A Large Crowdsourced Dataset of Security Questions

Jennifer Golbeck, Simon Li
College of Information Studies
University of Maryland
College Park, MD 20742
jgolbeck@umd.edu,simonli@gmail.com

ABSTRACT

Security questions are a common fallback authentication mechanism across the internet, from banking to e-commerce to social media. Previous studies have investigated issues with those questions, with a marked focus on their susceptibility to attack. Indeed, some questions have a small set of possible answers and others are easily guessable by people who know or know about the account owner, and this makes them less secure. However, there are many more important avenues of study regarding security questions: Is there bias inherent in some of these questions? Which are practically easier or more difficult to remember? How does this vary based on demographics? To support this type of usable security research, a dataset of security questions is required. We have created a public, living, crowdsourced dataset of questions with their sources. This paper introduces the first version, with analysis, and we plan to release regular updates that expand and enhance the collection to support future research.

KEYWORDS

security questions, usable security, dataset

1 Introduction

Security questions are ubiquitous across the web as a fallback authentication mechanism. There have been various studies on the security of these questions – essentially, how easily could an attacker guess or access the answer in order to penetrate an account. However, from a usable security perspective, there are many other research questions about Security Questions that have yet to be addressed. In practice, how well do people remember the answers to these questions? What biases lie in the questions? Are certain questions less memorable for certain demographics?

Who Are You?! Adventures in Authentication (WAY) 2020.
August 7, 2020, Virtual Conference..

To answer these questions, researchers need an up-to-date dataset of security questions. In this paper, we introduce a living, crowdsourced dataset of security questions. Built from user submissions, the dataset currently has 283 unique questions from 607 submitted questions. Our plan is to regularly release updates to this dataset with new security questions that people have observed in use and the sources of those questions.

Regularly updated versions of the dataset will be available on Github at https://github.com/jgolbeck/SecurityQuestions.

We also present a descriptive analysis of this dataset, identifying major themes, popular questions, and other characteristics of the questions in version 1 of this dataset release.

2 Related Work

Security questions have been studied in the past, though that work has largely focused on the security of the questions, i.e. how easy it is for an intruder to guess or obtain the answer to a secret question.

One of the most comprehensive studies of this type is [1] from 2008. This paper focused on security questions used on banking sites and analyzed security issues with those questions. The author identified several main problems with these questions:

- Inapplicable – questions like "Which high school did your spouse attend" that do not apply to people who are married.
- Non-Memorable – questions like "What is the last name of your kindergarten teacher" which many people may not remember
- Ambiguous – questions where a significant fraction of the population would truthfully have multiple possible answers they could give
- Guessable – questions where it would be statistically easy to guess an answer. For example, "What is the last name of your favorite President," is easily guessable both because 20% of the population answer "Lincoln" and because there are so few Presidents to choose from.
- Attackable – questions where an attacker who knows the victim is likely to know the answer to personal information

- Automatically attackable – questions with answers that can be automatically collected from Facebook or other sources to guess.

They analyzed questions from fifteen online financial institutions, with 215 total questions, 100 of which came from one bank. This type of analysis is important and refining the categories and applying them to more modern questions which may have been introduced over the last 12 years, requires a public dataset.

A 2009 looked at the security of questions from webmail providers [2] and found similar weaknesses. More recent research has looked at geographically based questions and found increased security [3]. However, a thorough analysis would compare these to a standard set of questions that have been thoroughly evaluated for both guessability/attackability and human factors like memorability and clarity. Again, a large public dataset would facilitate this.

Factors like memorability are a known problem, and researchers have addressed this through techniques like gamification [4] or indicators of the strength of their answers [5].

Other work has looked at creating dynamic questions that are user specific, easy to remember, and harder to guess [6]. Again, work like this would benefit from having a pool of existing security questions available for comparison, and having previous results on those questions available to compare their own participants' results to.

3 Dataset

3.1 Data Collection

To create our first version of this dataset, we put out a request on social media and mailing lists. We asked people to take screen shots of security questions they encountered and share them along with the source. Having screen shots ensured that we could capture the exact phrasing of each question. As this is a living dataset, we are continuing this collection at https://github.com/jgolbeck/SecurityQuestions.

We then transcribed each question and merged questions that asked for the same phrasing. For example:

- What is your favorite holiday destination?
- What is your favorite vacation destination?

And

- Name of First Pet
- What is the name of your first pet?

The sources of all phrasings were listed together with the standardized question in our dataset. We also included a file in the dataset that listed all the mappings from the original question to standardized phrasing.

We had a total of 607 security questions submitted and 283 unique questions in the dataset. These came from 52 unique sites.

The full set of security questions with counts, categories, and themes are included in the online dataset.

3.2 Categorization

Using a grounded theory open coding approach, we developed a set of categories and themes to classify the security questions. People's Names, Favorites, and Locations were the dominant categories, with several important but more minor categories.

The categories with examples are as follows:

- People's Names (94) – names of people and pets. Names of locations or things are under other appropriate categories
  What is your spouse/significant other's father's first name?
  What is your spouse's mother's maiden name?
  What was the name of the best man at your wedding?

- Favorites (77) – this includes "least favorite" items
  What is your favorite musical instrument?
  What is your favorite kind of cookie?
  What is your favorite musician's name?

- Locations (66)
  What street did you grow up on?
  What school did you attend for sixth grade?
  In what city or town does your youngest sibling live?

- Firsts (26)
  What was the first stage play you ever went to?
  What year did you first fly in an airplane?
  What was the first concert you attended?

- Time and Dates (14)
  What is the longest time you have lived in a city or town?
  When is your anniversary (mm/dd/yy)?
  What year did you get your first car?

- Numbers (13)
  What are the last four digits of your driver's license?
  What are the last four digits of your credit card?
  What are the last five digits of your student ID?

