

Messin' with Texas: Deriving Mother's Maiden Names from Public Records

Virgil Griffith Markus Jakobsson
School of Informatics, Indiana University Bloomington

April 11, 2006

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 guessability of a mother's maiden name from close to 6.67 bits down to a zero entropy (i.e., certainty of their mothers maiden name) for large numbers 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.

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.

Knowing someone’s MMN is useful for a variety of attacks. If an attacker knew your MMN, she could call up your bank pretending to be you using your MMN as proof that she is you. Secondly, many websites use one’s MMN as a “security question” in case you forget your password. If an attacker knew your MMN (or had a small list of possible MMNs) and knew you were a registered user of particular website (which could be determined using a variety of techniques), she could learn your password and takeover your account.

2 Mother’s Maiden Name: Plan of Attack

The ubiquity of birth and marriage records that are publicly available online constitutes a direct path to deriving MMN’s through 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 is linking a child to the appropriate parents, and outputting 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 in which MMN derivation is quite easy. For example, if a groom has a very uncommon last name, then it is very easy to match him with any of his children simply by their uncommon last name. Secondly, if the child’s last name is hyphenated, an attacker will seldom have any trouble matching the child with the appropriate marriage. Third, if the birth record denotes that the child is suffixed “Jr.”, “III”, etc., an attacker can drastically narrow down the list of candidate parents by knowing both the first and last name will be the same as one of the parents. While each of these special cases make up only a small portion of the population, on large scales even the most obscure special cases can result in thousands of compromises.

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 ensure that any conclusions pertaining to the sample would be worthy of attention in their own right. This left us with two prominent 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. In particular, the ethnic composition of Texas is closer to that of the nation than California. This is of spe-

cial relevance considering that last names, and therefore maiden names, are strongly influenced by ethnicity. Texas is also more representative of both the percentage of foreign-born residents, and the frequency of households emigrating 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 that are likely to be on the same order as our findings.

We address the possibility of performing these attacks on the national level in section 2.1.9, but we should reiterate that for this analysis we can only learn the mother's maiden name of people who were *born and whose parents were married* in the state of Texas.

3 Availability of Vital Information

In smaller states, vital information (such as marriage and birth records) is usually held by the individual counties in which the recorded event 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 many will gladly ship to you for a modest fee). However, as time progresses, public information stored at even the smallest county offices will invariably become digitally available.

The Texas Bureau of Vital Statistics website [17] lists all in-state marriages that occurred between 1966 and 2002; records from before 1966 are available from the individual counties. Texas birth records from 1926 to 1995 are also available online, but the fields containing the names of the mother and father (including the MMN) 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 made.¹ In our analysis we were able to acquire the unredacted or “aged” birth records for the years between 1923 to 1949.

¹It may seem obvious, but it's worth mentioning that the average American lives well beyond the age of 50, making this security measure insufficient to protect privacy.

From the aged birth records alone we are able to fully compromise 1,114,680 males. Married females² are more difficult. However, the connection can still be made. We matched up females born from 1923 to 1949 with brides married from 1966 to 2002 using first and middle names together with the person's age. We were able to learn the MMN of 288,751 women (27% of those born in Texas between 1923 and 1949). It is worth noting that MMN compromise from aged records is not only easier, but more lucrative! Older people are likely to have more savings than younger adults.

Here it is worth mentioning that in October of 2000, Texas officially removed *online access* to their birth indexes due to concerns of identity theft[2]. Death indexes were similarly taken down as of June 2002 [3]. Texas also 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 online access was taken down, partial copies of the state 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]. The other from the user-contributed grassroots genealogy site Rootsweb.com [12] which had an even larger compilation of user-submitted birth 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 on microfilm or through their websites [14]. It is worrisome that over three years after being "taken down", the full death indexes are *still* available (although not directly linked) from the Texas Department of Vital Statistic's own servers at *exactly the same URL they were at before* [20]! All of this is particularly relevant because, even though Texas is now doing a better job protecting their public records (although largely for reasons unrelated to identity theft), texans are just as vulnerable as they were before.

