Visualizing Semantics in Passwords: The Role of Dates

Rafael Veras
University of Ontario Institute of Technology
Oshawa, Canada
rafael.verasguimaraes@uoit.ca

Julie Thorpe
University of Ontario Institute of Technology
Oshawa, Canada
julie.thorpe@uoit.ca

Christopher Collins
University of Ontario Institute of Technology
Oshawa, Canada
christopher.collins@uoit.ca

ABSTRACT

We begin an investigation into the semantic patterns underlying user choice in passwords. Understanding semantic patterns provides insight into how people choose passwords, which in turn can be used to inform usable password policies and password guidelines. As semantic patterns are difficult to recognize automatically, we turn to visualization to aid in their discovery. We focus on dates in passwords, designing an interactive visualization for their detailed analysis, and using it to explore the RockYou dataset of over 32 million passwords. Our visualization enabled us to analyze the dataset in many dimensions, including the relationship between dates and their co-occurring text. We use our observations from the visualization to guide further analysis, leading to our findings that nearly 5% of passwords in the RockYou dataset represent pure dates (either purely numerical or mixed alphanumeric representations) and the presence of many patterns within the dates that people choose (such as repetition, the first days of the month, recent years, and holidays).

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Graphical User Interfaces (GUI); E.0 [**Data**]: General

General Terms

Design, Experimentation, Security, Human Factors

Keywords

Visualization, Passwords, Security, Patterns, User Choice

1. INTRODUCTION

Despite decades of password research, we still do not have a good grasp of how people choose passwords. It is well-understood that patterns in user choice exist [4, 12, 13, 21], and they have been characterized in terms of similarity to dictionary words and the types/positions of characters used [21], but the nature and presence of semantic patterns in user-chosen passwords remains somewhat of a mystery. Semantic patterns are important for usability as they might help people remember their passwords; they also have the potential to heavily impact security if the pattern defines a small number of passwords that an attacker can use in a guessing attack. In this paper, we begin a quest towards understanding the semantic patterns behind the passwords that people choose.

As semantic patterns are difficult to recognize computationally, we turn to visualization to aid discovery of interesting semantic patterns in user choice. Understanding such underlying semantic patterns can help us to better understand how people choose the passwords they do, which can help inform usable password policies and password creation guidelines. If we discover semantic patterns that do not lead to security vulnerabilities, they can be used as the building blocks of successful and usable new password guidelines and policies. On the other hand, if any of these semantic patterns do lead to security vulnerabilities, they are still useful as they can be used to help us build stronger and more appropriate password blacklists and proactive checks.

Understanding the semantics of passwords is not an easy task, thus we begin by focusing on one subset: dates and numbers. This is motivated by recent findings which indicate that numbers appear to be commonly used in passwords across language groups, nations, and other population groups [4]. Other recent findings indicate that dates are common amongst 4-digit sequences [5], but do not describe whether (and what) patterns exist within the dates themselves, and whether there are patterns between dates and other text within the passwords.

We found that in the RockYou dataset, which contains over 32 million passwords, over 15% of passwords contain sequences of 5–8 consecutive digits, 38% of which could be classified as a date. This represents significantly more dates than we would expect to parse from a randomly generated set of numbers of the same length. That is, a Chi-squared test showed that the frequency of date-like numbers parsed from 5–8 digit passwords is not equal to the frequency of date-like numbers parsed from a randomly generated set of 5–8 digit sequences ($p<2.2\times10^{-16}$).

We designed an interactive visualization for the detailed analysis of the dates that people choose within large datasets of text passwords. Using this visualization, we discover a number of interesting semantic patterns: (1) repeated days/months are popular, even when the repetition does not

VizSec '12, October 15 2012, Seattle, WA, USA

extend to the year portion of the date, (2) holidays are popular, (3) duplicating a year to create an 8-digit password is common, and (4) when non-digits are paired with dates, they are most commonly single-characters, or names of months.

Our contributions include a new visualization design to visualize date-related patterns in password databases, an analysis of these patterns in the 32-million password RockYou dataset [17], the creation and testing of a guessing dictionary based on these patterns, and a discussion of the security implications of these patterns.

The remainder of this paper is organized as follows. Section 2 discusses related research. Section 3 explains the data set we used and how it was pre-processed for input to the visualization. Section 4 describes the visual and interaction design of our visualization. Section 5 describes interesting patterns that we discovered using this visualization, and Section 6 describes their security implications. We conclude with a discussion of future work in Section 7.

