Analyzing User Password Selection Behavior for Reduction of Password Space

Roman V. Yampolskiy
Department of Computer Science and Engineering and IGERT in GIS
University at Buffalo, Buffalo, NY 14260, USA
rvy@buffalo.edu

Abstract- This paper presents a comprehensive survey of recent literature on the topic of password dictionaries for alphanumeric and graphical user authentication approaches including some password schemas proposed by the author. After different methods used for reduction of password space are introduced, they are analyzed and compared with the intent of finding a common flaw of user authentication mechanisms, which allows for the development of such password dictionaries by hackers. Our conclusion is that any user authentication system, which allows users to exercise choice in selection of their passwords, is vulnerable to the password space reduction methods presented in this paper and so should not be utilized.

Index Terms — authentication, password, password space reduction.

I. INTRODUCTION

In the past few years a lot of research went into analyzing and classifying user's choices of passwords [1-5]. Such research has benefits ranging from better understanding of personality types [5], to increasing security and reliability of computer systems and networks. Passwords are based on secret, often-personal information, which is frequently intrinsically linked to the person generating it. Clever hackers are learning to utilize such a connection to improve their chances of breaking into computer systems as well as increasing computational resources available to them [6, 7]. In order to combat such personalized attacks we need to better understand what types of password dictionaries currently exist, and which can appear in the near future. Such knowledge can allow us to improve our security procedures and perhaps develop superior novel user authentication mechanism resistant to dictionary based attacks [8].

II. PASSWORD SPACE REDUCING TECHNIQUES

The following sections describe different tendencies of user's in various authentication mechanisms, which can be utilized to reduce the overall password space, making the mentioned password systems unreliable for system protection.

A. Alphanumeric Passwords

It has been shown [5] that user's passwords can be grouped into four broad categories: family oriented, fans, fantasies, and cryptic. "Family oriented" users, which comprise 47.5% of users, select their own name or last name or other personal information such as pet or child's name as their password. Those are usually less experienced computer users. "Fans" utilize celebrities as passwords and comprise another 32% of users. Those are often younger computer users. "Fantasy" group relies on expressing their imagination and desire by selecting passwords such as "hot" and is comprised of about 11% of users. Remaining 9.5% of users are sophisticated computer users who understand security and tend to select truly difficult to guess random passwords. It is the first three groups, which are a target of concentrated dictionary attacks. The rest of this section goes into particular details of passwords, which can be found in those categories.

A lot of research has gone into identifying commonly used alphanumeric passwords [9-13]. Large dictionaries composing many possible passwords are freely available on the Internet. Those dictionaries taken together with personal information about the user are often sufficient to relatively quickly break between 20 and 30 percent of all passwords used in a given system by simple trial and error approach. By analyzing different types of passwords found in different systems the following suggestions can be made to attempt to reduce the passwords search space for alphanumeric-textual authentication approaches [1]:

- First try all the available personal information about the user such as: "one's names and initials; One's account name; Names of immediate family members; Names, breeds or species of pets; One's birthday; Family member's birthdays; One's vehicle make, model, year; Hobbies, interests and related words; One's job title; Employer's name; Job related words; Friend's names; Street numbers or names, city, county, state or zip code for home, work, family or friends; Phone numbers for home, work, family or friends; Social security numbers for self and immediate family; License plate numbers; Birthplace including street address; University or college name; College major; High school name; Student or employee

1-4244-0174-7/06/$20.00 ©2006 IEEE

ID numbers; Serial numbers from consumer products"

- Words from a number of dictionaries including technical and professional dictionaries
- List of personal names
- Listing of geographical terms such as countries, cities, lakes, rivers, etc.
- List of celebrity names, movie stars, writers, scientists, philosophers, athletes, sports team names
- Listing of characters from books, movies, cartoons, plays, mythology, religious texts
- Different numbers both spelled out and as numerals, repetitive strings of characters, common keyboard patterns,
- Short common phrases, word pairs and triples
- Common abbreviations and mnemonics
- For foreign users try a dictionary of foreign words in English transliteration.

