Messin' with Texas
Deriving Mother's Maiden Names Using Public Records

Virgil Griffith, Markus Jakobsson

April 18, 2005

Abstract

We have developed techniques to automatically infer mother's maiden names from public records. We demonstrate our techniques using publicly available records from the state of Texas, and reduce the entropy of a mother's maiden name from an average of close to 13 bits down to below 6.9 bits for more than a quarter of the people targeted, and down to a zero entropy (i.e., certainty of their mothers maiden name) for a large number of targeted individuals. This poses a significant risk not only to individuals whose mothers maiden name can easily be guessed, but highlights the vulnerability of the system as such, given the traditional reliance of authentication by mother maiden names for financial services. While our techniques and approach are novel, it is important to note that these techniques – once understood – do not require any insider information or particular skills to implement. This emphasizes the need to move away from mothers maiden names as an authenticator. Using the techniques described, during testing we were able to deduce the mother's maiden name for approximately 4,105,111 Texans.

1 Introduction

Within the security community the secrecy of your mother's maiden name (MMN) is known to not to be the strongest form of authentication. However, the MMN is frequently used by the commercial sector including banks, credit cards agencies, internet service providers, and many websites. This may be largely for convenience, but by and large the MMN is considered to be suitably secure against all but the most targeted attacks or those by

close family friends. However, our study shows that by mining and cross-correlating public records information (which is required by US law to be public), an attacker can determine or "compute" MMNs with startling accuracy. Utilizing large numbers of identities for the purposes of laundering is an immense asset to both terrorist organizations and other more traditional organized crime.

The ubiquity of birth and marriage information constitutes the most direct threat of MNN compromise by means of public records. Marriage records are a reliable way of obtaining large numbers of maiden names, while birth records provide the identities of offspring. By using them in conjunction, all that remains for a successful compromise is linking a child to the appropriate parents, and then printing the bride's maiden name as listed within the marriage record. The cross-correlation of birth and marriage data is not only effective as a general approach to MMN compromise, but also has numerous non-obvious special cases that make MMN derivation alarmingly easy. For example, if a groom has a very uncommon last name, then it becomes very easy to match him with any of his children simply by their uncommon last name. Secondly, if the birth record denotes that the child is suffixed "Jr.", "III", etc., an attacker can drastically narrow down the number of candidate parents. Third, if the child's last name is hyphenated, rarely will an attacker have any trouble matching the child with the appropriate marriage. While these special cases make up only a relatively small portion of the population, as we increase in scale, even the smallest tricks and statistical regularities will result in thousands of compromises. Moreover, for every victim that an attacker succeeds to infer the MMN, he narrows the number of choices for other potential victims (except for siblings). The ability to deduce secret information from supposedly innocuous information has been discussed previously [8]. However, we are not aware of any previous instances of deduction of personal authenticating information on this scale. Although no extensive survey has been done, the use of mother's maiden name as a security authenticator seems to be a practice unique to Canada and the United States. Other countries (particularly in Europe) use better security practices such as one-time use random numbers or insisting that they call customers back at their registered phone number.

The availability and exact information contained within birth and marriage records varies slightly from state to state. So, for purposes of illustration, we decided to focus on only one. Naturally, we wanted as large a sample size as possible to ensure that our methods scaled well to very large datasets, but also to assure that any conclusions pertaining to the sample would be worthy of attention in their own right. This left us with two promi-

nent choices for in-depth analysis: California and Texas. The most recent US Census [4] indicates that Texas is substantially more representative of the entire country than California. Particularly, the ethnic composition of Texas is closer to that of the nation than California. This is of special relevance considering that marriage patterns as well as last names (and therefore maiden names) are strongly influenced by ethnicity. Texas is also more representative in the percentage of foreign-born residents, and the frequency of households moving to other states. Overall, this made Texas a natural choice for our studies. It should be clear that although we chose Texas because of its statistical proximity to the national averages, these same techniques can be used to derive MMNs in other states (especially large states with digitized records) with success rates likely on the same order as our findings. California has also made their records digitally available; we anticipate very similar results to those presented here.

