On the Guessability of Resident Registration Numbers in South Korea

Youngbae Song[1], Hyoungshick Kim[1], and Jun Ho Huh[2]

[1] Department of Computer Science and Engineering,
Sungkyunkwan University, Republic of Korea
{youngbae, hyoung}@skku.edu
[2] Honeywell ACS Labs, USA
junho.huh@honeywell.com

Abstract. This paper studies a potential risk of using real name verification systems that are prevalently used in Korean websites. Upon joining a website, users are required to enter their Resident Registration Number (RRN) to identify themselves. We adapt guessing theory techniques to measure RRN security against a trawling attacker attempting to guess victim's RRN using some personal information (such as name, sex, and location) that are publicly available (e.g., on Facebook). We evaluate the feasibility of performing statistical-guessing attacks using a real-world dataset consisting of 2,326 valid name and RRN pairs collected from several Chinese websites such as Baidu. Our results show that about 4,892.5 trials are needed on average to correctly guess a RRN. Compared to the brute-force attack, our statistical-guessing attack, on average, runs about 6.74 times faster.

Keywords: Korean identification system; Resident Registration Number; Brute-force attack; Statistical-guessing attack.

1 Introduction

Just as Social Security Numbers (SSNs) are used in the United States to identify residents for various governmental purposes (e.g., tax or benefits), all Korean residents get a national identification number called the "Resident Registration Number" (RRN). It is a unique 13-digit number that is issued to every Korean at birth; the "Resident Registration Act," which came into effect on 1962, prohibits anyone from changing the issued RRN. Government, banking, and medical services all use RRNs to identify and track Korean residents. Hence, RRNs are very sensitive and their confidentiality must be protected. The reality, however, is that there are too many careless uses of RRNs both online and offline. As a result, millions of valid RRNs are actively being traded in the Chinese black-markets. In March 2010, a group of criminals were arrested for attempting to sell about 20 million RRNs [10].

Although online anonymity protects users' privacy, it can also be misused by those who try to spread rumors or defame others. To mitigate such undesirable online activities, the Korean government passed the "Real Name Verification Law" in July 2007, which regulates the following [4, 8]:

> *If a website that has more than 100,000 visitors per day, all users of that website must verify their real name in order to sign up or write posts.*

The most popularly deployed verification system verifies real names by asking users to enter their RRNs upon joining a website or writing posts. Hundreds and thousands of popular Korean websites started collecting users' RRNs without following any security standards for protecting them. Such trends exposed huge privacy risks, and as one would expect, there were several large-scale database breaches in the last few years that led to users' RRNs being compromised. For example, the Cyworld and Nate database breach [13] affected about 35 million users. Such incidents forced the government to amend the Real Name Verification Law. The amended law prohibits information and communication service providers from collecting RRNs from their users. That amendment did not discourage many websites though, and many websites still use RRN-based real name verification systems.

In this paper, we discuss a security risk associated with using RRN-based identification systems, demonstrating how easy it is to guess RRNs using commonly used name verification systems. At first glance, the theoretically possible space of 13-digits numbers looks sufficiently large to resist brute-force attacks. This is not true though. The actual RRN space is much smaller, making various types of guessing attacks feasible. Although similar guessing attack has been performed on the U.S. national identification number [1], our work is another valuable case study that provides further insight into the security implications of deploying a national identification number system. Our results further emphasize that deploying a secure and usable national identification number system is challenging because they can easily be misused to impersonate others.

To analyze the security of RRN-based identification systems, we adapt guessing theory techniques to demonstrate how robust existing systems are against a trawling attacker trying to guess correct RRNs using some publicly available personal information such as name, sex or location that can be obtained easily through popular social networks like Facebook or Linkedin. We used real-world datasets (collected from Chinese websites like Baidu) consisting of 2,326 valid name and RRN pairs to evaluate the feasibility of performing statistical guessing attacks which take into account the probability distribution of real RRNs. Using our statistical-guessing attack, only about 4,892.5 trials are needed on average to correctly guess an RRN, outperforming the pure brute-force attack by about 6.74 times on average. As a result, the actual security of RRN-based identification systems is worse than our hopeful expectation.