TABLE 1
AVERAGE LENGTH OF PASSWORDS

Length	Percentage [14]	Percentage [1]
1	0.4%	0.1%
2	0.6%	0.2%
3	1.53%	2.0%
4	3.25%	5.7%
5	9.86%	9.5%
6	22.01%	34.7%
7	21.15%	24.4%
8	41.86%	23.4%

For each type of password guesses given above a number of variations should also be tried such as: "append or prepend defined characters to a word; Reverse a word; Duplicate a word; Append the reversed word; Rotate a word either left or right, i.e. move the first letter to the end or the last letter to the front; Upper case a word; Lower case a word; Make only the first letter a capital; Male all but the first letter a capital; Toggle the case of all characters; Toggle the case of a character at a set position; Minimum and maximum word lengths can be set or long words can be truncated at a set length; Suffixes may be added to words; First, last or any specific character may be deleted; Characters can be replaced at a set location; Characters can be inserted at a set location; "Shift" the case, i.e. substitute the other character on the same key, e.g. 'a' and 'A' or '5' and '%'; Shift the characters left or right by keyboard position (so an 's' becomes an 'a' or 'd'); Replace all of one character with another; Replace all characters of a class (for example vowels, letters, non letters, digits) with a specific character; Remove all occurrences of any character from a word; Remove all characters of a class from a word. [15]"

Additional analysis of known passwords can allow for classification by frequency of passwords with regard to types described above. In terms of the character length Table 1 demonstrates password distribution which has been experimentally observed by [14] and [1] respectively.

With respect to the password makeup in terms of constituting characters the following statistics have been calculated:

TABLE 2
PASSWORD MAKEUP

Characters	Percentage
Lower-case only	28.9%
Mixed Case	38.1%
Some upper-case	40.9%
Digits	31.7%
Meta-characters	0.2%
Control characters	1.4%
Space and/or tab	4.1%
. , ;	6.1%
- _ = +	1.6%
!#$%^&*()	4.7%
Other non-alphanumeric characters	1.7%

In case of passwords based on the words in a foreign language the following distribution was observed [14]:

TABLE 3
PASSWORDS BY LANGUAGE OF ORIGIN

Language	Percentage
Australian/Aboriginal	.96%
Danish	2.8%
Dutch	2.2%
English	13%
Finnish	7.7%
French	2.6%
German	2.8%
Italian	7.9%
Japanese	4.5%
Norwegian	2.6%
Swedish	1.9%

Some languages were not considered which in environments rich with foreign speakers might be particularly beneficial such as trying Asian (Chinese, Korean), Indian and Russian languages in systems frequently used by computer scientists.

Other types of easily cracked passwords can be summarized as follows [1]:

TABLE 4
PASSWORDS BY TYPE OF INFORMATION

Type of Password	Percentage
User/account name	2.7%
Character sequence	.2%
Numbers	.1%
Chinese	.4%
Place names	.6%
Common names	4.0%
Female names	1.2%
Male names	1%
Uncommon names	.9%
Myths & Legends	.5%
Shakespearean	.1%
Sports teams	.2%
Science fiction	.4%
Movie and actors	.1%
Cartoons	.1%
Famous people	.4%
Phrases and patterns	1.8%
Surnames	.1%
Biology	.007%
/usr/dict/words	7.4%
Machine names	1%
Mnemonics	.014%
King James bible	.6%
Miscellaneous words	.4%
Asteroids	.1%

B. PassFaces

A system such as Passfaces [16] has an inherently small password space which creators claim is sufficient since no known dictionary exists for such type of password making further reduction of password space impossible. However, recent research [17] suggests otherwise.

While no pre-computed dictionary currently exists for Passfaces user preferences can be used to greatly reduce the size of search space in face recognition based authentication systems. It has been demonstrated that faces selected are affected by the race of the user and a strong preference is shown for attractive faces. Also, all users show preference for female faces. In case of male users preference for selecting attractive female faces is so strong that it makes password space manually searchable for Passface-like systems [17].

By knowing some demographic information about the user hacker can greatly reduce password space he has to search. If a user is male the search space for the worst 10% of passwords is equal to two. In case of Asian users of known gender search space is just one for the easiest 10% of passwords! Figures below demonstrate just how extreme user's facial biases can be [17].

TABLE 5
FACIAL PREFERENCES BY GENDER

User	Female Model	Male Model	Typical Female	Typical Male
Female	40.0%	20.0%	28.8%	11.3%
Male	63.2%	10.0%	12.7%	14.0%

Users of both genders tend to select female faces much more frequently (68% for females and 75% for males). Males also almost 5 times more likely to select attractive females as apposed to average looking ones.

TABLE 6
FACIAL PREFERENCES BY RACE

User	Asian	Black	White
Asian female	52.1%	16.7%	31.3%
Asian male	34.4%	21.9%	43.8%
Black male	8.3%	91.7%	0.0%
White female	18.8%	31.3%	50.0%
White male	17.6%	20.4%	62.0%

As far as race was concerned users tended to selected faces corresponding to their own race. Asian females and White females did so 50% of the time, White males 60% of the time and Black males 90% of the time.