Although these techniques are to the best of our knowledge completely novel, now that they've been discovered the replication and application of them can be done by anyone with Internet access. However, we do not believe that the publication of this information is immoral, and rather see it as a necessary alert of a problem bound to occur no matter what.

2 Availability of Texas Marriage, Birth, and Death Information

In smaller states, vital information is usually held by the individual counties in which the events took place, and in larger states there is an additional copy provided to a central state office. Texas is no exception to this pattern. Yet, regardless of where the physical records happen to be stored, all such records remain public property and are with few exceptions fully accessible to the public. The only relevance of where the records are stored is that of ease of access. State-wide agencies are more likely have the resources to put the information into searchable digital formats, whereas records from smaller local counties may only be available on microfilm (which they will gladly ship to you for a modest fee). However, as time progresses, public information stored at even the smallest county offices will invariably start being digitized.

The Texas Bureau of Vital Statistics website [16] lists all marriages state-wide from 1966—2002; records from before 1966 are available from the individual counties. Texas birth records are also available online but the *fields containing the names of the mother and father* are "aged" for 50 years

(meaning they are withheld from the public until 50 years have passed). This means that for anyone born in Texas who is over 50, a parent-child linking has conveniently already been done. It may seem obvious if we think about it, but it's worth mentioning that the average American lives well beyond the age of 50, making this security measure insufficient. By means of this policy alone, every single person born in Texas that was born from 1923–1949 currently has their MMN completely compromised in plaintext. From these records alone we are able to fully compromise 1,114,680 males. Females are somewhat more difficult because if they have been married we would not know their current last name. However, our marriage records from 1966–2002 contain the age of both the groom and bride, by matching brides not only by name but also by year of birth, we were able to compromise 288,751 women (27%). In many cases older people make much better targets for fraud as they are likely to have more savings than younger adults.

Here it is worth noting that in October 2000, Texas officially took down the online access to their birth indexes (death indexes were similarly taken down as of June 2002 [3]) due to concerns of adopted children discovering the identities of their biological parents [2] (which is illegal). Additionally, they increased the aging requirement for both the partially redacted and full birth records to 75 years, and even then will only provide birth and death records in microfiche. However, before they were taken down partial copies of the state and county indexes had already been mirrored elsewhere where we were able to find and make use of them. We found two sizable mirrors of the birth and death information. One was from Brewster Kahle's famous *Wayback Machine* [1], and the other from the user-contributed grass-roots genealogy site Rootsweb.com [11] which had a even larger compilation of partial indexes from the state and county level. Oddly, despite these new state-level restrictions, county records apparently do not require aging and many county level birth and death records all the way up to the present remain freely available in microfilm or through their websites [13]. Of particular amusement, even though the death indexes available on Rootsweb and the Internet Archive were put up before they were supposedly taken down in June 2002, the full death indexes are still available (although not directly linked) over 2 1/2 years later from the Texas Dept. of Vital Statistic's own servers at *exactly the same URL they were at before* [19]! All of this is particularly relevant because even though Texas is now doing a better job protecting their public records (although largely for unrelated reasons), the public is just as vulnerable as they were before.

3 Heuristics for MMN Discovery through Marriage Records

We have already described how a cursory glance over birth and marriage records reveals a more than ample supply of low-hanging fruit. However, if this is not enough to persuade the discontinuation of MMN-based authentication, the correlation of marriage data (perhaps the best source of MMNs) with other types of public information comprises an effective and more general approach to linking someone to his or her mother's maiden name. When given a list of random people whether it be produced by partially redacted birth records, phonebooks, or your favorite social networking service, there are at least seven general observations that an attacker could use to derive someone's MMN with high probability. Naturally, as each heuristic is applied, the chance of MMN compromise will be increased.

1. We do not have to link a child to a particular marriage record, only to a particular maiden name. There will often be cases in which there are repetitions in the list of possible maiden names. This holds particularly true for ethnic groups with characteristic last names. An attacker does not have to pick the correct parents, just the correct MMN! This observation alone makes guessing MMNs much simpler than one might think.