2. RELATED WORK

We discuss the related work from three different research areas that intersect our study. Regarding the problem domain of password patterns, there is extensive literature on password guessing and distribution, although the analysis of password patterns from large, real-world datasets is a recent advent, fueled by the leakage of millions of passwords from popular websites during the last few years. Security data visualization is another relevant area; we focus our discussion on related visualizations of large collections of passwords. Finally, we discuss the works that influenced the design of our visualization, in particular, contributions in the areas of time-oriented data and multiple coordinated views.

2.1 Password patterns

Perhaps the closest to our study, Bonneau and Preibusch [5] approach the guessing difficulty of human-chosen 4-digit PIN by analyzing the patterns in passwords collected through a survey and gathered from the subset of 4-digit passwords in two real datasets, including the RockYou dataset used in this work. They "explain" most of their dataset using a set of patterns, including five different date patterns (*e.g.*, MMDD). Our current research deals with a different subset of RockYou, namely dates that appear to exist in passwords of any length, rather than 4-digit PIN numbers. We focus on passwords characterized by sequences of 5 to 8 digits, which allow a greater variety of date patterns, consequently involving increased ambiguity. While Bonneau and Preibusch [5] mention the presence of overlapping patterns, deep analysis was out of the scope of their investigation; in our present paper, we discuss this challenge in Section 3.2. Additionally, in our work we introduce an exploration of the relationship between text and dates which co-occur in passwords.

Some other related works have made use of the RockYou dataset. For example, Weir et al. [21] use it as a target of attacks in order to account for the effectiveness of password creation rules, showing several statistics about the use of digits in passwords.

Studies involving large samples have also been published recently. Florencio and Herley [12] report on web password habits of half a million users. However, due to privacy concerns that hinder the access to raw passwords, there is no detailed analysis of semantic patterns, such as dates and numerical sequences. The same constraint was imposed to Bonneau [4], as a condition to access password data from 70 million Yahoo! users, the largest sample ever studied. In this paper, it is shown that the distribution of Yahoo! and RockYou passwords are considerably similar, a fact that might alleviate possible concerns regarding the reliability and relevance of leaked data.

2.2 Security data visualization

There is a substantial body of research in visualizing security data, spreading across diverse topics including network security (*e.g.*, [10]), system security (*e.g.*, [18]), and password strength meters (*e.g.*, as used by Google [9]). We do not aim to provide a comprehensive review of security visualization herein; rather we focus on work that visualizes patterns in user choice in text passwords and PIN numbers.

Schweitzer et al. [15] employ a visualization to guide the analytical process of detecting spatial keyboard patterns in passwords. The visualization is used in the early phase to visually detect the main keyboard patterns, preceding their formalization, then further counting and computing statistics independently.

Bonneau and Preibusch [5] present a basic two-dimensional matrix plot that visualizes user-chosen 4-digit PIN numbers. This visualization indicated a preference for numbers that could be dates, but does not show which days are popular (across groups of years), and what words and other symbols may be commonly related to individual dates (and groups of dates). Understanding such details, and consequently more about how people choose dates in their passwords, further motivates the visualization we present in this paper.

2.3 Time-oriented visualization and multiple coordinated views

Several works discuss the properties of time-oriented data and their implications for visualization design [2] [16]. We especially benefit from the "categorization of techniques" [1] and a massive catalog containing over a hundred visualizations of temporal data [3] to specify the requirements and design goals of both our temporal visualizations.

We combine time-oriented and textual views with filtering capabilities in a web-based, multiple coordinated views layout. In a similar way, Visgets [11] combines interactive query visualizations for spatial, temporal and topical aspects of data in a web application.

3. DATA PREPARATION

Starting with raw passwords, our data underwent a series of preparation steps: selecting, parsing, and counting in order to form the dataset used for the visualization.

3.1 Dataset

We make use of a password dataset that was leaked from the social gaming website RockYou in 2009, when hackers took advantage of an SQL injection vulnerability. It comprises over 32 million passwords stored in plain text and contains no user information—at least in the copy we had access to. This last fact implies that we can not base our analysis on demographic assumptions; thus, possible differences in the distribution of dates and use of numerical sequences depending on culture, age or gender are not considered. In particular, we could greatly benefit from birth date information, since it is regarded as having close relation to the choice of dates in passwords. On the other hand, a benefit