To mitigate such statistical-guessing attacks, we recommend using a security policy to limit the number of RRN verification attempts. Our analysis demonstrates that we can effectively prevent about 99.94% of guessing attack attempts by setting that number to 7.

The rest of the paper is organized as follows. Section 2 explains the structure of RRNs. Section 3 analyzes the guessability of RRNs with the collected RRN datasets. Our suggestions against guessing attacks are discussed in Section 4. Next, we explain how ethical issues were considered in Section 5. Related work is discussed in Section 6, and our conclusions are in Section 7.

2 Structure of Resident Registration Numbers

RRNs are pervasively used on the Internet, allowing the government, banking, and medical services to identify the Korean individuals. An RRN is validated by comparing the last digit against what it should be based upon the rest of the digits entered. In this section, we describe in detail how RRNs are validated.

2.1 Resident Registration Numbers

RRN is a 13-digit number issued to an individual by the Korean government for tracking individuals efficiently. RRNs are much like national identification numbers used in other countries (e.g., Social Security Numbers (SSN) used in the US), and are used by tax, banks, and websites to identify and authenticate the residents in South Korea. However, unlike SSN that is decoupled from individuals' personal data since 2011, RRNs contain residents' personal information such as “date of birth” and “place of birth”. That number has the following structure:

$$yymmdd{-}sccppnv$$

The first six digits, $yymmdd$, represent an individual's date of birth in the order of year, month, and day. For example, an individual born on March 21, 1987 would have an RRN that starts with 870321. The seventh digit, s, indicates the sex of an individual. The eighth through eleventh digits, $ccpp$, represent the place of birth. The eight and ninth digits, cc, signify an individual's city of birth (e.g., Seoul). The tenth and eleventh digits, pp, signify the “dong” of birth, which is the smallest region in a city that has its own government office and staff. The twelfth digit, n, is a sequential number used to differentiate those that have the same sex, born on the same day and in the same location (i.e., dong). The thirteenth digit, v, is used to verify the digit. It is generated from the rest of the digits using the following equation:

$$v = (11 - (\sum_{i=1}^{8}(i+1) \cdot ar[i] + \sum_{i=9}^{12}(i-7) \cdot ar[i] \bmod 11)) \bmod 10$$

In summary, RRNs consist of information about birth (i.e., date of birth, place of birth, and birth registration order). However, since some parts of the birth data and/or their statistical properties are already available to the public, the real RRN space is much smaller than that of the theoretical space—for example, Gross et al. [5] found that 87.8% of active Facebook users revealed their birth date—this is why guessing attacks can be effective on RRNs. The guessability of RRNs is discussed in the next section.

2.2 Example of Resident Registration Numbers Service

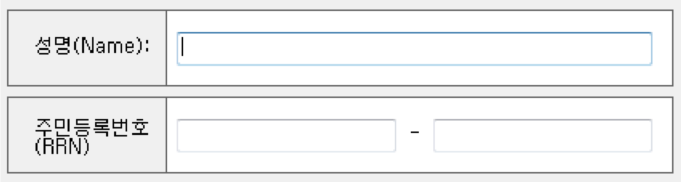


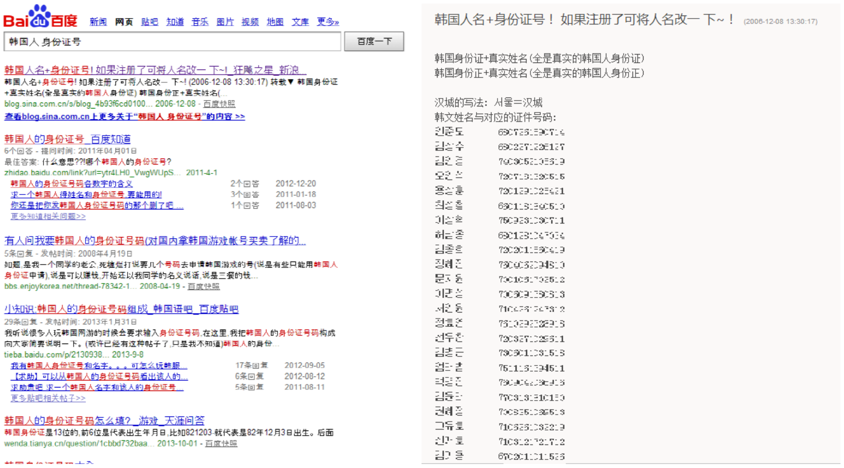
Fig. 1. Example of Resident Registration Numbers Service

A RRN is a government issued 13-digit identification number assigned to South Korean residents, which is used when residents register online or make online transactions. Many South Korean websites require users to enter their full name and a valid RRN to sign up and retrieve forgotten passwords. (see Figure 1).