4 Heuristics for MMN Discovery

So far we have shown that a cursory glance at birth and marriage records reveals an ample supply of low-hanging fruit. However, the correlation of marriage data (probably the best source of maiden names) with other types of public information comprises an effective and more general approach to

²We make the assumption that traditional naming conventions are used, i.e., that all women changed their last name to that of their husband.

linking someone to his or her MMN. 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 heuristics that an attacker could use to derive someone's MMN with high probability. As each heuristic is applied, the chance of MMN compromise is increased.

1. Children will generally have the same last name as their parents.
2. 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 (e.g. Garcia is a common hispanic name). An attacker does not have to pick the correct parents, just the correct MMN. This technique is described in more detail in section .
3. Couples will typically have a child within the first five years of being married. This is useful because knowledge of a child's age allows an attacker to narrow of the range of years in which to search for the parents' marriage.
4. Children are often born geographically close to where their parents were recently married, i.e., the same or a neighboring county. This is useful because if an attacker knows in which county a child was born (something readily available from birth records), she can restrict the search for marriage records to that and neighboring counties.
5. Parts of the parents' 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, Texas divorce records [19] 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 a child being born to a couple beyond a divorce date, but they also tell us how many children (if any) we should expect to find. In Texas, every divorce affects on average 0.79 children [18]. 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 9 months after the father's death. Texas death indexes are aged 25

years before release (full state-wide indexes for 1964–1975 are available online [20]). 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 the victim’s birth.

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

5 Experimental Design

With easy access to public records and no easy way to pull the records that have already been made public and widely mirrored, we should be asking ourselves, “How effective are the described attacks in leading to further MMN compromise?”, and “What percent of the population is at risk?” To answer these questions, we will use Shannon entropy to quantify the risk of MMN compromise from our attacks. Comparing the entropy of different distributions of potential MMNs is a suitable and illustrative measurement for assessing the vulnerability to these attacks. Entropy measures the amount of unpredictability within a distribution. In this case we use entropy to measure the uncertainty within a distribution of maiden names; the number of reoccurrences of each maiden name divide by the total number of names defines its probability within the distribution. The primary benefit of using entropy as a metric instead of simply counting the number of possible marriages that are possible after filtering is that entropy takes into account repetitions within the set of possible MMNs. For example, after applying all of our derivation rules there could be a set of 40 possible marriages from which a child could have come. However, 30 of these marriages may have the maiden name “Martinez”. Assuming that each name is equally likely, by guessing “Martinez” the attacker clearly has a far greater than a 2.5% chance ($1/40$) of correctly guessing the MMN as would be expected by simply counting the number of possible marriages.

After we have applied each of our heuristics, we can measure the chance of correctly guessing the victim’s MMN as follows. Let x be defined as the list of brides lastnames in marriage records that remain after applying our heuristics to a given person and let S be the most common name in list x .

Then define $|x|$ and $|S|$ as the number of items in list x and S , respectively. Then, we can define the chance of guessing the correct MMN as:

$$\text{Chance of guessing MMN} = \frac{1}{2^{\text{MinEntropy}(x)}}$$

$$\text{MinEntropy}(x) = -\log_2 \frac{|S|}{|x|}$$

To provide a baseline comparison for assessing the increased vulnerability due to attacks using public records, we calculated the entropy across all maiden names in our database (1966–2002) using no heuristics at all. By simply calculating the minentropy across all maiden names in our marriage records, we assess that the minentropy for randomly guessing the MMN is 6.67 bits, or about a 0.9% chance of guessing the correct mother’s maiden name simply by guessing the most common maiden name in the data set (Smith).

6 Assessing the Damage

Using our methods, we get the following graph (Figure 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 a pure guess, public records allow the attacker to take advantage of the fact that we know the victim’s last name (this is something the attacker would likely already know if attempting context-aware phishing). As previously mentioned, people with different last names will have different distributions of potential MMNs, and thus different entropies. 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 that the intended victim’s last name is “Smith” (resulting entropy = 12.18 bits), this reduces the entropy only by 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 correctly guessing the MMN. Note that for the absolute worst cases like