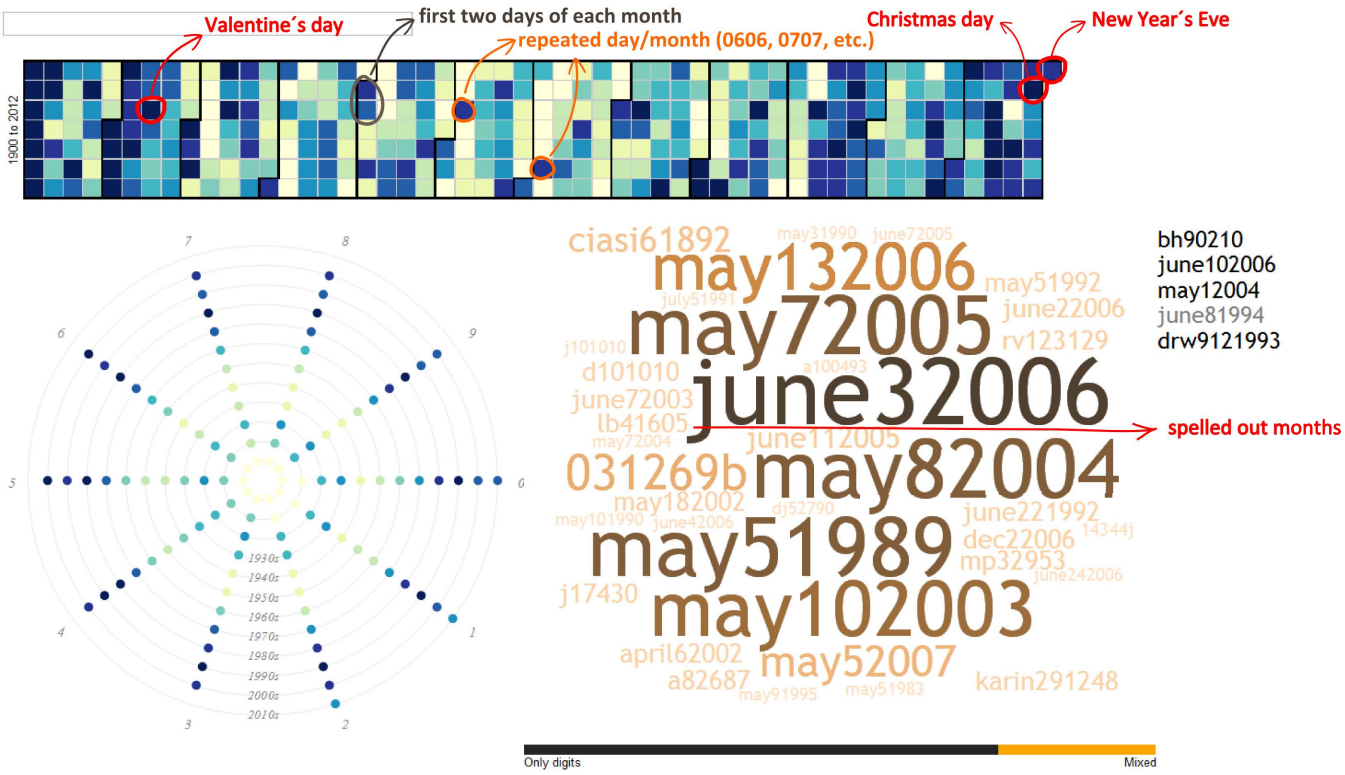

Figure 1: The coordinated visualization interface, with the Tile Map at top, the Radial Plot at left, and the Wordle of raw dates on the right. The Wordle is set to show only mixed numeric and text dates, and all other plots are coordinated to this filtering interaction. On the far right is a list of filtered passwords which the analyst has interactively removed from the analysis. Annotations were manually added to point out interesting patterns.

of using this dataset is that its large size allows us to explore more refined subsets that still contain reasonable password sample sizes.

3.2 Processing

Passwords come in a wide variety of forms. Since our principal goal is to characterize the occurrence of dates, we need to determine what will be considered as such. The everyday use of dates is supported by some important conventions and symbols meant to avoid ambiguity when a compact format is convenient. For example, separators (*e.g.*, '/', '-', '.') are normally used to delimit the elements of a date (year, month, and day); however, perhaps due to historical constraints in some password systems, password creation rules, and factors such as usability, memorability, and even portability—it is easier to re-use them as PINs—, people tend to avoid special characters in passwords.

Not less important, the order of the elements also helps to resolve ambiguity. Notably, the way people use it varies deeply across countries, and is source of confusion even within a single country, as is the case of Canada, where both DD/MM and MM/DD formats are used. Since we do not know the country where a password was issued, deciding between formats naturally becomes a challenge. Furthermore, the presence of leading zeros is also a source of variation and ambiguity. Even considering the separators, the date 01/02/99 can be parsed as February 1, 1999 or January 2, 1999. If we remove the separators and the leading zero (10299), the date February 10, 1999 is also introduced as a possibility.

