Summary of changes for the USENIX Security 2022 Revise-and-Resubmit submission (new paper #391, old paper #81)

This paper was submitted in Usenix Security 2022 summer cycle (old paper no. #81) and received a Reject and Resubmit decision. To address reviewers' concerns we greatly enhanced the paper including substantial rewrite, new results and an improved and new user study to demonstrate the efficacy of our method for generating passphrases.

The primary changes made in this version are (compared to the previous version):

- Reframed the motivation for developing our passphrase generation system, MASCARA
- Explored and investigated the passphrase generation methods available in the wild and identified their shortcomings.
- Included additional quantitative analysis to establish factors affecting the memoraribility model we are using
- Redefined and rationalized our approach to measure guessability using well-established methods of measuring security of authentication secrets
- Created additional models for extensive exploration of guessability.
- Compared our method with new baselines.
- Developed and deployed a new longitudinal user study for comparing the efficacy of MASCARA-generate passphrases with in-the-wild methods.

Next, we will first include our original anonymous reviews from Usenix Security 2022 and then we will present how we addressed each one of the reviewer's concerns.

USENIX Security '22 Summer Paper #81 Reviews and Comments

Paper #81 record from the adapter: On Systematic Generation of Memorable And Secure Passphrases by MASCARA

Review #81A

Review recommendation

2. Reject and resubmit

Reviewer expertise

3. Knowledgeable

Overall merit

1. Bottom 50% of submitted papers

Writing quality

1. Unacceptable

Paper summary

The paper present a method (Mascara) of generating passphrases from a Markov model generated from the 5% most popular English-language wikipedia articles. The authors claim that their passwords are more memorable than "other state-of-the-art machine-generated passphrases" and more secure than in-use user-generated passphrases (from password database leaks). The authors use their own measures of security and memorability.

Weaknesses

1. The authors do not compare to other methods in the wild for creating memorable (namely grammatically correct) passphrases, such as https://makemeapassword.ligos.net/Generate/ReadablePassphrase or https://github.com/ligos/readablepassphrasegenerator (and I'm sure there are others). Their simple model ("Markov") is essentially an unoptimized/basic version of Mascara based on bigrams, but it seems an easy thing to test something diceware like that pulls a sequence of words having the form noun-verb-adjective-noun (for example).
2. The method for measuring security is house made, and not based on measures of entropy, which are much more accepted. Their measure of security (guessability) may be okay, but is poorly defined (what are the $\Beta_i$'s and so what is A_i?) and the authors should reason as to why their method measuring security is superior to more accepted practices in the community.
3. The method for measuring memorability is house made, purports to measure "character error rate", which is an accepted method, but their computation of memorability is not based on any sources. For example, why are infrequent words difficult to remember? (My intuition is the opposite, and so without any scientific references, I have nothing else to go on.) Likewise for probability of the phrase. But each item in section 4.1 should be defended based on scientific literature.
4. "generating passphrases often involve selecting random words from an English wordlist" ... this is not true - there are

non-english language passphrase generators out there. The authors should make the case of why they are limiting their study to English language passphrases.
5. The user study to test memorability is really a test of preference. Users will almost always (I assume) prefer a song title or common phrase for a password unless they know about and are concerned about security. The user study design is not based on or built from other studies on the same topic.

Comments for author
----
* There are many grammatical errors throughout - too many to enumerate here. Future submissions should be better edited.
* You also mix tenses throughout sometimes using present tense, sometimes past (for activities of the written research). Make a choice and stick with it.

* The scale for guessability is counterintuitive. High guessability should mean easy to guess and so low security.
* Why is $\gamma_{oov}$ included in section 4.1 if it is never used? Is the reason you aren't using it valid? You talk later about in-corpus and out-of-corpus, but surely, the corpus contains infrequent words.
* "They used Europarl ..." - who does "they" refer to? (p 5)
* You refer a few times that user passphrases are "movie and song titles, common short quotes or phrases" - where is this coming from? A reference, or an observation of your dataset - in which case, you should measure it. For example, you could internet search your User passphrases to see how many are linked to titles, quotes and so on.
* What is a stop word? Can you give examples at least?
* The user study only compares Mascara to Diceware and User ... why not Markov?
* "We set the frequencies to be 1 for all words" ... WHY?
* What are the sizes of the word sets for User vs Wiki-5?
* The detailed explanation of how you preprocess the data is overkill.
* "from 8,210 Wikipedia articles, containing a total of 2,95,02,034 words (and 4,55,614 unique words)" ... are these errors?
* p. 9 under "diceware": "We ran- domly sample words based on the word distribution in User passphrases from the dictionary without replacement to create a passphrase" really? I am confused ... a sentence before you are talking about wiki-5 ...
* can you explain the plateaus in figure 5 for diceware?
* p. 10 "upassphrases"? p. 11 "User-bg"?