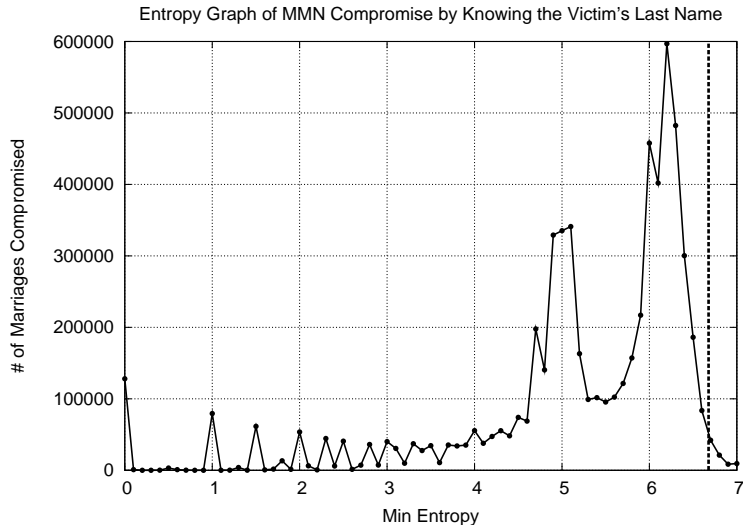


Figure 1: Ability to learn maiden names simply by knowing the victim’s lastname

“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, if an attacker has knowledge of the victim beyond his or her last name (such as age, place of birth, etc.), the attacker can eliminate large pools of candidate parents, and thereby improve the chances of determining the MMN. To allow effective comparison of different attacks, in Figure 2 we redraw Figure 1 as a cumulative percentage of marriage records compromised. We will then take the last names with the lowest entropies in the marriage analysis and look in birth records for children to compromise.

A guaranteed compromise (i.e. zero-entropy) of approximately 1% of marriages may not initially seem so terrible, but the table above shows that even the smallest percentages will lead to massive compromise.

7 Time and Space Heuristics

Although the first attack is the easiest and most assured route to MMN compromise, to gain more compromises there are times when an attacker would be willing to apply further heuristics, such as 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

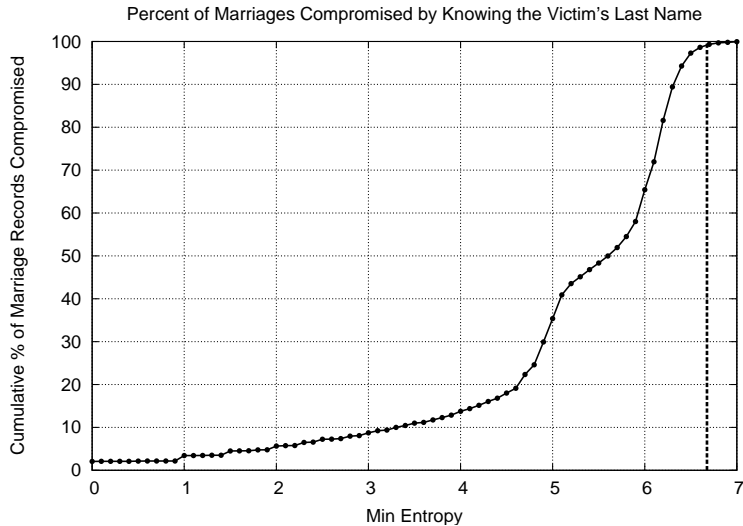


Figure 2: Drawing Figure 1 as a cumulative percentage

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

Table 1: Using birth records from 1966–1995 to search for children with highly unusual last names. The percentage of marriage records compromised (the graphs) does not necessarily reflect the percent of birth records compromised (the tables).

desires. Naturally, longer windows increase the chances of including the parents' marriage record, while shorter windows yield higher percentages of compromised MMNs (assuming the windows are correct). In this example we assume the attacker knows not only the victim's last name, but also 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 (which can be obtained from birth records and sometimes even social networks). This attack uses a five year window up to and including the year the victim was born to search for the parents' marriage record. Thus, it deduces MMNs

in accordance with the heuristic 3, that couples frequently have children within the first five years of being married. The statistics do vary from year to year, but for the reader’s convenience we have averaged all years into a single graph.