Since it is clear that information tends to be lost as a result of transforming dates to fit them in passwords, the parsing process is challenging. First, we select all passwords that contain a consecutive sequence of 5 to 8 digits. Passwords containing sequences of less than 5 digits are discarded, even though a date can be represented by 4 digits; we do this because we are only seeking dates which are fully specified with day, month, and year. At this point, the most common numerical sequences are 12345, 111111, 123123, 121212 and 112233, which, intuitively, seem not to represent dates, but "pure" numerical/keyboard patterns (see Appendix A). We remove all sequences that match any of the numerical patterns and some other highly frequent sequences not captured by the patterns.

The next step is to test the sequences against a comprehensive list of date formats (see Appendix B). This list captures a broad range of formats of 5–8 digits without special characters, including variations in use of leading zero. A valid date should match at least one of them and lie between the year range [1900, 2012].

A single password can match several formats, that might translate into different or repeated dates (*e.g.*, 030475 → MMDDYY, DDMMYY → April 3 and March 4, 1975). We

Table 1: Table of statistics of how numbers and dates appear in the RockYou (RY) dataset [17].

Subset Description	# of Passwords	% of RY Passwords
(1) Passwords containing sequences of at least 4 digits	8,056,329	24.72%
(2) Passwords from (1) above that match a numerical pattern (see Section 3.2)	1,346,410	4.13%
(3) Passwords containing 5–8 consecutive digits	4,974,602	15.26%
(4) Passwords that are exactly 5–8 digits (all numeric digits)	3,951,852	12.13%
(5) Passwords containing 5–8 consecutive digits and match a date	1,934,821	5.93%
(6) Passwords that are exactly 5–8 digits and match a date	1,469,662	4.51%
(7) Passwords that contain a date and other text	358,562	1.10%
(8) Passwords that are exactly 5–8 digits, match a date and numerical pattern	114,724	0.35%
(9) Passwords that are exactly 5–8 digits, match a date, no numerical pattern	1,354,938	4.16%

considered different approaches for dealing with this ambiguity when building the frequency distribution of dates. Counting all derived dates as independent events was discarded because it would overrate ambiguous dates. Counting just the first match based on a priority list of formats turned out to be impractical since we don't have solid basis on which to prioritize them. Hence, the most reasonable strategy is to divide the count of a single event between all matched dates. In the aforementioned case, for instance, both dates would receive an increase of 0.5 in their frequency value.

3.3 Testing the Dates Assumption

We performed an experiment to rule out that the matched date sequences in the observed data (RockYou dataset) could be observed by chance.

The experiment was divided in four parts, each corresponding to the sequence lengths considered. For each length, we randomly generated a list containing as many numerical sequences as found in the RockYou dataset. We then run the parsing algorithm over both samples, counting the event of a success (when a sequence is matched by at least one format). Finally, a Pearson's Chi-squared Test is performed to compare the results. The proportion of sequences that contain dates found in the random list corresponds to our expected value. The results show that for all considered lengths, the number of dates found in the RockYou dataset is significantly higher than in the random dataset ($p<2.2\times10^{-16}$]). While this test does not prove that numeric passwords which match date patterns are indeed intended to be dates, it does present intriguing evidence that the passwords may indeed represent dates, thus the semantic patterns of dates is of interest for further study using visualization.

3.4 Discussion of Numbers in Passwords

Overall, the RockYou dataset contains over 32 million passwords, approximately 25% of which contains a sequence of 4 or more digits. Of these sequences of at least 4 digits, approximately 62% contain 5 to 8 digits (which can represent a full date consisting of a month, day, and year).

Table 1 summarizes some interesting statistics on this password dataset. When we match the sequences of 5–8 digits against our date patterns, we notice that they can explain 38% of such sequences. Dates appear to be more popular in sequences that are completely composed of digits: of the sequences that contain a date pattern, 75% are entirely numerical digits. Of all passwords that are solely composed of digits, 37% match date patterns (or 34% when we remove the ones that may be due to a numerical pattern).

4. VISUALIZATION

To approach the problem of verifying whether dates really do play a significant role in passwords, and if so, discovering whether there are patterns of dates, or specific dates which stand out, we designed an interactive visualization to explore the dataset. We took a coordinated multiple views approach in order to provide several ways to look at the data (see Figure 1). The main goals which guided our design are:

Guide the investigation Drawing sound security recommendations from patterns observed in a dataset eventually requires rigorous statistical treatment; however, data manipulation at a low level is cumbersome and does not favour the exploration of data space necessary in the early stages of an investigation. The role of the visualization in this context is to support quick generation and early testing of hypotheses. It should enable insight on possible patterns and provide quantitative information to help deciding whether or not a statistical experiment is worthy. Thus, the formal procedures are left for validation in the final phase of the investigation.