Requested Changes

See other sections.

Questions for authors' response
----
1. In your measure of memorability, what assumptions are you making? Where are you "observ[ing] four distinct and important features of a sentence"?
2. Explain and defend your measure of guessability/security.
3. Mascara is SLOW. Is this even a reasonable method for that reason? Why or why not?
4. Why is 200 your user study size?
5. Defend your user study design and explain why you aren't replicating other memorability/usability study design used in the field.

Review #81B
===

Review recommendation

2. Reject and resubmit

Reviewer expertise

2. Some familiarity

Overall merit

1. Bottom 50% of submitted papers

Writing quality

3. Adequate

Paper summary

This paper proposes a passphrase generation system to improve upon the phrases produced by Diceware and those created by individual users. The goals are to provide passphrases that are easy to remember, yet hard to crack. The authors developed MASCARA, which uses a generative Markov model (with Wikipedia data) and some refinements to create passphrases with more memorable characteristics (as compared to Diceware) -- e.g., outputting only items with phrasal structure. They conducted a user study (n=175) on MTurk, to compare Diceware, user-generated, and MASCARA passphrases. Results indicated that users preferred MASCARA to Diceware, but also strongly preferred user-generated phrases (despite security issues).

Strengths

* trying to improve password + passphrase systems is a worthwhile research goal
* having a user study important for testing hypotheses
* results of MASCARA seemed to provide good results -- interesting to constrain the structure in a useful way (for both memorability and difficulty of guessing)

Weaknesses

* comparison with Diceware is a bit limited. Diceware is a commonly-used tool, yes, but others exist (e.g., tool built into Bitwarden)
* there is no strong argument made for the weaknesses of Diceware -- MASCARA sounds like an improvement, but it may not be significant enough to be worth it
* the user study was fairly simple -- no long-term memory testing, or use of multiple phrases, for example
* the memorability analysis (CER) is reasonable but it is not definitive; probably most useful for ensuring a baseline "floor" for MASCARA but not worth leaning on *too* hard for comparisons

Comments for author

It's good to see continued work in the area of improving authentication in a user-friendly way. The MASCARA tool looks like a reasonable passphrase generator -- some care was taken with the corpus and the refinements so that the resulting phrases seem like a good "tradeoff" in the passphrase space.

That being said, I do have a number of reservations. This approach sets itself up against Diceware, and while it may well represent a more memorable system, I failed to see that this was a significant problem; Diceware is often recommended and having some evidence of repeated user failure or frustration with it would help motivate this work. Similarly, there are several other systems available, such as the one in Bitwarden, or in several open-source projects on Github; there is no comparison with these. It would strengthen the paper if there was some indication that users are unsatisfied with current options.

The user study, while useful, also was somewhat rudimentary. It was short term and with a quite simple task; it was valuable to gain some insight into what users wanted, but ultimately it became a case of "I prefer my own" (despite your best efforts to make good passphrases). You also have data about the use of password managers, but haven't done much analysis with it. It would be helpful to know the degree to which people are already juggling multiple passwords + passphrases in their heads, or, for instance, if those who were happy with Diceware (for example) currently only had to recall one or two passphrases since everything else was in a password manager.

Requested Changes
----

* stronger evidence that this solution addresses a signifiant user problem that not supported by competing solutions: this includes evidence to motivate the problem as well as analysis of existing products in the passphrase space beyond Diceware
* a more robust user study that analyses more memory aspects (e.g., interference + medium/long-term use) to give stronger evidence about memorability
* more analysis on the users and their password manager experience, and how this may influence their preference towards passphrase choices

Below, we conceptually grouped the reviews and present how we addressed them in the new revision. We appreciate the thoughtfulness of the reviewers' comments, which helped us considerably improve this work in the major revision phase.