3 Guessability of RRNs

To evaluate how secure RRN-based identification systems are against guessing attacks, we first collected real-world RRNs and analyzed their statistical characteristics.

Interestingly, large volumes of Korean RRNs are available on Chinese websites such as Baidu (`http://www.baidu.com`), which we suspect are from



韩国人 身份证号

韩国人名+身份证号! 如果注册了可将人名改一 下~!_狂飚之星_新浪...

韩国人名+身份证号! 如果注册了可将人名改一 下~! (2006-12-08 13:30:17) 转载▼ 韩国身份证+真实姓名(全是真实的韩国人身份证) 韩国身份正+真实姓名(...

blog.sina.com.cn/s/blog_4b93f6cd0100... 2006-12-08 - 百度快照

查看blog.sina.com.cn上更多关于"韩国人 身份证号"的内容 >>

韩国人的身份证号_百度知道

6个回答 - 提问时间: 2011年04月01日

最佳答案: 什么意思??!哪个韩国人的身份证号?

zhidao.baidu.com/link?url=ytr4LH0_VwgWUpS... 2011-4-1

韩国人的身份证号码各数字的含义	2个回答	2012-12-20
求一个韩国人得姓名和身份证号,要能用的!	3个回答	2011-01-18
你还是把你发韩国人身份证号码的那个删了吧 ...	1个回答	2011-08-03

更多知道相关问题>>

有人问我要韩国人的身份证号码(对国内拿韩国游戏帐号买卖了解的...

5条回复 - 发帖时间: 2008年4月19日

如题,是我一个同学的老公,死缠烂打说要几个号码去申请韩国游戏的号(说是有些只能用韩国人身份证申请),说是可以赚钱,开始还以我同学的名义说话,说是三餐的钱...

bbs.enjoykorea.net/thread-78342-1... 2008-04-19 - 百度快照

小知识:韩国人的身份证号码组成_韩国语吧_百度贴吧

29条回复 - 发帖时间: 2013年1月31日

我听说很多人玩韩国网游的时候会要求输入身份证号码,在这里,我把韩国人的身份证号码构成向大家简要说明一下。(或许已经有这种帖子了,只是我不知道)韩国人的身份...

tieba.baidu.com/p/2130938... 2013-9-8

我有韩国人身份证号和名字。。。可怎么玩韩服...	17条回复	2012-09-05
【求助】可以从韩国人的身份证号码看出该人的...	6条回复	2012-08-12
求助贵吧 求一个韩国人名字和该人的身份证号...	5条回复	2011-08-11

更多贴吧相关帖子>>

韩国人的身份证号码怎么填? _游戏_天涯问答

韩国身份证是13位的,前6位是代表出生年月日,比如821203-就代表是82年12月3日出生。后面

wenda.tianya.cn/question/1cbbd732baa... 2013-10-01 - 百度快照

韩国人名+身份证号！ 如果注册了可将人名改一 下~！ (2006-12-08 13:30:17)

韩国身份证+真实姓名(全是真实的韩国人身份证)
韩国身份正+真实姓名(全是真实的韩国人身份正)

汉城的写法：서울=汉城
韩文姓名与对应的证件号码:

[redacted]

(a) Chinese websites (Baidu) (b) Korean residents' RRNs

Fig. 2. Examples of RRNs accessed through Chinese websites like Baidu.