Facilitate exploration of diverse scenarios The tool should enable one to easily delimit scenarios for investigation of localized patterns. This involves the ability to narrow the scope based on time dimension (*e.g.*, decades, years, days...) and password structure (*e.g.*, presence of a numerical pattern or letters).

Easily accessible We took a rapid-prototyping approach, refining the visualization to respond to the questions raised by every new hypothesis drawn, reflecting our increasing understanding of the data. As a consequence we needed a medium that provides easy and fast deployment of new versions and high accessibility to a distributed team.

4.1 Representation and Interaction Design

We opted for a layout with coordinated views that display the frequency of passwords at multiple aggregation levels

(decades, years, months, and days). To provide the analyst with confidence in our parsing algorithm, and to make use of the human ability to see patterns, we also provide a view of the raw passwords. There are three main components of the view. The Radial Plot shows the distribution of dates parsed from passwords along years and decades, the Tile Map depicts the distribution of passwords across days and months, while the raw passwords are shown in a Wordle view. Performing filtering in a high-level view, such as the Radial Plot, narrows the context of the lower level ones, in a top-down fashion; conversely, removing elements from the low level views triggers updates in the high level ones. Despite the huge amount of data, we strive for fluidity to support perception of changes resulting from transition between states.

4.1.1 Radial plot

This view represents years through circles positioned in a radial layout (see Figure 1, bottom left). All years of a certain decade are evenly distributed along a ring, in clockwise order. The rings, representing decades, are organized in ascending order from center to periphery. Each spoke represents years ending in a particular digit. The frequency of passwords in a given year is encoded by color, according to a quantile scale that maps the frequency values to the range [0,9], corresponding to the colors of a sequential multi-hue pallette published by Brewer [7]. This scale is meant to reduce the negative visual effect produced by outliers, which occurs with a linear color scale.

The radial view enables observation of cyclical patterns, while also giving us a sense of the linear growth of frequency over the decades; furthermore, it enables rich interaction through selection of rings, circles and labels. The most common cyclical representation is, however, the spiral [8, 19]. We choose instead the ring-based configuration because it allows selection of rings (aggregation by decade), which is an important task in this context.

The default state corresponds to the overview, where the whole dataset is shown in all views, and can be reached by clicking on a blank space in the Radial Plot. Selecting a year by clicking it updates the Tile Map to show the corresponding frequency distribution across days of that year, and the Wordle is filled with the corresponding passwords. In the same way, it is possible to aggregate the years by decade by selecting a ring. Cross-decade aggregation is supported by clicking on an external label at the end of a spoke, *e.g.*, clicking '2' would select the years 1902, 1912, 1922 and so forth.

4.1.2 Tile Map

The Tile Map (see Figure 1, top) uses a calendar layout to display the frequencies computed for each day in a particular year [14]. The color encoding is consistent with the Radial Plot; that is, frequent regions are evidenced by dark tiles. A click on a tile triggers an update in the Wordle, which will show the raw passwords associated with the selected day. We extend the original use of Tile Maps by plotting aggregated values from multiple years, much like as though several maps were stacked. When used in this way, the calendar nature of the visualization loses its meaning, so we discard the labels informing the days of week (Monday, Tuesday, etc.). Although simultaneous display of multiple Tile Maps in a vertical list eases comparison between years [22], aggregating them in a single unit allows better perception of patterns *accumulated* over a period of time.

4.1.3 Word cloud

This visualization builds on the idea of a Wordle diagram, a tightly packed version of a word cloud [20] (see Figure 1, bottom right). The view is populated with raw passwords which match the selected years (Radial Plot) and day, if any (Tile Map). The passwords are sized according to the number of times they occur in the underlying dataset. An indicator bar is used to show the proportion of matched passwords which are purely numerical compared to those which contain a date-like numeric sequence as well as words and other symbols. This bar is interactive and can be used to restrict the view to the corresponding subset by clicking the corresponding bar.

In order to allow a researcher to remove any passwords which are strong outliers, and to see patterns in the remaining data, we provide the ability to select and remove a password from the Wordle. The filtered word goes to a 'filtered' panel on the right side, then the Wordle is recomputed. When the computation is done, an animation smoothly reorganize the passwords.

Since it can be difficult to keep track of what has changed when a new layout is calculated (*e.g.*,, which passwords got more or less importance after a filter is adjusted), the duration of the transition is proportional to the frequency of the password. So, more frequent (bigger) passwords move slower. While we have not tested this, we feel that this appearance of the larger passwords moving more slowly helps to give stability to the view during the relayout process.

