A Model for the Creation of Biographical Dictionaries

Marcus Birath [1], Johan Ginman[1] and Joakim Kävrestad [1]

[1] *University of Skövde, Högskolevägen 1, Skövde, Sweden*

Abstract

The use of encryption is increasing, and while that is good for cybersecurity it is a core challenge for digital forensics. Encrypted information cannot be analyzed unless it is first decrypted, which is a complex and time-consuming process. Using a brute force attack to guess the password used for encryption is deemed impractical as even a simple password, being long enough, could take weeks, months, or even years to find. A more feasible approach is to use a dictionary attack where each word in a list is tested. However, a dictionary attack is only successful if the password is in the list, making the process of creating that list a crucial part of decrypting passwords. This research builds on existing literature showing that users commonly use strategies to create passwords, and the aim is to propose a method for creating dictionaries that are grounded in theories of password construction. An initial model was developed using a selective literature review with the purpose of identifying common elements included in biographical passwords, and in what order the elements are used. To improve the model, the study utilized semi-structured interviews with forensic experts from the Swedish police and the Swedish National Forensic Center (NFC). The main contribution of this research is a readily available model for creating dictionaries that can be used by practitioners. The model can also serve as a theoretical contribution that describes how users commonly construct biographical passwords.

Keywords

passwords, biographical dictionary, password cracking, digital forensics

1. Introduction

While current digital advancements generally are considered beneficial for the society, they also enable new opportunities for criminals, consequentially creating new challenges for law enforcement [1]. The digital world has rapidly become a platform for criminal activities such as selling drugs, child abuse, or extorsion. Thus digital evidence obviously plays a vital role in crimes conducted in a digital environment. However, digital evidence also plays an important role in prosecuting all types of crimes as the general use of personal digital devices has increased and are therefore likely to contain information of investigative importance (e.g. communication and chat data, position and location data, affiliations with other suspects, and images). Consequently, digital devices hold evidence that is important for modern criminal investigations. The process of securing and analyzing data from such devices is called digital forensics, and the need for digital forensics in criminal investigations is increasing rapidly [2].

One of the core challenges of digital forensics is encryption, as it is the main method used by criminals to restrict access to data stored on a digital device [3][4]. There are many encryption techniques available, and today many digital devices (e.g. smartphones or computers) come with encryption enabled by default which can encrypt a folder, an entire drive of a PC, an application, or a

8th International Workshop on Socio-Technical Perspective in IS development, August 19-21, 2022 (STPIS'22), Reykjavik, Iceland
EMAIL: marcus.birath@his.se (A. 1); a18johgi@student.his.se (A. 2); joakim.kavrestad@his.se (A. 3)
ORCID: 0000-0001-5692-4008 (A. 1); 0000-0002-6867-172X (A. 2); 0000-0003-2084-9119 (A. 3)


cloud service automatically or in just a few clicks [4]. In most cases, these encryption techniques are used in conjunction with passwords. As the encryption techniques themselves are often robust and standardized, they cannot be attacked directly, leaving the process of password recovery a necessity for forensic experts to decrypt data and collect evidence. Many encryption schemes are created to withstand brute force attacks, and it is, therefore, critical to find other methods in order to collect evidence in a timely manner [4]. One such method is to dump the memory and analyze it for passwords or encryption keys. Memory analysis is a technically advanced technique that requires the device to be acquired in a powered-on state and that precautions are taken not to alter the volatile data. A more feasible method is to find the password using a dictionary attack where each password in a list is tested. However, a dictionary attack is only successful if the password is in the list, making the process of creating that list a crucial part of decrypting passwords. As users tend to create easy to remember passwords, dictionaries that contain biographical data such as names, dates, hobbies, and other personal information are suggested to have a higher possibility of being successful as they relate to common strategies used for password creation [1][5].

Simple dictionary attacks usually include existing lists of previously leaked passwords which are unlikely to contain a password that is based on a targeted person's biographical data. To be more successful, biographical information about the targeted person needs to be collected by forensic experts and put into a text file which can be used with programs such as JohnTheRipper together with combination and modification rules to ultimately find the correct password. The success rate is thus dependent on identifying and collecting the correct biographical information.