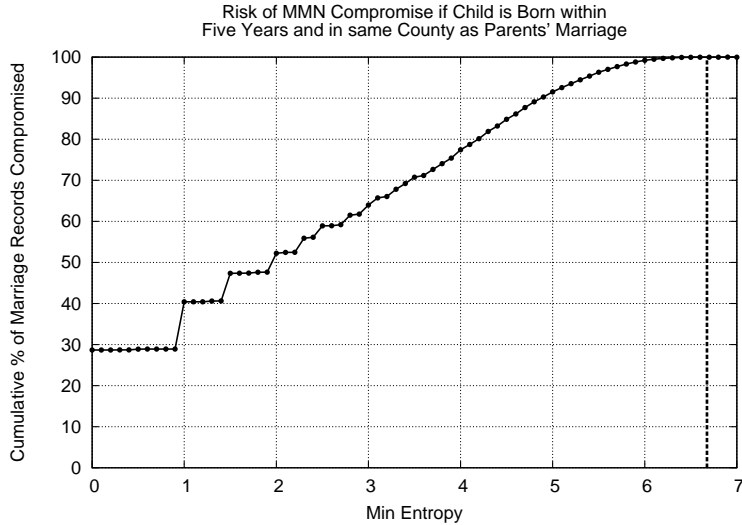


Figure 3: Risk of MMN compromise when parents’ marriage county is known and the marriage year known to within 5 years

Entropy	# Children Compromised	% Birth Records Compromised
= 0 bits	809,350	11.6
≤ 1 bit	1,278,059	18.3
≤ 2 bits	1,844,000	26.5
≤ 3 bits	2,459,425	35.3

Table 2: Using birth records from 1966–1995 to look for children – applying heuristics 1, 2, 3, and 4. The percentage of marriage records compromised does not necessarily reflect the percent of birth records compromised.

By narrowing our window in which to look for candidate marriages, the resulting entropies on the distributions of potential MMNs drop substantially. An attacker can increase or decrease the window size based upon the uncertainty of the marriage year. As the window increases, there are fewer “guaranteed” compromises (distributions with zero entropy), but any

“guaranteed” compromises are more reliable as there is a better chance that the correct marriage record being included within the window.

8 MMN Compromise in Suffixed Children

Our final 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’s name has a suffix is immensely valuable as it specifies the first name of one of the groom listed in the marriage record.

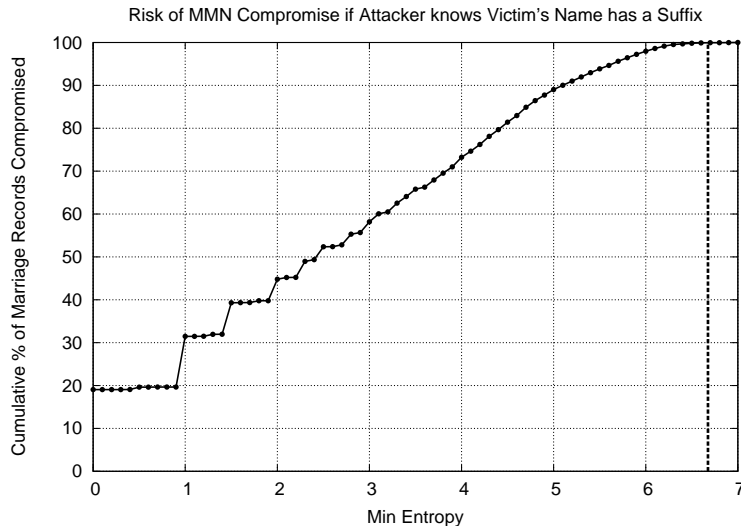


Figure 4: Ability to determine MMNs for suffixed children

Entropy	# Children Compromised	% Birth Records Compromised
= 0 bits	78,197	13.7
≤ 1 bit	126,153	22.1
≤ 2 bits	178,234	31.3
≤ 3 bits	231,678	40.7

Table 3: Using birth records from 1966–1995 to look for suffixed children. The percentage of marriage records compromised does not necessarily reflect the percent of birth records compromised.