2. Children will generally have the same last name as their parents.

3. Couples will typically have a child within the first five years of being married.

4. Children are often born in the same county in which their parents were recently married.

5. Parts of the parents' first, last, and middle names are often repeated within a child's first or middle name. (Conveniently, this is especially true for the mother's maiden name and the child's middle name.)

6. Children are rarely born after their parents have been divorced. In addition to this rule, all Texas divorce records [18] list the number of children under 18 bequeathed within the now dissolved marriage. So, divorce records are helpful not only by eliminating the likelihood of children being born to a couple beyond a divorce date, but they also tell us how many children (if any) we should expect to find, as well as the general birth range to expect for them. In Texas, every divorce

affects on average 0.79 children [17]. As nation-wide divorce rates average about half that of marriage rates, divorce data can significantly complement any analysis of marriage or birth records.

7. Children cannot be born after the mother's death nor more than a year after the father's death. Texas death indexes are aged 25 years before release (full state-wide indexes for 1964–1975 are available online [19]). Death records are useful in that they not only contain the full name (First/Last/Middle/Suffix) of the deceased, but also the full name of any spouse. This seemingly innocuous piece of information is useful for easily matching up deaths of husbands and wives to their marriages, thus narrowing the list of possible marriages that can still produce offspring by the time of a victim's birth.

For our preliminary statistics, we have taken into account observations 1, 2, 3, and 4. The heuristics listed above certainly are not the only viable attacks an attacker could use, but they serve as a good starting point for the automated derivation of MMNs.

4 Experimental Design

With easy access to public records and no easy way to put the cat back in the bag, we should now be asking ourselves, "How effective are the above described attacks/heuristics in leading to further MMN compromise?", and "What percent of the population is at risk?" To answer these questions, we will use data entropy to measure the risk of MMN discovery from our attacks. Comparing the entropy of different sets of potential MMNs is a suitable and illustrative measurement for accessing the vulnerability to these attacks. Data entropy measures the amount of unpredictability within a distribution of potential MMNs. Its primary benefit over simply listing the number of possible marriage records after filtering is that entropy takes into account repetitions within the set of possible MMNs. For example, after correlating records you could have a set of 40 possible marriages from which the child could have come from. However, 30 of these marriages may have the maiden name "Martinez", and 5 of the remaining 10 marriages the maiden name "Lopez." Clearly, in this case there is a far greater than a 2.5% chance (1/40) of correctly guessing the MMN. (In this example, the entropy would be 1.351 bits.)

To provide a baseline comparison for assessing the increased vulnerability due to accessing attacks using public records, we calculated the data entropy