This research aims to create a model to support the creation of dictionaries used when cracking passwords based on biographical data. The model should aid forensic experts in the acquisition of biographical data during a forensic investigation and in the composition of a dictionary that is likely to contain the targeted password and do so in the most time-efficient manner. The explicit focus just described implies that this research will neither cover the technical details of cracking passwords nor how to create dictionaries to crack passwords based on other elements than biographical information.

2. Method

This study has an action-based approach that, according to [6], is a collaborative method where systematic action is used to reach a specific goal and improve specific practices. This research includes collaborative development with professional forensic experts from law enforcement. First, based on current research, the facts about the creation of biographical passwords were collected. Second, an initial model was created based on the output from the previous step. Finally, an iterative process began where the model was revised based on interviews with field experts.

2.1. Selective Literature Review

To build a foundation for the initial model this study utilized a selective literature review which, according to [7], is an excellent way to start action-based qualitative research. This also ensures that the research is grounded in relevant previous research [8]. The selective literature review was used to identify components that commonly make up biographical passwords and in which order they are used. The study used the databases Sciencedirect, Wiley, and ACM with the following search terms:

- "biographical" AND "password"
- "password" AND "personal information"
- "password" AND "building"
- "biographical" AND "dictionary" AND "password"
- "password"
- "password" AND "police"
- "cracking passwords"

The found articles' abstracts were analyzed to filter out articles relevant to the research topic; user strategies for the creation of biographical passwords. Relevant articles were read in their entirety and analyzed to identify what biographical elements are commonly used and categorize that data thematically. The articles were also analyzed in order to find patterns as to how common the elements are and how passwords created with such data are usually structured. Based on the findings, an initial model was created, which was used as input for the interviews.

2.2. Semi-structured Interviews

To improve the initial model the study utilized semi-structured individual interviews with three forensic experts from the Swedish police and the Swedish National Forensics Center (NFC). The experts were purposefully selected as they encounter encryption as part of their daily work. The purpose of the interviews was to validate and improve the model based on the experience of field experts. It was deemed that three interviews, independent of each other, would sufficiently cover the potentially different experiences of forensic practitioners to aid the purpose of this study. The interviews were conducted in a semi-structured way utilizing open-ended questions. Thus the interviewees were invited to speak freely about the subject and add their own opinions. The interviews were recorded and transcribed in their entirety. The transcribed material was then analyzed and categorized using a thematic approach. The changes consequently led to a revised model, which was, again, presented to the interviewees. This iterative process continued until the interviewees all considered the model complete.

3. Results and Discussion

This chapter presents the results of the research process. First, the results from the selective literature review are presented, and then the opinions from the interviewees and how they revised the model are presented. Finally, the resulting model for creating a dictionary based on biographical data is presented.

3.1. Selective Literature Review

This section presents the results of the literature analysis. First, the components commonly used when creating passwords based on biographical data are presented. Then, the order in which they are commonly used is discussed. The publications identified during the search and selection process are cited throughout this section.

Names of different types are common components of passwords. Of approximately six million leaked passwords analyzed by [9], more than 25% included names. [5] analyzed the passwords of 100 participants of different groups (students, IT professionals, and the general population), where 67% of the students and 42% of the other groups used names in their passwords. [10] present a summary of a previous study consisting of 1200 participants where 26% of the passwords contained names or nicknames. In their own study, [10] found that 32% of 218 students' passwords contained their own, a pet's, partner's, or relative's name.

The analyzed literature [10][11][12] shows that multiple kinds of geographical and positional data may be included in passwords. The identified geographical and positional entities are found in Table 1.

Requirements on password complexity commonly force users to incorporate numbers in their passwords. Consequently, users also tend to incorporate numbers that have some kind of personal meaning. [13], [9] and [14] find that birth dates are common when combining letters and numbers as well as personal identification numbers and phone numbers. It is also common to use a friend's or relative's phone number or birth date [14].