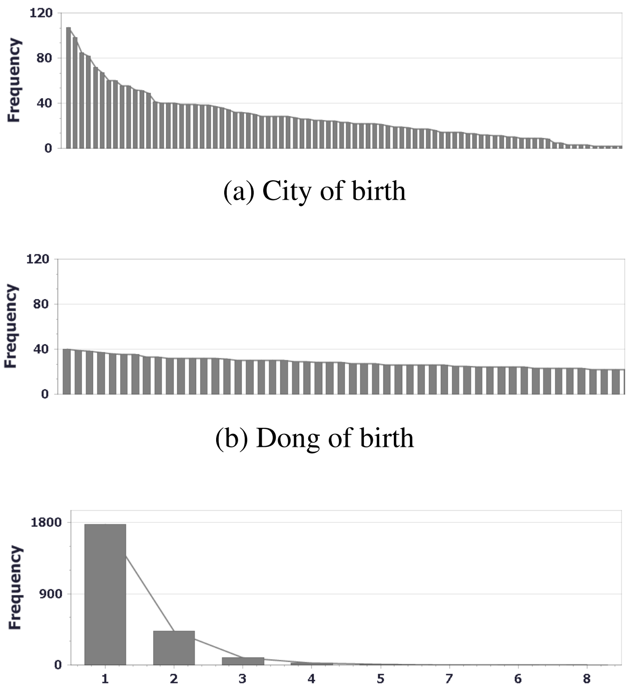
previous RRN leaks. RRNs are lucrative targets for Chinese hackers too, because RRNs are often needed to impersonate and create user accounts on Korean websites [7, 10]. With increasing Korean pop and drama popularity in China, there is also a rapidly growing interest in Chinese population to access Korean websites, especially Korean online shopping malls, and to purchase trendy Korean items (e.g., clothes and accessories worn by Korean celebrities). Figure 2 shows some examples of RRNs accessed through Chinese websites like Baidu.

During a two-days period, we found several Chinese webpages containing RRNs and collected a total of 3,007 name and RRN pairs. Through a Korean website using RRN-based authentication, we tested the validity of the collected pairs and finally obtained 2,326 valid ones. We examine the statistics of those RRNs as follows.

3.1 Occurrence frequency of birth data in RRNs

First, we analyzed the occurrence frequency of the city of birth in the collected RRNs. The frequencies are graphed in a descending order (see Figure 3 (a)). The histogram shows the occurrence frequency of the city of birth decreasing dramatically, which indicates that the city of birth distribution is heavily skewed in favor of a small number of common birth places (e.g., Seoul).

Also, the occurrence frequencies of the dong of birth (fine-grained location), sorted in a descending order, are shown in Figure 3 (b). In contrast, the dong of



(a) City of birth

(b) Dong of birth

(c) Birth registration order

Fig. 3. Occurrence frequency of birth data in RRNs.

birth was more evenly distributed. It would be relatively easier for an attacker to guess the birth city than the birth dong.

Last, the occurrence frequency of the birth registration order was analyzed. The frequencies are also sorted in a descending order (see Figure 3 (c)). The histogram shows the birth registration order occurrence frequency decreasing dramatically from the 2nd order, which indicates that the birth registration order distribution is heavily skewed. The most popular order, ‘1st’, alone, accounted for 76.3% (1,775 out of 2,326) of the total number of the collected individuals.

3.2 Correlation between city of birth and dong of birth

We also analyzed the correlation between the birth city and the birth dong. Plotting the relationship between them in a 2-dimensional grid (see Figure 4; darker the color the higher the number of combinations found) highlighted that there are some combinations of city and dong that appear more frequently in the collected dataset. The most frequently appearing combination had a city code of 93 and a dong code of 21. There were also a few combinations that did not occur at all (see the blank parts in Figure 4). This trend indicates that the city-dong combination distribution might also be highly skewed in favor of a small number of popularly occupied locations. As a result, we can see that the real RRN space is much smaller than that of the theoretical space, which would make dictionary attacks more effective.

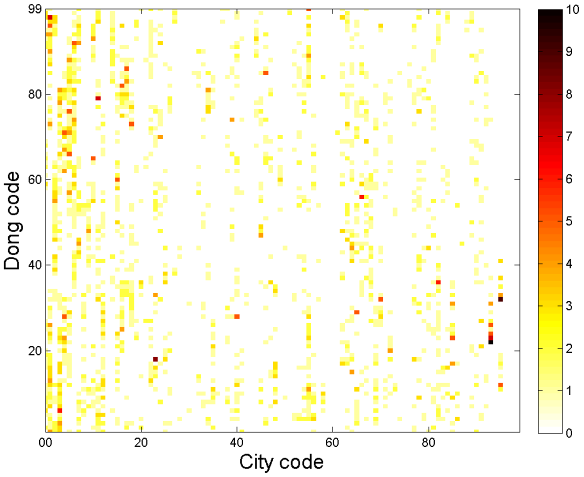


Fig. 4. Relationship between city code and dong code.

