many brothers and sisters a particular parent has as well. Together, both the parent and child generations comprise an independent combinatoric. For example, a mother could have 4 siblings, as well as have 4 children. In this case, there are 5*5, or 25, possible family structures of this type. And since there are 10 people in each structure, there are 25*10, or 250, individuals that can be re-identified. In general, the maximum number of people that can be re-identified for a structure with n variables in the first generation and m variables in the second generation is $(n + 1) * (m + 1) * (n + m)$. Thus, if we consider all of the two-generation family structures up to 5 siblings in each generation, there are on the order of 10^5 people that can be uniquely re-identified. In reality, we find pedigrees are more robust than simple nuclear families, which increases the number of family structures. For example, a three generation family of two children per family permits on the order of 10^6 distinct family structures and 10^7 individuals that could be uniquely characterized. It is a larger number of combinations when supplementary information, such as living status or ordering siblings by age, is provided.

Not all family structures will be realized in a population, and certain variants are more probable than others. Nonetheless, the magnitudes make for daunting statistics. Moreover, the number of family disease structures is larger still, considering that this analysis does not account for additional features, such as the whether or not certain family members are deceased. This is yet another feature that can be communicated in both pedigrees and family structures. Additionally, supplementary medical information may be known.[6] In some of the publicly available information we studied, it is stated how an individual died, such as "after a long battle with breast cancer". Clinical-related information is of especial concern because many polygenic trait studies are interested in learning which factors are the most influential in disease severity or occurrence.

RESULTS

There are many websites that post obituaries and death notices. Often, the sites use a standard format url for database backed queries. We manually queried websites with obituaries and generated templates and scripts for *IdentiFamily* to automate the extraction of obituaries from databases within a certain locale.

We found online newspapers are abundant and provide an excellent alternative resource. For instance, Legacy.com provides the website and database backing for death records databases in over 250 newspapers in the United States. Unfortunately, some newspapers charge a fee to post a death record, such as the *Detroit Free Press* which costs $12 per record. By charging to post a record, the reported population will be a biased sample. Thus, we focused on a local newspaper that does not charge for death record postings, the Wyoming Tribune-Eagle (WTE).

IdentiFamily downloaded the death notices from the WTE for the 2000-2005 period.[†] *IdentiFamily* extracted the names of the deceased for 1007, 925, 1015, 997, 1031, and 1033 records for the years 2000 through 2005, respectively. In each death record, names were reported for the deceased, and often for siblings, children, and parents.

Population Coverage. To test if the list of extracted names provided coverage of the local population, *IdentiFamily* cross-referenced the death

[†] http://www.wyomingnews.com/news/obits/.

records with the Social Security Death Index (SSDI) Database. One of the challenges of dealing with obituaries from newspapers is the posted records can correspond to individuals that lived in the local region at one time, but are no longer reside there. Such individuals are not considered to be a part of the local region according the SSDI, since it was not their residence at the time of death.

Year	Death Records	SSDI Records	Underreported
2000	513	469	8.5%
2001	521	513	1.5%
2002	501	487	2.68%
2003	547	513	6.2%
2004	492	481	2.2%
2005	576	553	4.0%

Table 1: Cheyenne population coverage.

Thus, we used *IdentiFamily* to further extract obituaries on a well-defined population. We concentrated on where the majority of the records (approximately half) were submitted from, which was Cheyenne, Wyoming. The number of extracted death records from Cheyenne per year is shown in Table 2. The number of records extracted from the SSDI for Cheyenne is also shown in this table.

Interestingly, we found that it is the SSDI which underestimates the population in a region and not the WTE death records. Per year, the death records had almost 100% coverage of the SSDI records for Cheyenne. In contrast, over the time period studied, the SSDI underreported by an average 4.18%, with a standard deviation of 2.69%. Underreporting per year is shown in Table 1. Therefore, we can conclude the death records correspond to most of the deceased individuals in Cheyenne during this period. This is due to the fact that the SSDI does not contain the names of all deceased. For instance, the underreporting can derive from people who die without being issued a social security number. In fact, upon manual inspection, we discovered that almost all of the underreported deaths in the Cheyenne dataset corresponded to infants

Family Structure Uniqueness. Next, we analyzed the two-generation family structures *IdentiFamily* extracted from the death records. As mentioned above, the two generations correspond to the parental siblings (first generation) and the children (second generation) of the deceased individual. From the 2581 obituaries, *IdentiFamily* extracted approximately 12,000 named individuals.

There were approximately 1500 distinct family structures extracted from the death records. The family structures were organized by frequency and the result is shown in Figure 3. We discovered that around 780 family structures were unique. Thus, 30.1% of the 2581 death records correspond to unique structures. Moreover, almost half of all structures were had a frequency of two or less.

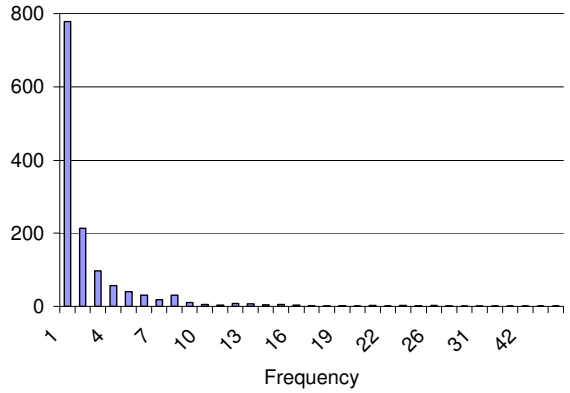

Figure 3. Uniqueness of two-generation family structures.