C. Story

In the story authentication approach a user has to select a sequence of themed images [18]. The differences between males and females are not extreme, but still statistically significant. Females tend to choose animals twice as often as males do, while males show preference for choosing pictures of women twice as much. Other less significant differences are presented in the Figures below and show how the two genders and three different races differ in terms of preferences for different themes.

TABLE 7
TOPIC PREFERENCES BY GENDER

User	Female	Male
Animals	20.8%	10.4%
Cars	14.6%	17.9%
Women	6.3%	13.6%
Food	14.6%	11.0%
Children	8.3%	6.8%
Men	4.2%	4.6%
Objects	12.5%	11.0%
Nature	14.6%	17.2%
Sports	4.2%	7.5%

TABLE 8
TOPIC PREFERENCES BY RACE

User	Asian	Hispanic	White
Animals	10.7%	12.5%	12.5%
Cars	18.6%	12.5%	16.8%
Women	11.4%	25.0%	13.0%
Food	11.4%	12.5%	11.5%
Children	8.6%	0.0%	6.3%
Men	4.3%	12.5%	11.5%
Nature	17.1%	12.5%	11.1%

Utilizing the demographic information presented above for the easiest 10% of passwords, which belonged to Asian males, it was shown to be possible to break the Story authentication mechanism in just twenty attempts [18].

D. Draw-a-Secret

Recent investigation of types of drawings users tend to select as their draw-a-secret passwords revealed some common properties which can be taken advantage of by a clever hacker in order to reduce password space of a draw-a-secret approaches [19, 20]. Drawings can be classified into three groups based on the following characteristics: global symmetry, number of strokes and location within the grid.

TABLE 9
SYMMETRY FOR DRAW-A-SECRET PASSWORDS

Vertical Reflective	19%
Horizontal Reflective	8%
Diagonal Reflective	4%
Total Reflective	31%
Rotational	7%
Repetitive	7%
Total Symmetric	45%
Total Asymmetric	55%

TABLE 10
NUMBER OF STROKES IN A DRAW-A-SECRET CODE

1-3 strokes	4-6 strokes	> 6 strokes
80%	10%	10%

TABLE 11
LOCATION OF A D.A.S. PASSWORD ON A GRID

Centered	Approximately Centered	Not Centered
56%	30%	14%

As can be see from the above figures, 45% of users tend to choose symmetric passwords, with 66% of those being reflective. A large majority of users (80%) have a relatively short password of 1 to 3 strokes and another 56% of users locate their drawings in a center of a grid or almost the center for an additional 30% of users [19].

E. PassPoints

In an investigation of the PassPoints system it has been demonstrated that accurate recollection of the password is strongly reduced if a small tolerance region is used around the user's password points. But if a large region is used the password space of PassPoints is being reduced. Additionally it was established that not all images are suitable as PassPoints graphics. In particular images with few memorable points such as images with large expenses of green grass or overly complicated images should be avoided. One reason being that large regions with little character such as blue sky can be safely eliminated as potentially containing pass points. The opposite is also true, a really memorable region such as a small boat in an ocean is very likely to be chosen by the users as the pass point. So the overall password space of PassPoints can be greatly reduced by not considering large fields with monotonous information, and concentrating on potential regions of interest to a user, such as faces, outlier objects based on color or size and other easy to remember sub-parts of an image [21].

F. Pronounceable Passwords

Random pronounceable passwords are automatically generated by computer programs to assist users in obtaining a secure password, which is also relatively easy to remember. It has been shown by Ganesan et al. [22] that based on a particular implementation of a random pronounceable password generator it might be possible to greatly reduce the overall password space.

While different passwords have an equal probability of being generated in each of the possible categories, because of the additional knowledge about distribution of vowels, consonants, etc. in pronounceable words it is possible to select a category with a small overall number of possible passwords, but which nonetheless has the same total number of passwords as all the other categories. As a result average density of passwords per unit of password space in such a sub-space is extremely

high. Unlike in other password dictionaries hacker does not generate a list of likely passwords, but rather determines a relatively small region in a password space, which is likely to contain many user passwords. So while the overall password space of the generator may be practically un-searchable, a small sub-space may be a fruitful ground for a brute force attack [22].

As long as the objective of the hacker is not to break into a particular account but into any account in the system this approach works extremely well, to the point there continuous use of password generators of this type is not recommended by authors of the study [22].