9 Other Ways to Derive Mother’s Maiden Names

Hereto we have focused on the use of birth and marriage records in compromising MMNs. Although birth and marriage information probably constitute the greatest threat to large-scale MMN compromise, it is by no means the only viable route. The following is a list of more creative public-records attacks that have worked in our in sample tests, but which so far remain largely unexplored.

Social Security Death Index The Social Security Death Index (SSDI) [15] provides up-to-date information on 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 by the public. 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. (However, there is still always the possibility for name overlap, particularly as you increase in scale.) By taking information from the Texas state death index 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 verify his membership in a particular marriage as the marriage record contains the bride and groom’s age. This is a reminder of the ability of an attacker to combine different many types public records into much stronger attacks.

Voter Registration Records. In efforts to prevent voter fraud voter registration records are, by U.S. law [10], 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 an unknown or missing birth or marriage record. Voter registration contain the full name, “previous name” (the maiden name), date of birth, and county of residence [21]. In 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.

Genealogy Websites. Not only a source for mirrored public records data, Rootsweb [13] 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 last name (as opposed to her maiden name), an attacker can simply look at the last name of the bride’s father or one of her brothers. If for some reason this information is not listed, the bride’s 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 almost 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 Latter-day Saints’ FamilySearch.org, avoids listing information about living people.

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, or even within the same obituary. Because the children may have moved to other parts of the country, simply looking them up in the local phonebook will 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].

Property Records. At our current scale, property records are of relatively little value. However, 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 [16]. For the purpose of deriving maiden names, property records can be thought of as phonebooks that owners are legally required to be in.

10 Conclusion

Our analysis shows that the secrecy of MMNs is vulnerable to the automated data-mining of public records. New data-mining attacks show that it is increasingly unacceptable to use a documented fact as a security authenticator. 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. If these actions have accomplished anything of significance, 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. We can only speculate as to the long term effects of policies that would routinely restrict access to 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 separate mother's maiden name, much less a secret one, is becoming obsolete. At one time, the mother's maiden name served as a convenient and reasonably secure piece of information. However, sociological changes have made it socially permissible for a woman to keep her original name. As time progresses, this will further weaken the secrecy of MMNs. Today, there are far more secure ways to authenticate oneself either online or over the phone. Given the current and future state of the MMN, we encourage their speedy adoption.

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.htm> 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>
- [7] Google Phonebook Search for “Smith” in zipcode 75201 (Dallas, TX)
<http://www.google.com/search?pb=r&q=Smith+75201>
- [8] Jakobsson, Markus; “Modeling and Preventing Phishing Attacks,” Phishing Panel at Financial Cryptography ’05. 2005.
<http://www.markus-jakobsson.com>
- [9] Sweeney, Latanya; Malin, Bradley: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics* pages 179-192
- [10] National Voter Act of 1993
<http://www.fvap.gov/laws/nvralaw.html>
- [11] Texas State Property Records
<http://www.txcountydata.com>
- [12] Rootsweb.com FTP server with complete copies of both the marriage and death indexes
<ftp://rootsweb.com/pub/usgenweb/tx/>
- [13] RootsWeb.com Home Page
<http://www.rootsweb.com>

- [14] SearchSystems.net listing of Texas Counties' online public record offerings
<http://searchsystems.net/list.php?nid=197> <http://searchsystems.net/list.php?nid=344>
- [15] Social Security Death Index <http://ssdi.genealogy.rootsweb.com/>
- [16] Texas State Property Records <http://www.txcountydata.com>
- [17] Texas Department of Health, Bureau of Vital Statistics, Marriage Indexes
<http://www.tdh.state.tx.us/bvs/registra/marridx/marridx.htm>
- [18] Texas Department of Health, Divorce Trends in Texas, 1970 to 1999
www.tdh.state.tx.us/bvs/reports/divorce/divorce.htm
- [19] Texas Department of Health, Bureau of Vital Statistics, Divorce Indexes
<http://www.tdh.state.tx.us/bvs/registra/dividx/dividx.htm>
- [20] Texas Department of Health, Bureau of Vital Statistics, General and Summary Death Indexes
<http://www.tdh.state.tx.us/bvs/registra/deathidx/deathidx.htm>
- [21] TX Secretary of State Voter Information
<http://www.sos.state.tx.us/elections/voter/index.shtml>