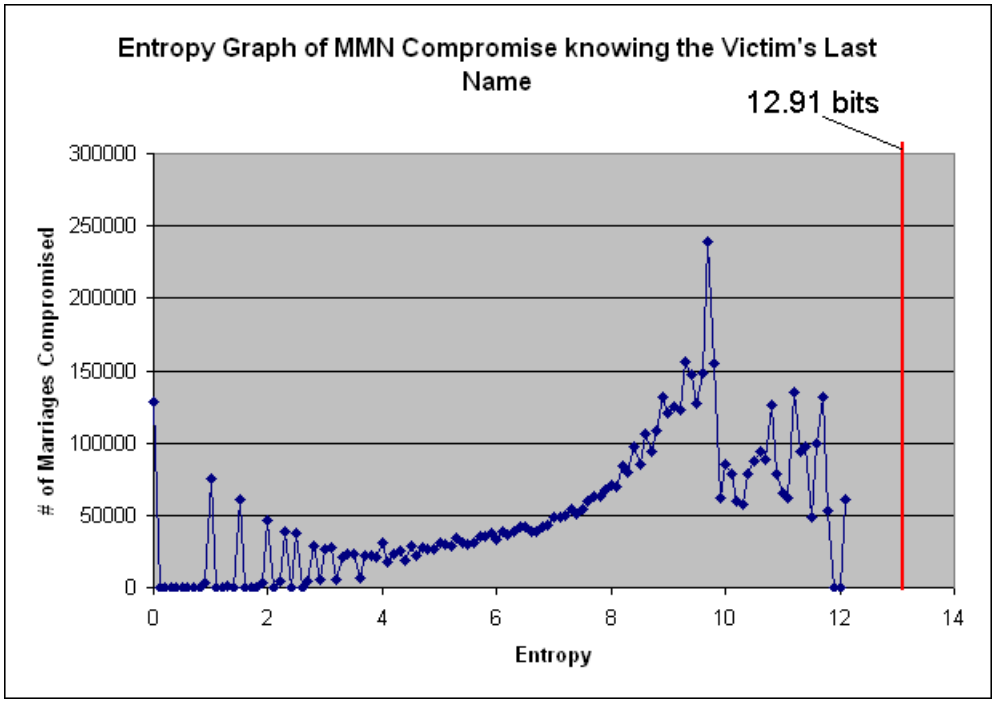

Figure 1: Ability to determine MMNs from knowing the victim's last name

across all maiden names in our database (1966–2002). This measurement is equivalent to an example situation in which an attacker does not use any public records data, but despite this tries calling a local bank and saying the target's MMN is "Smith", just hoping to guess correctly. ("Smith" is the most common last name in America consisting of about 0.9% of the population). The resulting baseline entropy for this attack is 12.92 bits.

5 Analysis of MMN Discovery in Marriage Records

By our methods, we get the following graph (Fig. 1) gauging the risk of MMN compromise from an attacker who makes use of marriage data and makes the assumption that the parents' marriage took place anytime from 1966 to 2002, but who knows nothing more than the victim's last name (i.e., has no knowledge of the victim's age, first or middle name, place of birth, etc.).

Unlike the entropy corresponding to a pure guess, public records allow

the attacker to take advantage of the fact that we know the victim's last name (something the attacker would have to know anyway). Therefore, we will have different entropies, one for each last name. Naturally, deriving someone's MMNs based solely on the their last name will be more difficult for common last names than for uncommon last names given the larger pool of possible parents.

For example, if the attacker only knows the intended victim's last name is "Smith" (resulting entropy = 12.18 bits), this reduces the entropy only 0.74 bits from the original 12.91 bits. However, if it is a less common last name like "EVANGELISTA" (resulting entropy = 5.08 bits), or "AADNESEN" (resulting entropy = 0 bits), the attacker is immensely increasing the chances of correcting guessing the MMN. Note that for the absolute worst cases like "Smith" (12.18 bits) or "Garcia" (9.811 bits), these entropies will still be too high to compromise their bank accounts over the phone. However, these numbers quickly fall into the range of making brute-force an increasingly viable option for gaining assess to their web accounts. Moreover, knowledge of the victim beyond his or her last name(such as age, place of birth, etc.) can help the attacker eliminate large pools of candidate parents, and therefore improve the chances of determining the MMN. In summary, the use of public records to inform the most trivial search for vulnerable MMNs statistically increases the risk for everyone while enabling complete MMN compromise for children with the rarest of names. To allow effective comparison of different attacks, we will redraw Fig. 1 as a cumulative percentage of marriage records compromised.

Table 1: Using the unusual last names attack against our local birth records

Entropy	# Children Compromised	% Birth Records Compromised	Chance to Guess MMN
= 0 bits	82,272	1.04	= 1/1
≤ 1 bit	148,367	1.88	≤ 1/2
≤ 2 bits	251,568	3.19	≤ 1/4
≤ 3 bits	397,457	5.04	≤ 1/8

A full zero-entropy compromise of approximately 2% of marriages may not initially seem so terrible, but the table above shows that even the smallest percentages will lead to massive compromise. The graph above is an accurate assessment of the risk of MMN compromise to an attacker armed with marriage records and Google phonebook [6].