- Dreams and Aspirations (13)
  What is your dream job as a child?
  What city would you most like to visit?
  What is your dream car?

- Other (14)
  What was your high school mascot?
  What is your most unique characteristic?
  If you were an animal, what animal would you be?

Our of our 283 questions, 35 had two categories. There were three frequent combinations:

- Favorites and Name (14 instances)
  What is the name of your favorite high school teacher?
  What is the name of your favorite pet?
  What is the name of your favorite writer?

- Firsts and Location (7 instances)
  Where did you go the first time you flew on a plane?
  What was the first foreign country you visited?
  What was the first major city that you visited?

- Location and Aspirations (6 instances)
  What country would you most like to visit?
  What city would you most like to visit?
  What famous landmark would you like to visit?

Note that 13 questions fell into the Aspirations category and nearly half of these also asked about a Location. The other combinations included Favorites and Location (4 instances), First Times (2), and Names of Firsts (2).

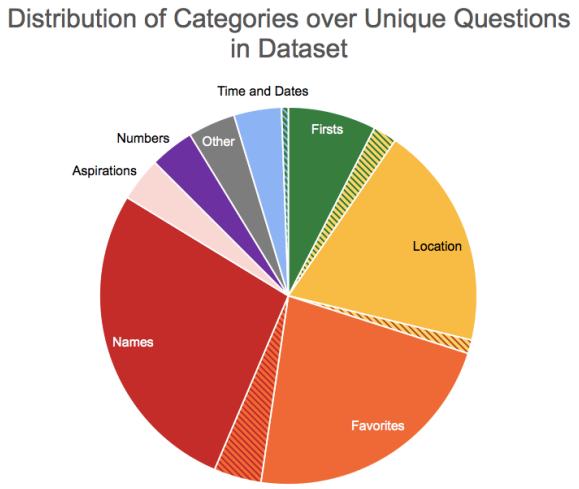

Figure 1. Distribution of themes over unique security questions in the dataset

We also looked at themes that cut across categories. Childhood was the most significant with 73 questions, and Family closely followed with 69.

Childhood themed questions were broadly distributed across categories, with 25 related to Location, 21 to Names, and 14 to Favorites. Family questions were more focused on Names, with 42 of 69 questions in this pairing. Locations were part of Family themed questions 16 times.

Minor themes included cars (7) and pets (6). However, note that these numbers are for the unique questions. Both car and pet questions were used frequently, with 19 car questions and 25 pet questions in the 607 question dataset.

3.3 Popularity

There was great variety in our dataset, as evidenced by the fact that 607 original questions were de-duplicated down to a large remaining set of 283 unique questions.

Out of these 283 questions, 98 appeared on more than one site. Figure 2 shows the frequency distribution of questions.

The 10 most popular questions were also the only questions that appeared on 10 or more sites. They are as follows:

1. What is the name of your first pet? (21)
2. What is your mother's maiden name? (19)
3. In what city did you meet your spouse/significant other? (14)
4. What is the name of your childhood best friend? (14)
5. What was your childhood nickname? (12)
6. In what city did your parents meet? (11)
7. What is the name of the first company you worked for? (11)
8. What was the make and model of your first car? (11)
9. In which city were you born? (10)
10. What was your high school mascot? (10)

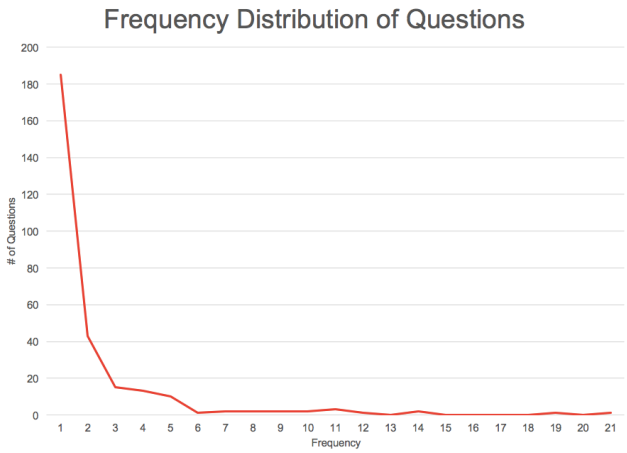

Figure 2. Frequency Distribution of Questions

4 Conclusion and Future Work

We believe this living dataset will provide the foundation for important future work on usability security research related to security questions, including investigations of usability, bias, and security.

REFERENCES

[1] Rabkin, A. (2008). Personal knowledge questions for fallback authentication: Security questions in the era of Facebook. *In Proceedings of the 4th symposium on Usable privacy and security (pp. 13-23)*. ACM.

[2] Schechter, S., Brush, A. B., & Egelman, S. (2009, May). It's no secret. measuring the security and reliability of authentication via "secret" questions. In 2009 30th IEEE Symposium on Security and Privacy (pp. 375-390). IEEE.

[3] Addas, A., Thorpe, J., & Salehi-Abari, A. (2019). Geographical security questions for fallback authentication. arXiv preprint arXiv:1907.00998.
[4] Micallef, N., & Arachchilage, N. A. G. (2017, November). Changing users' security behaviour towards security questions: A game based learning approach. In 2017 Military Communications and Information Systems Conference (MilCIS) (pp. 1-6). IEEE.
[5] Senarath, A., Arachchilage, N. A. G., & Gupta, B. B. (2017). Security strength indicator in fallback authentication: Nudging users for better answers in secret questions. *arXiv preprint arXiv:1701.03229.*
[6] Hang, A., De Luca, A., & Hussmann, H. (2015, April). I know what you did last week! do you?: Dynamic security questions for fallback authentication on smartphones. *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1383-1392). ACM.