3.3 Effectiveness of guessing attacks

Given an individual's name, sex, and birth date, which can be attained through social networks like Facebook, Twitter, or Linkedin, the goal of the guessing attack is to find his or her corresponding RRN. Stolen RRNs can be used to help criminal activities, e.g., allowing rogue accounts to be created on Korean websites and impersonating innocent people, or accessing Korean residents' personal records or confidential documents maintained by the Korean government. We consider the following, specific adversary.

The adversary already has access to a victim's publicly attainable information (i.e., name, sex, and birth date) and is capable of accessing an RRN validation service, which allows one to submit a name and RRN pair through an online form and validate it. Given that accessibility, the adversary tries to guess the victim's RRN by enumerating through every possible combination of RRN until a valid one is found. In theory, the adversary needs to try 10^5 possible RRNs.

Based on such statistical characteristics of the collected set of RRNs, we designed a "statistical-guessing attack" that is highly effective on RRNs. With the two strongly skewed city-dong and birth registration distributions, our RRN guesses were sorted in a descending order by the frequency of city-dong combinations and also by the frequency of birth registration order.

To evaluate the effectiveness of the proposed statistical-guessing attack, the ten-fold cross validation was performed ten times. In each run, one of the folds was used for validation, while the remaining folds were used for obtaining statistical properties of RRNs. We compared the performance of the statistical-guessing attack against a basic brute-force attack that sequentially enumerates every possible RRN combination—this is the lower bound strategy of a guessing attack since it can be performed without any knowledge of RRNs. Those attack results are plotted in Figure 5.

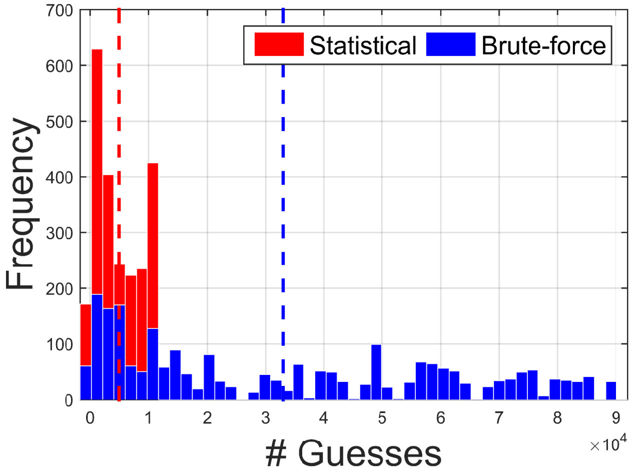


Fig. 5. Histograms of the numbers for successfully guessing a RRN for statistical-guessing (red) and brute-force (blue) attacks. The x-axis and the y-axis represent the number of guesses to successfully infer a RRN and the number of frequency in each bin, respectively. The dotted lines represent the mean of guess numbers.

The graph shows that the number of guessing attempts made in the statistical-guessing attack is significantly lower than the number of guessing attempts made in the brute-force attack ($p < 0.001$, two-tailed unpaired t-test). The brute-force attack has a mean number of 32,974.9 guessing attempts with a standard deviation of 27,905.9, while the statistical-guessing attack has a mean number of 4,892.5 guessing attempts with a standard deviation of 3,913.4. The statistical-guessing attack is about 7 times faster. Moreover, unlike the brute-force attack which often failed even with 50,000 guesses, all of the statistical-guessing attack attempts completed successfully (that is a valid RRN was found) within about 10,000 guesses.

4 COUNTERMEASURES

We suggest two defense mechanisms to mitigate the statistical-guessing attack discussed in the previous section.

4.1 Limiting the number of RRN verification fail attempts

To mitigate this type of attack, a straightforward solution is to limit the number of RRN verification attempts. The idea of limiting the number of attempts from a particular user (e.g., with an IP address) or imposing a minimum time interval between failed attempts is not new [2]. However, the vast majority of the real-world RRN verification systems that we investigated did not limit the number of attempts, and did not seem to be considering the threats associated with brute-forcing RRNs.