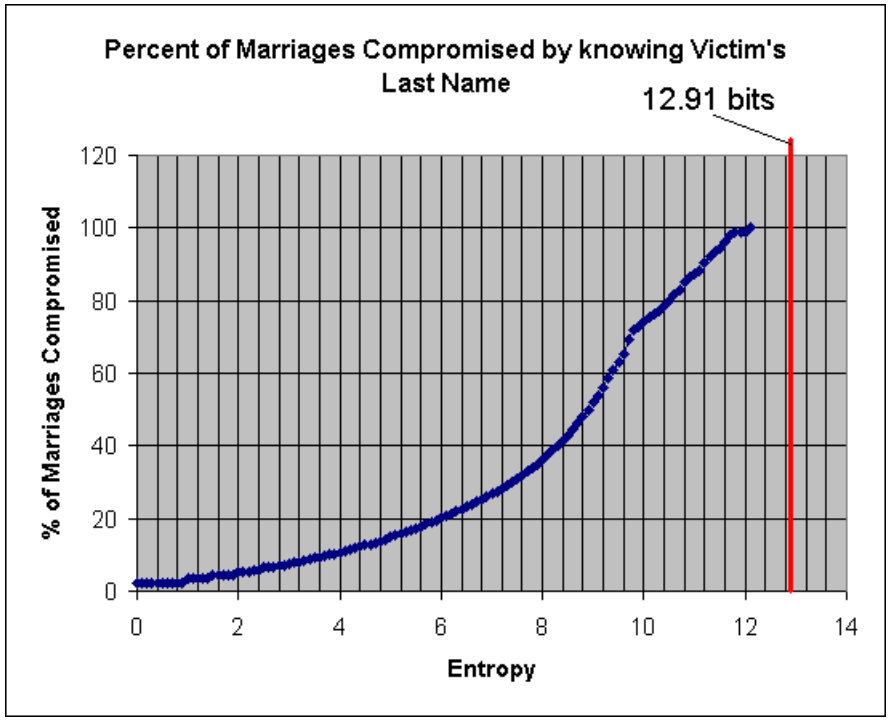

Figure 2: Redrawing of Fig. 1 as cumulative percentage of marriages compromised

5.1 MMN Compromise looking within Five Years and County of Victim's Birth

Although the first attack is the safest route to MMN compromise, in efforts to gain a greater yield there are times in which an attacker would be willing apply further assumptions, such as by creating a "window" of time in which it is reasonable to assume the victim's parents were married. This window of time could be as long or as short as the attacker desires. Naturally, longer windows increase the chances of including the parents' marriage record, while shorter windows yield higher percentages of compromised MMNs. In this example we assume the attacker knows not only the victim's last name, but his or her age (this information can be obtained from birth records or online social networks), and the county in which the victim was born (can be obtained from birth records). This attack uses a five year window up to and including the year the victim was born and deduces MMNs in accordance with the observation that couples frequently have a child within the first five

years of being married. Naming statistics do vary from year to year, but for the reader's convenience we have averaged all years.

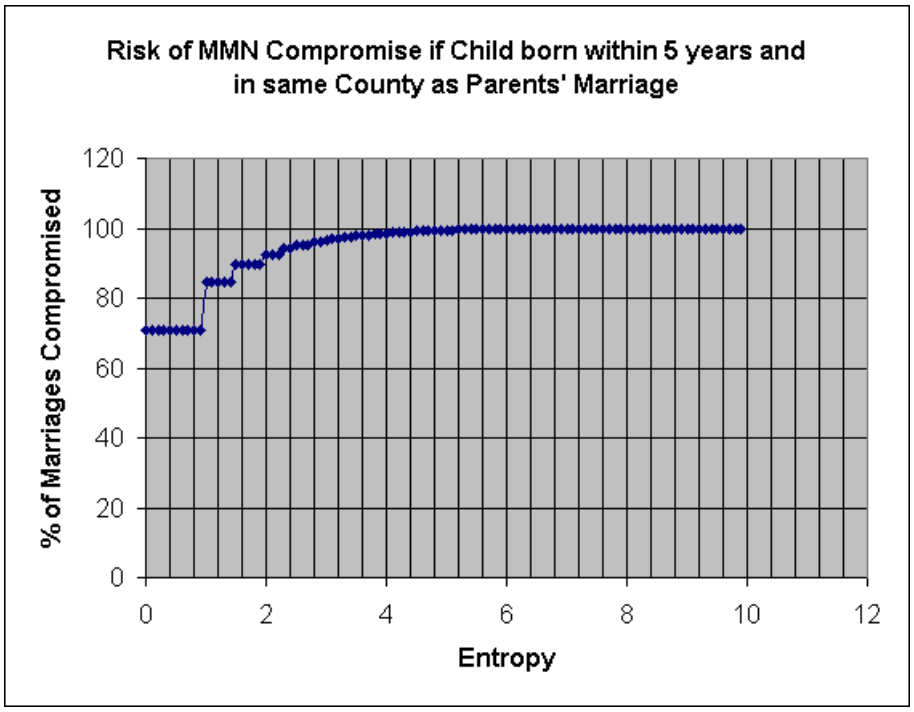

Figure 3: 1970-2002: Risk of MMN compromise when parents' marriage county is known, and the marriage year known within five years