Reviewer's feedback	Author's response
Concerns regarding scope and motivation	
R2: there is no strong argument made for the weaknesses of Diceware -- MASCARA sounds like an improvement, but it may not be significant enough to be worth it	Shay et al. [54] showed that passphrases generated by this approach are error-prone for users and the resulting passphrases are not memorable with memorability being similar to randomly generated passwords.
R2: Evidence to motivate the problem as well as analysis of existing products in the passphrase space beyond Diceware	The motivation for the work has been reworked with a new section being introduced for the same (Section 3). To summarize: ● User passphrase has been analyzed by extracting it from existing password leaks and their shortcomings are elaborated ● A survey of famous password managers has been carried out and their passphrase generation algorithm was examined, if any ● Two primary system generation algorithms surfaced, Diceware and TemplateDice. Diceware passphrases have a severe drawback in the aspect of memorability. ● TemplateDice aims to improve it by making the passphrases memorable, but compromises on security. It is also not scalable and has portability issues. Generated passphrases are also finitely bound. ● We design Mascara to overcome these issues.
R1: The authors should make the case of why they are limiting their study to English language passphrases.	In line with the previous works (and for a better comparison with state of the art), our exploration considered analyzing and generating English passphrases [18, 21, 34].
Comparing MASCARA with methods from the wild	
R1: The authors do not compare to other methods in the wild for creating memorable (namely grammatically correct) passphrases, such as https://makemeapassword.ligos.net/Generate/ReadablePassphrase or https://github.com/ligos/readablepassphrasegenerator (and I'm sure there are others). Their simple model ("Markov") is essentially an unoptimized/basic version of Mascara based on bigrams, but it seems an easy thing to test something diceware like that pulls a sequence of words having the form noun-verb-adjective-noun (for example).	We have carried out a survey on the current state of the art system generated passphrases and have introduced them as proper baselines for comparison as well. The https://github.com/ligos/readablepassphrasegenerator is based on the TemplateDice algorithm used for comparisons. We have tried our best to ensure that all current state of the art passphrase generators have been considered in our evaluations.

R2: comparison with Diceware is a bit limited. Diceware is a commonly-used tool, yes, but others exist	We have carried out a survey on the current state of the art system generated passphrases and have introduced them as proper baselines for comparison as well. The https://github.com/ligos/readablepassphrasegenerator is based on the TemplateDice algorithm used for comparisons. We have tried our best to ensure that all current state of the art passphrase generators have been considered in our evaluations.
Concerns regarding user generated passphrase creation process	
R1: "We set the frequencies to be 1 for all words" ... WHY?	This has been removed.
R1: What are the sizes of the word sets for User vs Wiki-5?	There are 21,945 unique words in the word set of User passphrases, whereas Wiki-5 has 4,55,614 unique words.
R1: The detailed explanation of how you preprocess the data is overkill.	We moved the process to the appendix.
R1: "from 8,210 Wikipedia articles, containing a total of 29,502,034 words (and 455,614 unique words)" ... are these errors?	We double checked and the numbers are correct.
Concerns regarding guessability computation	
R1: The method for measuring security is house made, and not based on measures of entropy, which are much more accepted. Their measure of security (guessability) may be okay, but is poorly defined (what are the $\Beta_i$'s and so what is A_i?) and the authors should reason as to why their method of measuring security is superior to more accepted practices in the community. **R1:** The scale for guessability is counterintuitive. High guessability should mean easy to guess and so low security.	In the revised version, we have added citations and rationale for why we used guessrank---it is an established metric to quantify security in the context of guessing by an offline attacker as identified by multiple recent related works [49, 57, 58, 59, 60, 62] . In this version, we also used min-auto, an established method used widely in multiple prior works and proposed by Ur et al.[59]. Ur et al. demonstrated that taking the minimum guessrank of all the automated approaches (called min auto) is a reasonable approximation of the real world cracking scenario. We also mentioned right in the introduction that the higher the guessrank (the established metric to measure security) the lower the guess rank. **We significantly updated Section 2.2 to introduce these concepts and then again detailed the exact computation in enlarged and enhanced Section 4.2.** We used this formulation of guessrank with min auto throughout this paper. According to normal convention,