G. PassText

A recently developed system proposed by the author in [23] relies on manipulation of free-form text by the user. It has been investigated with respect to common tendencies of users. Results similar to those from investigation of alphanumeric passwords have been obtained. Users typically added personal information such as name or date of birth in case they chose to add text. As far as removing parts of text, the actual text being removed depended on the text itself, but easy to remember locations were often selected, such as title of the text, and first or last sentence of the text.

H. PassMap

Another user authentication mechanism proposed by the author [24] is based on selection of routes in a map. It has been put through a preliminary testing and while the set of test users was limited to just a dozen, it was shown that users tended to select locations, which they have visited or wanted to visit. Particularly places of birth showed a high degree of being selected. Also famous locations were chosen more often then less known ones.

III. CONCLUSIONS

Users have long been considered the weakest link of any security system [25-28]. Across different authentication schemas users tend to choose passwords, which are easy to remember, and so are easy to guess. In addition users show strong preference for egocentric passwords based on their own names in case of alphanumeric passwords, faces of the same race as they are in case of face-recognition based authentication schemas and personal information in case of PassText or PassMap systems. Table 10 summarizes the types of information which is used in order to significantly reduce the size of a password space for a given user authentication system.

TABLE 12
APPROACHES TO PASSWORD SPACE REDUCTION

Authentication Schema	Password space reduction based on:
Alphanumeric passwords	Personal and public well known information, which is already remembered by the user
PassFaces	Faces of people who are same race as the user, good looking faces
Story	Variations in gender preferences
Draw-a-Secret	Symmetry, number of strokes, in grid location
PassPoints	Dismiss large uniform areas, concentrate on regions of interest
Pronounceable Passwords	Phonetic rules of user's language
PassText	Personal information, outlier locations
PassMap	Familiar locations

By analyzing information presented in Table 10, we can suggest a way of reducing password space size for any currently existing user authentication mechanism as well as for any such future system. This is true as long as the system allows the user to select the password as apposed to assigning one randomly selected from the full set of possibilities by the system itself. The general approach for alphanumeric or graphical passwords is to utilize demographic information about the user including but not limited to personal information, race, gender, age, and interests. Such information was already known and remembered by the user prior to the enrolment with the authentication system. So it requires no additional effort to commit to memory and as a result is provided as the password.

In case a user has to select among multiple types of information, to be used as his password, it should be assumed that the user would select the easiest to remember, most symmetric, less complicated object with fewest possible number of components. In a field of many similar items users will select the most distinguished items either in terms of color or shape or some other property.

Users tend to select passwords, which are relatively easy to remember. As we showed in this paper for any authentication system a diligent hacker can be expected to be successful at taking advantage of the reduced password space. Assigning passwords to users can solve this problem, but this in turn results in forgotten passwords and increased costs of system administration. This motivates as to suggest a move towards biometrics-

based user authentication systems as the only solution for secure and user-friendly person identification.

IV. ACKNOWLEDGEMENTS

This paper is based upon work supported by National Science Foundation Grant No. DGE 0333417 "Integrative Geographic Information Science Traineeship Program", awarded to the University at Buffalo.

V. REFERENCES

[1] D. V. Klein, "Foiling the cracker: A survey of and improvements to password security," presented at USENIX Conference Proceedings, 1990.

[2] D. C. Feldmeier and P. R. Karn, "UNIX Password Security - Ten Years Later," presented at CRYPTO, Available at: citeseer.ist.psu.edu/188968.html, 1989.

[3] M. Bishop, "Proactive Password Checking," presented at 4th Workshop on Computer Security Incident Handling, Available at: citeseer.ist.psu.edu/bishop92proactive.html, August 1992.

[4] R. Morris and K. Thompson, "Password Security: a Case History," presented at CACM, 1979.

[5] B. J. Brown and K. Callis, "Computer Password Choice and Personality Traits Among College Students," Available at: http://cstl-cla.semo.edu/callis/xResearch /PasswordsBettyBrown/PasswordsRevs5.30.04.doc, Retrieved December 12, 2005.

[6] A. Narayanan and V. Shmatikov, "Fast dictionary attacks on passwords using time-space tradeoff," presented at Conference on Computer and Communications Security archive Proceedings of the 12th ACM conference on Computer and communications security, Alexandria, VA, USA, 2005.

[7] T. Perrine and D. Kowarch, "Teracrack: Password cracking using TeraFLOP and PetaByte Resources," Available at: http://security.sdsc.edu/publications/teracrack.pdf, Retrieved December 15, 2005.