Table 2: Compromises from five year window + county attack against within local birth records

Entropy	# Children Compromised	% Birth Records Compromised	Chance to Guess MMN
= 0 bits	2,355,828	29.8	= 1/1
≤ 1 bit	3,750,798	47.5	≤ 1/2
≤ 2 bits	3,750,798	47.5	≤ 1/4
≤ 3 bits	3,750,798	47.5	≤ 1/8

By narrowing our window in which to look for candidate marriages, the resulting entropies drop significantly. An attacker can increase or decrease the window size based upon the uncertainty of the marriage year. As the window increases, there are fewer zero-entropy compromises, but any compromises are more reliable as there is a better chance of the correct marriage record being included within the window.

5.2 MMN Compromise in Suffix'ed Children

Our final quantitative analysis is for an attack using public records in which the attacker has no knowledge of the victim's age but instead knows the victim's first name, last name, and suffix. Knowing that the victim has a suffix is immensely valuable as it tells the first name to look for within the parents' marriage record. Once again, naming statistics do vary from year to year, but for the reader's convenience we are printing only the average across all years.

Table 3: Compromises of suffix'ed children in our local birth records

Entropy	# Children Compromised	% Birth Records Compromised	Chance to Guess MMN
= 0 bits	344,463	60.5	= 1/1
≤ 1 bit	345,211	60.6	≤ 1/2
≤ 2 bits	345,223	60.6	≤ 1/4
≤ 3 bits	345,223	60.6	≤ 1/8

6 Other Means for Deriving Mother's Maiden Names

Hereto we had focused on the use of birth and marriage records in compromising MMNs, and although birth and marriage information constitute the

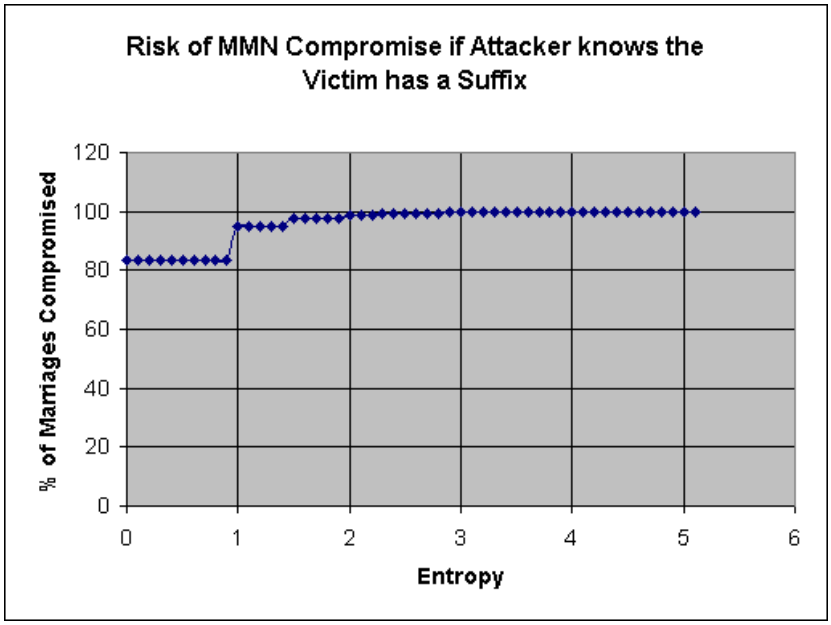

Figure 4: Ability to determine MMN's for suffix'ed children

greatest threat to large scale MMN discovery, it is by no means the only viable route. The following is a sample of some of the more creative public records attacks that we have confirmed to work in our sample cases, yet remain largely unexplored.