Almost all words that have some kind of meaning to the user may be used in password creation [15]. [10] found that almost a third of their analyzed passwords contained entities that can be categorized as *Hobbies & Interests*, such as athletic teams, celebrities, fictional characters, and sports. [16] also mentions favorite foods as potential password elements. Further, [11] suggests that it is reasonable to include, for example, the entire collection of the Lord of the Rings books if the user has

such an interest. The identified interests and hobbies found are presented in column 3 of Table 1, which presents and overviews the biographical entities found in the literature.

Table 1
Identified biographical components

Names	Geographical Information	Hobbies & Interests	Numbers & Dates
Personal names	Addresses	Athletic teams	Year of birth
Relatives and friends	Cities	Celebrities	Date of birth
Partners	Countries	TV shows	Personal identification number
Pets	Locations	Fictional characters	Phone numbers
Other name related entities (e.g. initials, nicknames)	Regions	Other interests and hobbies	Related persons' numbers & dates
		Foods	

Figure 1 introduces the order in which the entities should be tested. The literature says very little about how the different entities should be prioritized. Apart from the category names, which the literature identifies as the most commonly used [13] and is therefore placed first, the order is based on the researchers' interpretation of how often they occur in the literature. Also, as described by [17], most passwords are created using alphanumeric characters (both letters and numbers), and it is, therefore, reasonable to assume that users utilize important dates and numbers in conjunction with the other categories. Consequently, the entities in the category dates and numbers should be tested together with the other categories, as illustrated in Figure 1.

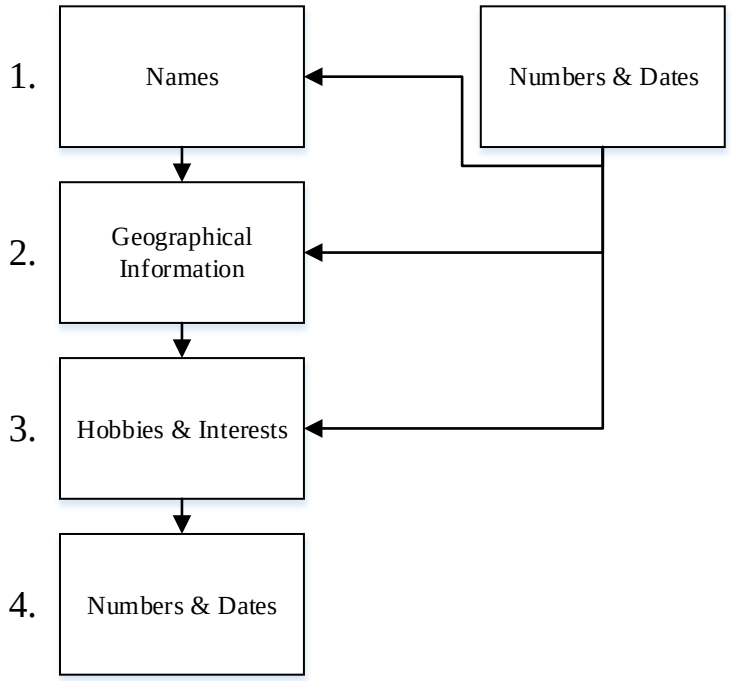

Figure 1: Initial prioritized order of categories

3.2. Semi-structured Interviews

Semi-structured interviews were conducted with the Swedish police and the Swedish National Forensics Center (NFC) for the purpose of refining the initial model. The results of the analyzed interviews are presented below.

Interview 1 was conducted with an IT forensics expert from the Swedish police. The categories and the biographical components from the initial model are acknowledged and cover what the interview subject would currently use to generate a dictionary. Improvement suggestions are to include current and previous workplaces and schools in the category *Geographical Information*. Regarding the category order, the interview subject believes it is more common to use interests and hobbies than geographical information. Thus the category *Hobbies & Interests* should be tested prior to *Geographical Information*. The interview subject also acknowledges the utilization of *Numbers & Dates* in conjunction with the other categories as commonplace for password creation.