4.2 Implementation

The tool is a web-based application that runs entirely in the browser, is written in JavaScript, and built on top of a set of web technologies standardized by W3C; namely, HTML, CSS and SVG. In additon, we use the D3 library [6] to manipulate data and the page's elements, to control animation, map data values to visual attributes and deal with events.

The web platform is convenient to our distributed and multi-disciplinary team, since it allows easy and fast deployment of new releases. It is also suitable for the rapid-prototyping development model that was undertaken.

5. SEMANTIC PATTERNS DISCOVERED

When using our date visualization tool, we noticed a number of interesting patterns in user choice (Figure 1). To summarize, there appears to be a preference for the following:

- Years after 1969. The popularity of a year is indicated by the darkness of the color in the radial portion of the visualization. See Section 5.1 for further details.
- Text words that spell out the name of a month (*e.g.*, "May12009"); see Section 5.2.
- Two years immediately after one another (*e.g.*, "20082008" or "19391945").
- The first two days in each month (*e.g.*, "010989").
- Repeated months/days (*e.g.*, "August 08").
- Holidays (*e.g.*, Valentine's day, Christmas day, and New Year's day); see Section 5.3.

We use each of these observations to specify patterns, which we use to compile a dictionary used to analyze security implications (discussed in Section 6). We investigate these patterns further in the following subsections.

5.1 Recent Years

The radial plot indicates that recent years, in particular after 1969, are the most popular. Years in the 1980's, followed by 1990's and then the 2000's appear to be the most popular. There are still a fair number in the 1970's and 2010's, and the popularity noticeably drops after 1969. We investigated this effect further and found that 1,160,801 (86% of purely numeric date passwords) represent dates after 1969. Some possible reasons for this preference are that the dates correspond with: (1) the birthdays of people using these accounts, (2) the dates of significant events for the people using these accounts, and (3) the dates that people created these accounts.

5.2 Text Combined with Dates

Using the Wordle portion of the visualization, we examined the most popular text strings that co-occur with dates. We observed that single-characters and initials appear the most frequently, and when full words are used, they are often the months of the year. This motivated us to examine how many passwords match date patterns, where the month is spelled out as opposed to being in a purely numerical format. We generated a set of formats for such dates, for example, MonthDDYY (see all formats in the Appendix). In all cases where the day is a single digit, we assume no leading zero is present. Our results are shown in Table 2.

Years considered	# of passwords	% of **all** passwords
1900-2012	124460	0.38%
1969-2012	117436	0.36%

Table 2: Passwords in the RockYou dataset that are in a mixed characters and digits representation of a date (*e.g.*, "1May1990").

We found these numbers to be quite surprising, given that dates written in this format are rather specific. Table 3 combines this result with the pure number results that are dates, showing that nearly 5% of users *choose a date as their password*, and nearly 4% of users *choose a date on or after 1969 as their password*. As indicated in Table 1, the number should be even higher when considering users who choose dates as *part* of their passwords.

Years considered	# of passwords	% of all passwords
1900-2012	1479398	4.54%
1969-2012	1278237	3.92%

Table 3: Passwords in the RockYou dataset that match a date pattern (*e.g.*, "1May1990" or "01051990"). Note that dates which can also be considered a numerical pattern (*e.g.*, "112233") are not included in this result.

5.3 Holidays

Through exploring using our visualization, we discovered that some familiar dates "pop out", which correlate with holidays such as Valentine's Day, New Year's Day, New Year's Eve, and Christmas Day (see Figure 1). While exploring the decades individually, we also noticed a number of other noteworthy dates appearing more frequently than expected, including:

- March 21 (First day of spring; Persian new year)
- December 21, 2012 (date associated with the "2012 phenomenon")
- August 17, 1945 (Indonesian Independence Day)
- April 14 and 15, 1912 (when the Titanic sank)

6. SECURITY IMPLICATIONS

Our observations using our visualization tool provide deeper understanding of user choice relating to the semantic category of dates. It provides information regarding how an attacker might perform an offline attack against a system in which he or she has no knowledge of the users, their spoken languages, and the dates they might choose (*e.g.*, does *not* know the user's birthday). Our analysis can also inform password policies and guidelines.

6.1 Date-based Guessing Attacks

Here we focus on purely numeric passwords, showing the results of building a dictionary based on each of the patterns discussed in Section 5. Our results are provided in Table 4. Of particular interest are the bolded values in the last two rows. In the second last row ("combined"), we see that by creating a dictionary which combines all of our visualization-observed patterns, we would be able to guess over 27% of date-based passwords using a dictionary composed of only approximately 15% of the possible dates. The final row shows that we can guess over 22% of date-based passwords using a dictionary composed of only approximately 7% of the possible dates.