6.1 Social Security Death Index

The Social Security Death Index (SSDI) [14] provides up-to-date information of people who have passed away. The SSDI was created as a security measure to prevent the mafia from selling the identities of deceased infants to illegal immigrants. As such, it is comprehensive, digitally available, and fully searchable. In the case of Texas, the SSDI can be used to verify the connection between a groom's death and marriage record. The state death record provides the full name of the deceased person and his or her spouse, but there is still always the possibility for name overlap, particularly as you increase in scale. By taking information from the Texas state death index information and plugging the it into the SSDI, we are able to learn the groom's date of birth, a fact that was unknowable from the state records alone. By knowing the groom's date of birth, an attacker is able to strongly verify the

connection to a particular marriage as the marriage record contains the bride and groom's age. This is a reminder of the ability of different records to "interlock" (also called database aggregation) which allows for much stronger conclusions.

6.2 Voter Registration Records

In efforts to prevent voter fraud (a concern especially of late) voter registration records are by U.S. law [9] required to be public. But despite the good intentions, next to marriage and birth information, voter information constitutes the greatest threat to automated MMN discovery and can perhaps fill in the place of either a birth of marriage record. They contain the Full Name, "previous name" (the maiden name), date of birth, and county of residence [20]. Texas voting records for individual counties are sometimes available from the county websites, but for any significant coverage an attacker would have to purchase them from the state bureau. The database for voter registration records across the entire state costs approximately $1,100. As of 2000, 69% of voting-age Texans were registered to vote; this percentage has almost certainly increased since then due to efforts to "get-out-the-vote" during the 2004 elections.

6.3 Genealogy Websites

Not only a source for mirrored public records data, Rootsweb [12] is an all-purpose user-contributed genealogy website. Amazingly, more often than not, MMNs of currently living people can be read directly from the submitted family trees with no further analysis required for successful MMN compromise. In the off-chance that a security conscious genealogy researcher lists a mother under her husband's name, her first name, middle name, marriage date, date and place of birth are always given. With this much information already in hand, a marriage or birth record will allow for certain recovery of the maiden name. Online user-contributed family trees currently do not cover a large fraction of the population, but the submitted trees are still a complete map for MMN compromise and are available to anyone with Internet access. In our analysis we found Rootsweb.com to contain full family trees for 4,499 living Texans. Some genealogy resources such as the Church of Later-day Saints' FamilySearch.org avoids listing information about living people.

6.4 Newspaper Obituaries

Local newspapers frequently publish, both in print and online, obituaries of those who have recently died. Regardless of whether these obituaries happen to be analyzed by hand or via some clever natural language analysis, an obituary entry will generally give an attacker the deceased's name, date of birth, name of spouse, as well as the names of any children. The recently deceased is of no interest to an attacker, but the recent departure of a parent is a convenient opportunity for attacking any children. With the information contained in an obituary, the maiden name can be gotten easily from either the marriage or voting record. However, the children may have moved to other parts of the country, so simply looking them up in the local phonebook may not work. However, an attacker can look up the deceased's SSDI entry which lists a "zipcode of primary benefactor," which will almost invariably be the zipcode of one of the children. The combination of a name and zipcode is a surprisingly unique identifier and the location of the child can be easily queried using Google Phonebook [7].

6.5 Property Records

At our current scale property records are of relatively little value. But if we wanted to expand these techniques to a national scale, property records are a good option for tracking people who have moved to another state. Property records are required by law to be public and are usually freely available online [15]. In the absence of property records, mass aggregation of phonebooks from different years is also a viable option.

7 Conclusion

Just as cryptographic algorithms do with time require to be replaced, so do authentication mechanisms. Unlike the time when mother's maiden names first started being used as an authenticator, our analysis shows the MMN is increasingly vulnerable to the automated data-mining of public records. New data-mining attacks make it increasingly unacceptable to use documented facts as authenticators. Facts about the world are not true secrets. As a society, there are many ways to respond to this new threat. Texas' response to this threat was by legislating away easy and timely access to its public information. This approach has been largely ineffective, and has accomplished exceedingly little in diminishing the threat of MMN compromise to the public at large. If these actions have accomplished anything of sig-

nificance, it is only the creation of a false sense of security. Access to public records of all types was created to strengthen government accountability and reduce the risk of government misconduct by allowing the public to watch over the government that it supports with its tax money. Some states find this governmental oversight so vital as to even write public records provisions directly into their state constitution [5]. We can only speculate as to the long term effects of policies which would routinely restrict access to otherwise valuable public information simply because it might also be valuable to those with less-than-noble intentions.