Interview 2 was conducted with an IT forensics expert working with the Swedish police. The interview subject recognizes the current categories and their elements as highly relevant as they have been a part of almost all biographical passwords cracked or encountered. Suggested refinements are to include books and historical events in the category *Hobbies & Interests*. As an example, the title or the first word of a book is not uncommon. The final suggestion is to include the element "other" in all categories. This is to encourage the investigator using the model to feel free to include other information relevant to the category. The most commonly used biographical password is, according to the interview subject, a person's name in conjunction with a birth date. Thus the category *Names* is very common and should be tested first. In an attempt to distance the password from themselves, users create passwords based on interests and hobbies more often than geographical information and dates or numbers, making the *Hobbies & Interests* category prioritized. The model should also include the transformation of passwords using leetspeak and special characters. Other opinions expressed by the interview subject are that many passwords are believed to be cracked using the categories and elements from the model and that the prioritization would complete the process more efficiently. Finally, the interview subject underlines the importance and usefulness of the model from a forensic standpoint.

The third interview was conducted with a person from the Swedish National Forensics Center (NFC) working with re-creation of passwords, encryption, and digital forensics. Regarding the categories, the elements nicknames and internet aliases should be added to the category *Names*. It was also pointed out to include the mother's maiden name as it is common for the mother to change name when married, which makes the maiden name less traceable to the targeted person. Further, the category *Numbers & Dates* should be changed to *Important Numbers* to reflect a more general view of what numbers could represent and include vehicles' license plate numbers and postal codes. *Geographical Information* should include vacation locations and user origins. If the targeted person originates from another country, that country and geographical areas in that country are possible elements. Related to the geographical information, one should also consider other languages that the targeted person may be familiar with and use the translated versions of gathered information. The category *Hobbies & Interests* should be updated with the element paraphilias and vehicles, which could include vehicle types, brands, and models. Regarding the order, the interview subject suggests that *Geographical Information* and *Hobbies & Interests* should switch places as hobbies and interests are more personal and thus more likely to be used as a password. It is also recommended to combine all categories with each other to create as many passwords as possible and then apply modifications such as leetspeak or special characters. The interview subject explicitly pointed out the importance of having a historical perspective of all the categories. This means to keep in mind what the elements could have been historically, for example, a childhood friend or the phone number of the local pizza place where the targeted person grew up or studied. To not make the user of the model miss valuable information, the elements should be open and general to be as inclusive as possible. Finally, utilizing a biographical dictionary to crack passwords is considered very successful as most passwords turn out to be based on biographical information.

Based on the information gathered from existing research and the interviews, a revised model is presented in Figure 2.