Next, we looked at how many individuals would be re-identified given the extracted family structures. This result as a cumulative function is depicted in Figure 4. Here, we found a logarithmic growth curve, such that almost 8300 people, or 67% of the population, were uniquely re-identifiable. And by a frequency of two or less, almost 83% of the population was identified. In other words, 83% of the population resided in a family structure that characterized two or less death records.

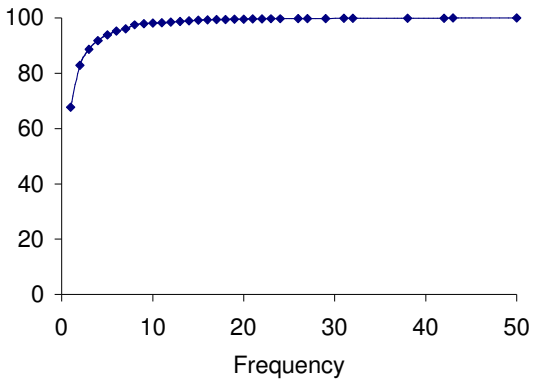

Figure 4. Cumulative identifiability of extracted population.

Our analysis was performed with respect to a two-generation family structure; however, the structures can be made more robust given available data. For many of the extracted death records, names were found for many other relations, such as grandparents, nieces, nephews, grandchildren, and great-grandchildren. Often, when the names of these relatives were not provided, the number of each relation (e.g. "survived by eleven grandchildren") was often reported. As a consequence, the identifiability statistics that we present are an underestimate of the total number of

family structures, as well as the total number of people that can be re-identified.

DISCUSSION AND CONCLUSIONS

Publicly available records, such as death records, are not going to be removed from the World Wide Web. Thus, it is necessary to develop protection methods for genealogical records with provable guarantees. Robust family structures, with the names of individuals, exist in publicly available resources and simple structures of two generations are often uniquely identifying. *IdentiFamily* illustrates that *ad hoc* de-identification schemes, such as name removal or encoding, are not sufficient to protect privacy against and automated strategies can lead to significant privacy breaches. At the same time, the *IdentiFamily* program provides administrators and organizations with the ability to determine the re-identification risk associated with disclosed information. Nonetheless, *IdentiFamily* does have limitations, several of which we briefly discuss.

One of the drawbacks to *IdentiFamily* is its dependence on the Social Security Death Index (SSDI) database. The SSDI does not contain the names of all deceased. The underreporting is not necessarily due to errors in the SSDI. For instance, the underreporting occurs when people without social security number expire. Furthermore, in addition to excluding individuals without social security numbers, such as infants, the completeness in the SSDI is dependent on the actions of the deceased relatives and funeral homes. If the relatives, or a funeral home, fail to report an individual's death to the Social Security Administration, then the individual's record will not appear in the death index. In future research we intend on limiting *IdentiFamily*'s reliance on the SSDI. This may be achieved through alternative sources of public records, such as through census accountings, or other online repositories.

A second limitation to our research is that each family structure extracted from a death record is considered to be independent. However, multiple individuals from the same family can have death records in the same population. Recent research has shown that an individual can be traced over multiple public records databases, including death records, birth records, and marriage records.[17] We anticipate integrating these models into *IdentiFamily* to remove redundancies and construct more robust family structures.

Acknowledgements

This research was conducted while the author was at the School of Computer Science, Carnegie Mellon University. The author thanks Latanya Sweeney and Edoardo Airoldi for insightful discussions. This research was funded in part by NSF IGERT grant 9972762 in CASOS.

References

1. Ginsburg G, Haga S. Translating genomic biomarkers into clinically useful diagnostics. *Expert Rev Mol Diagn*. 2006; 6(2): 179-91.
2. Gusella J et al. DNA markers for nervous system diseases. *Science*. 1984; 225: 1320-1326.
3. Sax U, Schmidt S. Integration of genomic data in electronic health records: opportunities and dilemmas. *Meth Info Med*. 2005; 44 : 546-550.
4. Federal Register, 45 CFR, 160-164. Standards for privacy of individually identifiable health information, Final Rule. Aug 12, 2002.
5. Gaudet D et al. Procedure to protect confidentiality of familial data in community genetics and genomic research. *Clin Genet*. 1999; 55: 259-264.
6. Gulcher J et al. Protection of privacy by third-party encryption in genetic research. *Eur J Hum Genet*. 2000; 8: 739-742.
7. Ferris T et al. A proposed key escrow system for secure patient information disclosure in biomedical research databases. *Proc AMIA Symp*. 2002: 245-249.
8. de Moor G, et al. Privacy enhancing techniques. *Meth Info Med*. 2003; 42: 148-153.
9. Kruse R, Ewigmen B, Tremblay G. The zipper: a method for using personal identifiers to link data while preserving confidentiality. *Child Abuse & Neglect*. 2001; 25: 1241-1248.
10. Hara K et al. Establishment of a method of anonymization of DNA samples in genetic research. *J Hum Genet*. 2003; 48: 327-330.
11. Sweeney L. Uniqueness of simple demographics in the US population. *Data Privacy Laboratory Working Paper LIDAP*-4, CMU. 2000.
12. Malin B, Sweeney L. Determining the identifiability of DNA database entries. *Proc AMIA Symp*. 2000: 547-551.
13. Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network. *J Biomed Info*. 2004; 37(3): 179-192.
14. Lin Z, Owen A, Altman R. Genomic research and human subject privacy. *Science*. 2004; 305: 183.
15. Malin B. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *JAMIA*. 2005; 12(1):28-34.
16. Mann G, Yarowsky D. Unsupervised personal name disambiguation. *Proc CCNL*. 2003: 33-40.
17. Griffith V, Jakobsson M. Messin' with Texas: extracting mother's maiden name using public records. *Proc Applied Cryptography and Network Security Conference*. 2005.