Our findings approximate the extent to which these patterns dominate user choices of dates. The breakdown of each individual sub-dictionary, and the combined dictionary (with duplicates removed) is provided in Table 4.

Table 4 shows that these patterns correctly capture approximately 27% of date passwords, which corresponds to approximately 1% of all RockYou passwords. We emphasize that we have eliminated our identified numerical patterns (*e.g.*, "121212") from these results, and that by combining raw numerical patterns with this dictionary, even more passwords could be guessed; however our purpose in the present paper is to quantify the effect of popularly-chosen dates. The results of the combined dictionary show that we could guess nearly 1% of all RockYou passwords in approximately 15,000 guesses defined by "popular-looking" dates.

Given that this dictionary uses only purely numerical passwords, it could model an attack under the following threat model — when an attacker only wishes to obtain access to a single account, account-lockouts are not implemented (or the attack is offline), and the attacker knows nothing about the target user group (*e.g.*, language, birthdates, etc.). Of course, numerical patterns appear to be more popular and would pose more of a threat, but on some systems such obvious passwords may be blacklisted.

Dictionary (1900–2012, unless otherwise specified)	dictionary size	% of full dictionary	# passwords guessed	% of all date passwords	% of all RockYou passwords
(1) All days	**206658**	100.00 %	1354938	100.00%	**4.16%**
(2) Valentine's day	752	0.36%	6020	0.44%	0.02%
(3) Christmas day	426	0.21%	5675	0.42%	0.02%
(4) New Year's Eve	426	0.21%	4562	0.34%	0.01%
(5) New Year's Day	539	0.26%	9835	0.73%	0.03%
(6) First days of every month	11193	5.41%	105493	7.79%	0.32%
(7) All days in December	15501	7.50%	94957	7.01%	0.29%
(8) Repeated days/months	5490	2.66%	71709	5.29%	0.22%
(9) Repeated days/months/years	81	0.04%	16058	1.19%	0.05%
(10) Year1Year2	12769	6.21%	29976	–	0.09%
(11) Repeated years	113	0.05%	10490	–	0.03%
(2–11) Combined	**31856**	15.49%	372640	**27.50%**	**1.14%**
(2–11) Combined (only 1976–2012)	**14914**	7.26%	303334	**22.39%**	**0.93%**

Table 4: Passwords in the RockYou dataset that were guessed by dictionaries representing each of the patterns that we found in our visualization.

6.2 Password Policies and Guidelines

We use the presented visualization to gain further understanding of how people choose dates in passwords. The date subset appears worthy of investigation as it is apparently a common semantic category within user choice; nearly 5% of all user passwords in the RockYou dataset can be considered a pure date. A dictionary that would be able to guess all of these pure dates would consist of approximately 508, 492 entries, which is feasible to guess in a short amount of time in an offline attack. This alone creates patterns that are easy for attackers to guess, implying that it would be prudent to recommend that users do not choose a pure date as their password, even when it adheres to all other password rules (*e.g.*, "May1/2009" would satisfy common password requirements, but likely should be disallowed).

Our findings also strongly suggest the presence of certain patterns in user choice of dates. These patterns tell us something about user preferences, which provide further insight into the password selection process. For example, users seem to prefer dates that fall on the first day of the month, or are a partial repetition. This raises a question of whether users might prefer passwords that can be characterized by multiple patterns? It also raises the question of whether certain numbers are more memorable than others? If either is so, this could have implications for creating better password guidelines to aid users in choosing a more secure yet memorable password.

7. CONCLUSIONS

We have designed and created a visualization to aid the detection of date-related semantic patterns in user choice of passwords. Our visualization enabled discovery of a number of semantic patterns, including preferences for: years after 1969, text words that spell out the name of a month, two years immediately after one another, the first day in each month, repeated months/days, and holidays.

These semantic patterns have security implications — most notably, they enable the creation of language-independent password guessing dictionaries, which require no a-priori knowledge of the users. These dictionaries could be successful in an offline attack or against systems that do not implement account lock-out policies. We created one dictionary of approximately 15,000 popular dates that guessed approximately 1% of passwords from the RockYou dataset. We also found that approximately 4% of RockYou passwords were purely numeric dates, which can be guessed in a dictionary of approximately 200,000 entries. Finally, we found that over 4.5% of RockYou passwords can be characterized as dates (either purely numeric dates or dates that spell out the name of the month).

Our findings suggest it would be prudent to recommend that users do not choose a pure date numeric sequence as their password. Our findings also strongly suggest the presence of certain patterns in user choice of dates. These patterns tell us something about user preferences, which provide further insight into the password selection process. Future work includes exploring other semantic categories that exist in user-chosen text passwords.