[8] M. Bishop, "Comparing Authentication Techniques," Available at: citeseer.ist.psu.edu/bishop91comparing.html, Retrieved December 15, 2005.

[9] C. Blundo, P. D'Arco, A. D. Santis, and C. Galdi, "Hyppocrates: A New Proactive Password Checker," presented at The Journal of Systems and Software, 2004.

[10] M. Bishop and D. Klein, "Improving System Security Through Proactive Password Checking," presented at Computers and Security 14 (3), May/June 1995.

[11] E. Spafford, "Opus: Preventing Weak Password Choices," presented at Computers and Security, Available at: citeseer.ist.psu.edu/spafford91opus.html, May 1992.

[12] R. E. Smith, "The Strong Password Dilemma," presented at Authentication: From Passwords to Public Keys, 2002.

[13] G. C. Kessler, "Passwords - Strengths and weaknesses," presented at Internet and Internetworking Security, Available at: http://www.garykessler.net/library/password.html , January 1996.

[14] E. Spafford, "Observing Reusable Password Choices," Available at: citeseer.ist.psu.edu/spafford92observing.html, Retrieved November 3, 2005.

[15] G. Shaffer, "Good and Bad Passwords How-To," presented at GeodSoft, Available at: http://geodsoft.com/howto/password/, Retrieved December 12, 2005.

[16] R. U. Corporation, "The Science Behind Passfaces," presented at Real User, Available at: http://www.realuser.com/, June 2004.

[17] F. Monrose, M. K. Reiter, and S. Wetzel, "Password Hardening based on Keystroke Dynamics," presented at International Journal of Information Security, 1(1):69--83, 2001.

[18] D. Davis, F. Monrose, and M. K. Reiter, "On user choice in Graphical Password Schemes," presented at In Proceedings of the 13th USENIX Security Symposium, San Diego, August 2004.

[19] D. Nali and J. Thorpe., "Analyzing User Choice in Graphical Passwords.," presented at Tech. Report TR-04-01, School of Computer Science Carleton University, Canada, 2004.

[20] J. Thorpe and P. v. Oorschot, "Graphical Dictionaries and the Memorable Space of Graphical Passwords," presented at 13th USENIX Security Symposium.

[21] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon, "Authentication using graphical passwords: effects of tolerance and image choice," presented at ACM International Conference Proceeding Series; Vol. 93, Proceedings of the 2005 symposium on Usable privacy and security, Pittsburgh, Pennsylvania, 2005.

[22] R. Ganesan and C. Davies, "A new attack on random pronouncable password generators," presented at In 17th NIST-NCSC National Computer Security Conference, 1994.

[23] R. V. Yampolskiy, "PassText the Latest Step in the Evolution of Passwords," Available at: http://www.acsu.buffalo.edu/~rvy/papers.htm, Retrieved November 4, 2005.

[24] R. V. Yampolskiy, "Better Network Security Via Imroved User Authentication," Available at: http://www.acsu.buffalo.edu/~rvy/papers.htm, Retrieved December 12, 2005.

[25] S. Brostoff, M. A. Sasse, and D. Werich, "Transforming the 'Weakest Link' -- a Human/Computer Interaction Approach to Usable and Effective Security," presented at Technological Journal, July 2001.

[26] D. Weirich, M. A. Sasse, and, "Pretty good persuasion: a first step towards effective password security in the real world," presented at Proceedings of the 2001 workshop on New

security paradigms, Cloudcroft, New Mexico, 2001.

[27] A. B. Jianxin (Jeff) Yan, Ross Anderson and Alasdair Grant, "The Memorability and Security of Passwords -- Some Empirical Results.," presented at Technical Report No. 500, Available at: http://www.ftp.cl.cam.ac.uk/ftp/users/rja14/tr500.pdf, 2000.

[28] M. Hertzum, "Remembering Multiple Passwords by Way of Minimal-Feedback Hints: Replication and Further Analysis," presented at Proceedings of the Fourth Danish Human-Computer Interaction Research Symposium, Aalborg University, Aalborg, DK, November 16, 2004.

VI. VITA

Roman V. Yampolskiy holds an MS in Computer Science degree from Rochester Institute of Technology and is a PhD candidate in the department of Computer Science and Engineering at the University at Buffalo. His studies are supported by the National Science Foundation IGERT fellowship. Roman's main areas of interest are Artificial Intelligence and intrusion detection. Roman has a number of publications describing his research in neural networks, genetic algorithms, pattern recognition and behavioral profiling.