To find an optimal number of RRN verification fail attempts that should be allowed (i.e., a number that would effectively prevent online guessing attacks), we calculated a range of successful rates of guessing attacks by varying the “maximum fail attempts allowed” from 1 to 10,000. Figure 6 shows the results. 7 seems to be a reasonable number to be used as the maximum fail attempts allowed since it can successfully stop 99.94% of statistical-guessing attacks.

4.2 Anomaly detection

Another promising approach to explore is *anomaly detection*. Guessing attacks that try to brute-force RRNs must inherently generate and send a large volume of query (RRN-based real name verification) messages in a very short time period. Such a spike of query messages will result in unusual traffic patterns and can be treated and flagged as anomalies. For instance, when a service provider receives a series of RRN-based real name verification messages from one IP address, we can classify such messages as a potential guessing attack because that traffic will look significantly different from normal user traffic. Our suggestion is to design and deploy such an anomaly detection system on the RRN verification server. Implementation techniques for anomaly detection are already widely available (e.g., see [3]).

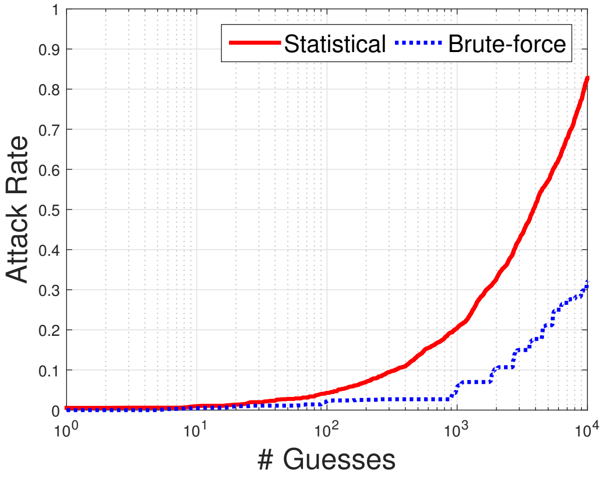


Fig. 6. The successful attack rates of statistical-guessing (red) and brute-force (blue) attacks with varying maximum allowed number from 1 to 10000. The x-axis and the y-axis represent the maximum allowed number for guessing attacks and the successful attack rate, respectively.

5 Ethical Considerations

Our analyses were conducted based on publicly available RRN dataset that we attained through Chinese websites such as Baidu. We clarify that it was not our intention to collect or misuse personal information in any way, or use the collected data for commercial purposes. As can be seen from the previous sections, our goal is to identify security risks associated with current RRN-based identification systems, and recommend practical mitigation solutions to make it difficult for adversaries to perform effective guessing attacks.

6 Related Work

Our primary focus is on analyzing security risks associated with Korean RRN-based name verification systems by performing both simple and more advanced statistical guessing attacks on real-world RRNs. Understanding security issues of RRN-based verification is important because it is still widely used online. In fact, the Korean government passed the *Real Name Verification Law* in July 2007 so that only RRN-verified users (with their real identity) may post comments on websites. Cho [4] discussed both the positive and negative effects of passing the *Real Name Verification Law*. For instance, Cho mentions that identifying posts had a significant positive impact on reducing uninhibited behaviors.

In response to the growing concern about RRN database leaks and severe misuse of the stolen RRNs [10], Pak et al. [11] investigated alternative methods for reliably verifying users' identity and age. In particular, they proposed a new national identification service called "i-PIN," where the Korean government issues and manages an i-PIN identifier and a matching password for Korean residents. The Japanese government also introduced a similar identification service called "My Number" [9].

National identification programs as such, however, have often been exposed to security breaches. The Korean i-PIN system was hacked in 2015, with 750,000 Korean citizens' i-PIN accounts being disclosed [6].

Oh et al. [10] showed that partial RRN information (the first 6 digits representing the date of birth) can be used to design sophisticated phishing attacks. Our work demonstrates how successful statistical-guessing attacks (that use some publicly available personal information) can be on guessing RRNs from real-world name verification services.