	higher the guessrank, higher the security (lower guessability).
Concerns regarding memorability computation	
R1: The method for measuring memorability is house made, purports to measure "character error rate", which is an accepted method, but their computation of memorability is not based on any sources.	We use CER to quantify memorability. Prior works [41, 44, 48] noted that CER is widely accepted as a proxy for memorability. To decide the parameters of a phrase (passphrase in our case) that has to be considered for memorability, we experiment. We used a dataset of 2,230 sentences, each of which has been annotated with the character error rate (CER) determined from a user survey [41]. We calculated various parameters for each phrase in the dataset like frequency of occurrence, out of vocabulary words, the average frequency of occurrence, etc. With these data, we find the statistically significant correlation of each feature concerning CER and found three statistically significantly correlated signals. They are:- ● Unigram probability (L1) :- High negative correlation (r = -0.83, p approximately 0). This has also been proven by prior works [36] ● Bigram probability (L2):- High negative correlation (r = -0.84, p approximately 0) ● Standard deviation of characters of the words:- Positive correlation (r=0.25, p approximately 0)
R1: why are infrequent words difficult to remember? (My intuition is the opposite, and so without any scientific references, I have nothing else to go on.) Likewise for the probability of the phrase. But each item in section 4.1 should be defended based on scientific literature.	Prior works have shown that phrases consisting of frequent words are easier to memorize [36]. This can also be seen from the fact that L1 is statistically significant wrt to CER (p approx 0) and high negative correlation (r = -0.83) which we learnt from the experiment.
R1: Why is γ_{oov} included in section 4.1 if it is never used? Is the reason you aren't using it valid? You talk later about in-corpus and out-of-corpus, but surely, the corpus contains infrequent words.	This has been changed after observing that γ_{oov} is not statistically significant with respect to CER when the experiment was carried out
R1: "They used Europarl ..." - who does "they" refer to? (p 5)	This section has been changed and this part was removed
R2: the memorability analysis (CER) is reasonable but it is not definitive; probably most useful for ensuring a baseline "floor" for MASCARA but not worth leaning on *too* hard for comparisons	The memorability analysis is supported with a lot of experimental datas as mentioned above and prior works. Moreover, a user study has been carried out to compare the memorability of passphrases across the different

	models, and data from the study corroborated the results we obtained.
Concerns regarding MASCARA algorithm **R1:** What is a stop word? Can you give examples at least? (section 4.5)	 This has been addressed. Stop words are a set of commonly used words in a language. Examples are “the”, “have”, “and”, “is”, etc.
Concerns regarding offline evaluation **R1:** p. 9 under "diceware": "We randomly sample words based on the word distribution in User passphrases from the dictionary without replacement to create a passphrase" really? I am confused ... a sentence before you are talking about wiki-5 ... **R1:** can you explain the plateaus in figure 5 for diceware? -->it compared the distro of passphrases	 This entire section 5 has been revamped (with new experiments and new text) and this concern has been addressed by fixing the writing oversight. The model of guessrank calculation has changed (min auto now), and hence the corresponding figures have changed.
Concerns regarding user study **R1:** The user study to test memorability is really a test of preference. Users will almost always (I assume) prefer a song title or common phrase for a password unless they know about and are concerned about security. The user study design is not based on or built from other studies on the same topic. **R1:** The user study only compares Mascara to Diceware and User ... why not Markov? **R1:** a more robust user study that analyses more memory aspects (e.g., interference + medium/long-term use) to give stronger evidence about memorability	 We redesigned and redeployed a longitudinal two-part survey-based user study to actually calculate memorability in a real deployment. The experiment and the results of the user in **Section 6 (a fully new addition and a new study in this revised version)**. We added Markov in our new user study. Please refer to Section 6 for the full result showing efficacy of MASCARA. Our new survey instrument is in Appendix D. We incorporated the reviewer suggestion and created a user study at par with other works on password/passphrase memorability. The new study uses a longitudinal design to actually measure memorability of the users for different algorithms. Our user study demonstrated that in practice MASCARA is able to generate passphrases which gives a higher recall than state of the art methods for creating system-generated passphrases.

Writing changes **R1:** There are many grammatical errors throughout - too many to enumerate here. **R1:** What are “upassphrases” **R1:** What is “User-bg”	We made a deep grammatical pass over the draft and aside from rewriting very large parts of the paper with new experiments, we also made very detailed changes to the paper for handling grammar, spelling issues and problems with the writing. We removed the typo. We removed the typo.