In today's society, the existence of a seperate mother's maiden name, much less a secret one, is slowly becoming obsolete. At one time, the mother's maiden name served as a convenient and reasonably secure piece of information. However, as sociological changes have made it increasingly socially permissible for a woman to keep her original name and have additionally brought about hyphenated names for children, new technologies have made for comprehensive and accurate record keeping as well as easy searching of these records. Using one of our methods (but expanding our search beyond the state of Texas), we established that the mother's maiden name of the current president of the United States is "Pierce," and the mother's maiden name of his two children is "Welch."

8 Acknowledgements

The primary author wishes to thank Henry Strickland for his suggestions on the entropy graph presentations.

References

[1] Archive.org 21-Jun-2001: Bureau of Vital Statistics General and Summary Birth Indexes
http://web.archive.org/web/20000621143352/http://www.tdh.state.tx.us/bvs/registra/birthidx/birthidx.htm

[2] Archive.org 20-Nov-2001: Bureau of Vital Statistics, General and Summary Birth Indexes
http://web.archive.org/web/20001120125700/http://www.tdh.state.tx.us/bvs/registra/birthidx/birthidx.htm

[3] Archive.org Birth/Death Index mainpages for 19-Nov-2001 and 05-Jun-2002

Comparing http://web.archive.org/web/20011119121739/http://www.tdh.state.tx.us/bvs/registra/bdindx
to http://web.archive.org/web/20020605235939/http://www.tdh.state.tx.us/bvs/registra/bdindx.htm

[4] Census 2000 Briefs www.census.gov/population/www/cen2000/briefs.html

[5] Florida State Constitution, Section 24.
http://www.flsenate.gov/Statutes/index.cfm?
Mode=Constitution&Submenu=3&Tab=statutes#A01S24

[6] Google Phonebook Search
http://www.google.com/search?hl=en&q=phonebook%3A&btnG=Google+Search

[7] Google Phonebook Search for "Smith" in zipcode 75201 (Dallas,TX)
http://www.google.com/search?sa=X&oi=rwp&pb=r&q=Smith+75201

[8] Sweeney, Latanya; Malin, Bradley: Journal of Biomedical Informatics. 2004; 37(3): 179-192 How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems.

[9] National Voter Act of 1993 http://www.fvap.gov/laws/nvralaw.html

[10] Texas State Property Records http://www.txcountydata.com

[11] Rootsweb.com FTP server with complete copies of both the marriage and death indexes
ftp://rootsweb.com/pub/usgenweb/tx/

[12] RootsWeb.com Home Page http://www.rootsweb.com

[13] SearchSystems.net listing of Texas Counties' online public record offerings
http://searchsystems.net/list.php?nid=197 http://searchsystems.net/list.php?nid=344

[14] Social Security Death Index http://ssdi.genealogy.rootsweb.com/

[15] Texas State Property Records http://www.txcountydata.com

[16] Texas Department of Health, Bureau of Vital Statistics, Marriage Indexes
http://www.tdh.state.tx.us/bvs/registra/marridx/marridx.htm

[17] Texas Department of Health, Divorce Trends in Texas, 1970 to 1999
www.tdh.state.tx.us/bvs/reports/divorce/divorce.htm

[18] Texas Department of Health, Bureau of Vital Statistics, Divorce Indexes
http://www.tdh.state.tx.us/bvs/registra/dividx/dividx.htm

[19] Texas Department of Health, Bureau of Vital Statistics, General and Summary Death Indexes
http://www.tdh.state.tx.us/bvs/registra/deathidx/deathidx.htm

[20] TX Secretary of State Voter Information
http://www.sos.state.tx.us/elections/voter/index.shtml