8. ACKNOWLEDGMENTS

The authors would like to thank the Natural Sciences and Engineering Research Council (NSERC) for funding through the Discovery Grant program.

References

[1] W. Aigner, S. Miksch, W. Muller, H. Schumann, and C. Tominski. Visualizing time-oriented data:a systematic view. *Computers & Graphics*, 31(3):401–409, 2007.

[2] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Trans. on Visualization and Computer Graphics*, 14(1):47–60, Jan 2008.

[3] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*, chapter 7, pages 15–44. Number 1997 in Human-Computer Interaction Series. Springer London, 2011.

[4] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *IEEE Symp. on Security and Privacy*, 2012.

[5] J. Bonneau and S. Preibusch. A birthday present every eleven wallets? the security of customer-chosen banking pins. In *FC '12: Proc. of the Int. Conf. on Financial Cryptography*, 2012.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Trans. on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[7] C. A. Brewer. Colorbrewer. URL `http://colorbrewer2.org/`. Last accessed July 09, 2012.

[8] J. V. Carlis and J. a. Konstan. Interactive visualization of serial periodic data. In *Proc. of the ACM Symposium on User Interface Software and Technology - UIST '98*, pages 29–38. ACM Press, 1998.

[9] R. Chapman. Google password strength api. URL `http://www.codeproject.com/Articles/19245/Google-Password-Strength-API`. Last accessed June 24, 2012.

[10] G. Conti. *Security Data Visualization:Graphical Techniques for Network Analysis*. No Starch Press, 2007.

[11] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. VisGets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Trans. on Visualization and Computer Graphics (Proc. of the IEEE Conf. on Information Visualization)*, 14(6):1205–1213, Nov./Dec. 2008.

[12] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proc. of the Int. Conf. on World Wide Web*, WWW '07, pages 657–666. ACM, 2007.

[13] C. Herley and P. Van Oorschot. A research agenda acknowledging the persistence of passwords. *IEEE Security Privacy*, 10(1):28–36, 2012.

[14] F.-S. T. Mintz, D. and M. Wayland. Tracking air quality trends with sas/graph. In *Proc. of the 22nd Annual SAS User Group Int. Conf.*, pages 807–812, 1997.

[15] D. Schweitzer, J. Boleng, C. Hughes, and L. Murphy. Visualizing keyboard pattern passwords. *Information Visualization*, 10(2):127–133, 2011.

[16] S. F. Silva and T. Catarci. Visualization of linear time-oriented data: A survey. In *Proc. of the Int. Conf. on Web Information Systems Engineering (WISE)*, pages 310–, 2000.

[17] SkullSecurity.org. Leaked passwords. `http://www.skullsecurity.org/wiki/index.php/Passwords`, Last accessed June 27, 2012.

[18] J. Stoll, C. S. Tashman, W. K. Edwards, and K. Spafford. Sesame: Informing user security decisions with system visualization. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[19] C. Tominski. Enhanced interactive spiral display. In *Proc. of the Annual SIGRAD Conf., Special Theme: Interactivity*, pages 53–56, 1999.

[20] F. B. Viégas, M. Wattenberg, and J. Feinberg. Participatory visualization with Wordle. *IEEE Trans. on Visualization and Computer Graphics*, 15(6):1137–1144, Nov./Dec. 2009.

[21] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proc. of the ACM Conf. on Computer and Communications Security*, CCS '10, pages 162–175, 2010.

[22] R. Wicklin and R. Allison. Congestion in the sky: Visualising domestic airline traffic with sas. ASA Statistical Computing and Graphics Data Expo 2009, 2009.

APPENDIX

A. NUMERICAL PATTERNS

Pattern	Examples
Repeated digits	123123, 112233, 111222
Progression	12345, 02468, 654321
Palindrome	45754, 33633, 045540
Partially Repeated	080875, 010189, 121204

Table 5: List of numerical patterns

B. DATE FORMATS

Format
Textual Month
MonthDDYY, MonthDDYYYY, DDMonthYY, DDMonthYYYY, MonthDD, DDMonth YYYYMonth, YYMonth, MonthYYYY, MonthYY
8 digits
ddMMyyyy, MMddyyyy, yyyyMMdd, yyyyddMM
7 digits
ddMyyyy, Mddyyyy, dMMyyyy, Mdyyyy, yyyyddM, yyyyMdd, yyyyMMd
6 digits
ddMMyy, MMddyy, dMyyyy, Mdyyyy, yyyyMd
5 digits
ddMyy, Mddyy, dMMyy, MMdyy

Table 6: List of date formats