Sweeney et al. [12] studied the security of encrypted RRNs, showing that 23,163 encrypted RRNs can be successfully revealed using two de-anonymization methods.

Acquisti et al. [1] analyzed the correlation between the U.S. social security numbers and the individual's birth data, demonstrating the feasibility of statistically inferring social security numbers. We conducted a similar study to show the security risks of another national identification number system that is being used in Korea, showing that current Korean RRNs can be guessed with a high chance based on the personal information collected through public data sources.

7 Conclusions

In this paper, we explore an efficient guessing attack that can be performed using publicly available personal information (e.g., obtained through social networks) to reveal national individual identifiers, such as the Resident Registration Numbers (RRNs) used in South Korea. Upon joining websites, Koreans are often required to enter their RRN to prove their identity. Inadequate security protections deployed on such websites expose holes that can be exploited by adversaries to steal RRNs and impersonate innocent users.

We designed a moderately advanced statistical-guessing attack that uses some publicly available personal information such as name, sex, and birth date, to efficiently and accurately guess RRNs. We evaluated the effectiveness of the proposed statistical-guessing attack with a real-world RRN dataset that consists of 2,326 valid name and RRN pairs. Our experimental results showed that the statistical-guessing attack can guess RRNs with a much smaller number of guessing attempts compared to pure brute-force attacks (that try all the possible combinations without any smart rule). Intriguingly, with the proposed attack, only about 4,892.5 trials were needed on average to successfully guess an RRN.

We suggested two possible mitigation techniques that would work well against statistical-guessing attack. One technique is to limit the number of consecutive RRN verification fail attempts allowed, significantly slowing down online attacks. Our feasibility analysis demonstrated that a carefully set fail attempts allowed number, e.g., 7, would prevent about 99.94% of statistical-guessing attacks. Most legitimate users should be able to enter their correct RRN within 7 attempts. We strongly recommend that a policy for limiting the number of consecutive fail attempts allowed be enforced on RRN-based name verification systems.

8 Acknowledgements

This work was supported in part by the NRF Korea (No. 2014R1A1A1003707), the ITRC (IITP-2015-H8501-15-1008), and the MSIP/IITP (2014-PK10-28). Authors would like to thank all the anonymous reviewers for their valuable feedback.

References

1. Acquisti, A., Gross, R.: Predicting Social Security numbers from public data. Proceedings of the National Academy of Sciences **106**(27) (2009) 10975–10980
2. Alsaleh, M., Mannan, M., Van Oorschot, P.: Revisiting defenses against large-scale online password guessing attacks. IEEE Transactions on Dependable and Secure Computing **9**(1) (2012) 128–141
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41**(3) (2009) 15
4. Cho, D.: Real Name Verification Law on the Internet: A Poison or Cure for Privacy? In: Proceedings of the 10th Workshop on Economics of Information Security. (2011)
5. Gross, R., Acquisti, A.: Information Revelation and Privacy in Online Social Networks. In: Proceedings of the ACM Workshop on Privacy in the Electronic Society. (2005)
6. Kovacs, E.: Personal Details of 27 Million South Koreans Stolen by Hacker (2014)
7. Lee, R.: Korean national ID numbers spring up all over Chinese Web (2011)
8. Lee, T.B.: South Korea's "real names" debacle and the virtues of online anonymity (2011)
9. Miyata, S., Suzuki, K., Morizumi, T., Kinoshita, H.: Access Control Model for the My Number National Identification Program in Japan. In: Computer Software and Applications Conference Workshops. (2014)
10. Oh, Y., Obi, T., Lee, J.S., Suzuki, H., Ohyama, N.: Empirical analysis of internet identity misuse: case study of south korean real name system. In: Proceedings of the 6th ACM Workshop on Digital Identity Management. (2010)
11. Pak, H., Kim, C., Choi, H.: Preparation a Study on the Use of the Resident Registration Number and Alternatives for RRN. World Academy of Science, Engineering and Technology **6**(11) (2012)
12. Sweeney, L., Yoo, J.S.: De-anonymizing South Korean Resident Registration Numbers Shared in Prescription Data. Technology Science (2015)
13. Yang, S.: 35m Cyworld, Nate users' information hacked (2011)