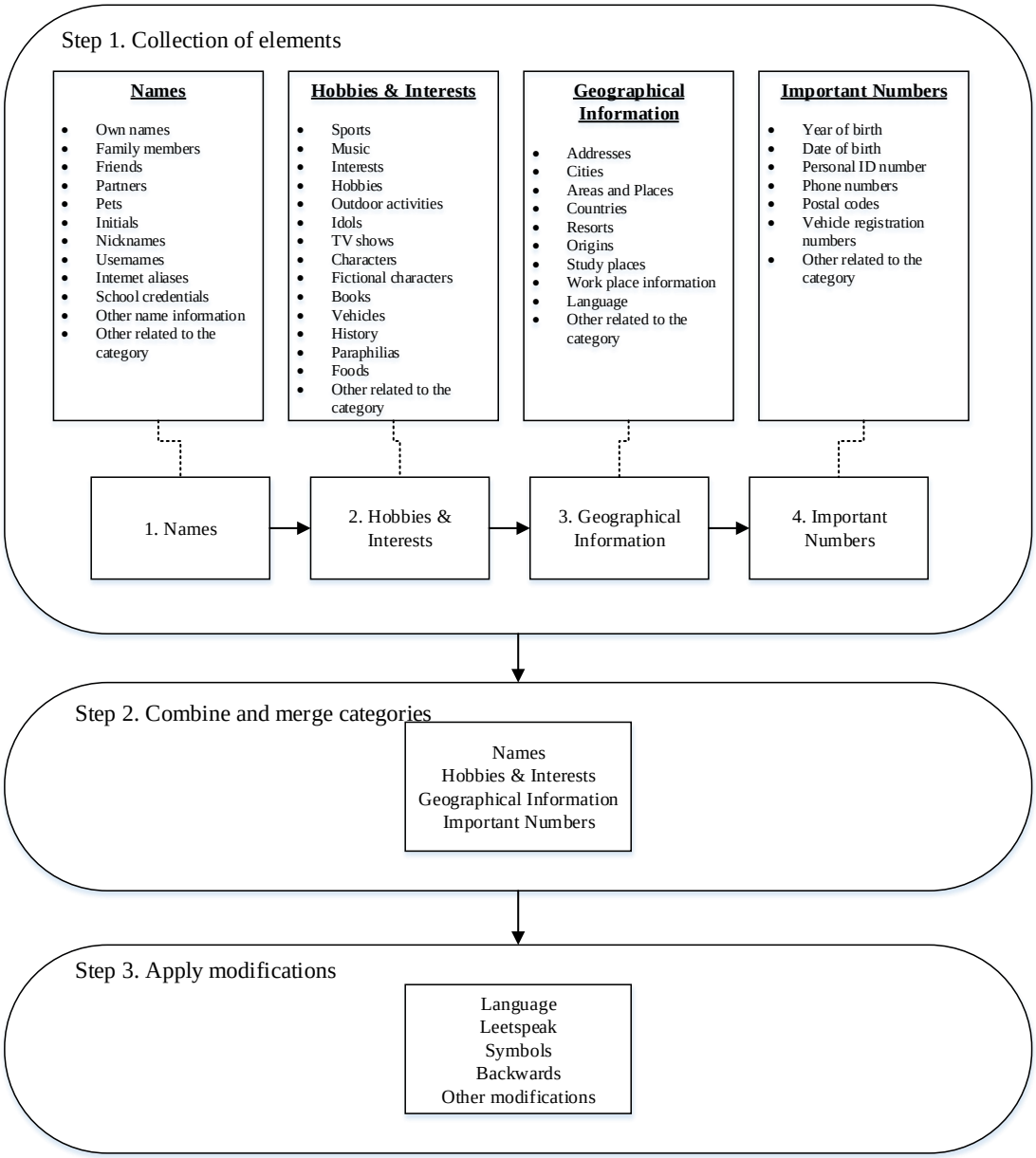

Figure 2: Final model

Apart from the added elements, the final model also reflects structural changes recommended from the interviews. The final model now includes combining all categories with each other (step 2), which replaces combining just numbers and dates with all other categories. The interviewees also agreed that *Hobbies & Interests* are more common than geographical information, which is now reflected in the final model. The revised model was sent to the interviewees for assessment and received satisfactory responses.

4. Conclusions

This research reviewed the literature for elements commonly included when creating passwords that are based on biographical information. Based on current research, an initial model was created, which was then refined through semi-structured interviews with practitioners from the field. The interviewees deemed the final model to be useful and likely to be successful as passwords, in many cases, are based on personal information.

We present a ready to use model to support the creation of biographical dictionaries used when cracking passwords. The model is based on previous research and practical knowledge of what elements

are used in biographical passwords and in what order they are most often used. Consequently, dictionaries resulting from the use of the model will contain likely passwords ordered by probability. The resulting dictionary will thus contain the most commonly used elements in an order that could improve efficiency.

The model suggests that information in four different categories should first be collected, then combined with each other, and finally modified using various techniques (e.g. leetspeak). In addition to the three-step process, the model also presents commonly used elements and the order in which they should be tested based on previous research and the experience of professional practitioners. The list of biographical elements can thus serve as a guide for what information that should be collected during a forensic investigation to increase the chance of cracking upcoming passwords. Our model could aid law enforcement when passwords need to be cracked, consequently collecting evidence in a more efficient and time-saving manner. This research also highlights common password behavior that may be taken into consideration by system administrators when creating password policies. The results are, in that regard, a summary of current research around password behavior.

To be successful, the user of the model should also consider the languages known to the targeted person and have a historical perspective, ultimately including both present and past versions of the elements in multiple languages. It should also be noted that the model is not only useful for cracking passwords used for encryption but also to crack passwords to gain access to devices, accounts, and systems in order to collect more evidence, regardless of encryption, where the legislation allows.

Finally, using the model is likely to result in more cracked passwords and potentially solving crimes. However, the model is not mutually exclusive to other methods, and we want to highlight the importance of using other tools and methods as well to be successful in the field of digital forensics.

5. Future Work

This study provides the research community with additional insight into password creation. To validate the effectiveness, a practical evaluation of the model is necessary. This could be achieved using a collection of cracked passwords that are known to be created using biographical data and see whether information collected based on the model would generate those passwords. The effectiveness should also be compared to other methods of cracking passwords, such as using lists of leaked passwords or brute force attacks. Furthermore, this study was conducted from the perspective of Swedish forensic experts. Similar studies could be conducted to investigate potential differences in password behavior between countries and cultures to further improve the model.

6. References

[1] A. Kanta, I. Coisel, and M. Scanlon, A survey exploring open source Intelligence for smarter password cracking, Forensic Science International: Digital Investigation 35 (2020). doi:10.1016/j.fsidi.2020.301075.

[2] S. Belshaw, and B. Nodeland, Digital evidence experts in the law enforcement community: understanding the use of forensics examiners by police agencies, Security Journal 35 (2021) 248-262. doi:10.1057/s41284-020-00276-w

[3] Forensicfocus.com, Current Challenges in Digital Forensics, 2016. URL: https://www.forensicfocus.com/articles/current-challenges-in-digital-forensics/.

[4] R. Montasari, and R. Hill, Next-Generation Digital Forensics: Challenges and Future Paradigms, in: 2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3), 2019, pp. 205-212. doi:10.1109/ICGS3.2019.8688020

[5] R. Alomari, and J. Thorpe, On password behaviours and attitudes in different populations, Journal of Information Security and Applications 45 (2019) 79–89. doi:10.1016/j.jisa.2018.12.008

[6] E. T. Stringer, Action Research, 4th. ed., SAGE Publications, CA, USA, 2014.

[7] R. Yin, Qualitative Research from Start to Finish, Guilford Publications, NY, USA, 2011.

[8] H. Snyder, Literature review as a research methodology: An overview and guidelines, Journal of Business Research 104 (2019) 333-339. doi:10.1016/j.jbusres.2019.07.039

[9] M. AlSabah, G. Oligeri, and R. Riley, Your culture is in your password: An analysis of a demographically-diverse password dataset, Computers and Security 77 (2018) 427–441. doi:10.1016/j.cose.2018.03.014.
[10] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas, Generating and remembering passwords, Applied Cognitive Psychology 18 (2004) 641-651. doi:10.1002/acp.1014
[11] J. Kävrestad, Fundamentals of Digital Forensics – Theory, Methods and Real-life Application, Springer, Switzerland.
[12] K. Al-Wehaibi, T. Storer, and W. B. Glisson, Augmenting password recovery with online profiling, Digital Investigation 8 (2011) 25–33. doi:10.1016/j.diin.2011.05.004.
[13] K. Renaud, R. Otondo, and M. Warkentin, "This is the way 'I' create my passwords" … does the endowment effect deter people from changing the way they create their passwords?, Computers and Security 82 (2019) 241–260. doi:10.1016/j.cose.2018.12.018
[14] G. B. Duggan, H. Johnson, and B. Grawemeyer, Rational security: Modelling everyday password use. International Journal of Human Computer Studies 70 (2012) 415–431. doi:10.1016/j.ijhcs.2012.02.008
[15] R. Alomari, M. V. Martin, S. MacDonald, A. Maraj, R. Liscano, and C. Bellman, Inside out - A study of users' perceptions of password memorability and recall, Journal of Information Security and Applications 47 (2019) 223–234. doi:10.1016/j.jisa.2019.05.009
[16] F. Ghiyamipour, Secure graphical password based on cued click points using fuzzy logic, Security and Privacy, 4 (2021). doi:10.1002/spy2.140
[17] V. Zimmermann, and N. Gerber, The password is dead, long live the password – A laboratory study on user perceptions of authentication schemes, International Journal of Human Computer Studies 133 (2020) 26–44. doi:10.1016/j.ijhcs.2019.08.006