Vol.2016-DPS-168 No.16
Vol.2016-SPT-21 No.16
Vol.2016-EIP-74 No.16
2016/11/18

実利用されているパスワード強度メーターの分析と検証

菅井 琢[1,a)]　金岡 晃[1,b)]

概要：Web 上のサービス利用等で会員登録時のパスワード設定をする際に、パスワードの安全性を測るパスワード強度メータを導入しているサイトは多い。しかしそのメータによる強度スコアの計算については一貫した計算の基準が欠如しており、本来パスワードが持つ情報量との乖離がある可能性が指摘されていた。一方で、現在広く利用されているサイト群におけるパスワード強度メータの実態調査はされてこなかった。そこで本研究では Alexa Top 100 サイトからパスワード強度メータ利用をしているサイトを抽出し、それらの挙動を分析し、その実態を明らかにする。さらに、3000 万件を超えるパスワードデータセットを用いて実際にメータが算出するスコアの分布などの統計情報を分析し、その違いを深く明らかにする。

Analysis and Verification of Actual Use Password Strength Meters

Taku Sugai[1,a)]　Akira Kanaoka[1,b)]

1. はじめに

パスワードよりも強固な認証手法へ移行しようという追及の旅は、パスワードが持つ強力な配備のしやすさ (Deployability) により阻まれている [1]。さまざまな認証手法が次から次へと提案されているものの、社会への展開や実装の難しさなどから、適用が安易なパスワードはいまだ広く使われている。そういったパスワード認証に対し、利用者により強いパスワードの設定を促す仕組みを採用することにより、パスワード認証自身をより強固なものにしようというアプローチがいくつか存在する。

1 つはパスワード構成ポリシであり、ある一定の制限を超えなければそのサービスを利用するために設定するパスワードとしてサービス側が受け付けないとするものである。代表的なポリシには「パスワードは 8 文字以上」や「アルファベットの大文字が少なくとも 1 文字使われること」などがある。もう 1 つはパスワードの強度メータである。ユーザが入力したパスワードがどの程度の強度を持つかのスコアを計算し表示するパスワード強度メータは、ユーザに対してより強いパスワードを設定させる効果があることが認められている。

パスワード構成ポリシやパスワード強度メータの適用については、いくつかの方面から研究がされており、より効果的に利用できる環境が整いつつある。一方で、そういった研究成果の社会展開は進んでいるとは決して言えず、まだ多くのサイトは適切なポリシ設定や強度メータの適用がされていないと考えられている。Dell'Amico らはパスワード強度メータに焦点をあて、多くのパスワード強度メータ利用サイトが実際のパスワードが持つ情報量を反映したスコアになっていないことを指摘し、その近似を高速にできる手法を提案した [2]。Dell'Amico らによる指摘は十分に受け入れられるものであった一方で、代表的なサイトを含め、具体的にどのサービスがどういったスコア計算をし、実際の情報量とどれだけの差異があるかについては調査されていなかった。

本稿では、Alexa Top 100 の Web サイトからパスワード強度メータを採用しているサイトを抽出し、さらにそれぞれのスコアの計算手法を分析する。各サービスのメータがどういった挙動をし、どういう手法でスコアを計算しているかを明らかにし、分類と比較を行う。また実際に利用されていたパスワードのデータセットを用いて、それぞれのメータによるスコアの分布を調査しその差を明らかにする。

[1] 東邦大学
Toho University
[a)] 5513054s@nc.toho-u.ac.jp
[b)] akira.kanaoka@is.sci.toho-u.ac.jp

Vol.2016-DPS-168 No.16
Vol.2016-SPT-21 No.16
Vol.2016-EIP-74 No.16
2016/11/18

2. 関連研究

パスワードの強度についての文献と言えば、NIST が発行している NIST SP 800-63 においてその強度がビット数であらわされていることがまず挙げられよう [3]。NIST SP 800-63-1 ではパスワードが持つ情報量についての記述があるが、ユーザがパスワードを設定する場合のパスワードの統計的な偏りといった特徴を調査が行われ [4], [5]、これらの結果から情報量は短いパスワードではより小さく、長いパスワードではより大きいものであることが示された。その後、そういった特徴をパスワード推測攻撃に応用すること [5] などがされてきた。さらに、推測攻撃を応用し耐性を評価する手法として転用しパスワードを強化するための方策の評価として用いられることも多くなってきた [6], [7], [13]。

パスワード強度メータについては、メータ設置によりユーザのパスワード強度が上がることが Egelman らにより示されている [9]。しかし Carnavalet らはそれらの疑問を呈している [10]。その疑問については、Ur らが多くのパスワード強度メータの種類の整理とその効果についての大規模な調査を行い、効果を示したことで解消された [7]。メータの表示方法の違いやメータと強度の紐づけ方法の違いなど、複数のメータについてユーザ実験を行い、それらのメータ表示のもとで作成されたパスワード群がどれほど推測攻撃に耐性があるかの分析がなされた。

パスワード強度メータにおける強度の計算手法については、Ur らの研究 [10] ではパスワード構成ポリシに従ったポイント制、そして Egelman らの手法 [9] ではパスワード長と文字種数の対数の積、といったように手法が定まっているわけではない。これらの手法では、その強度の算出がそのパスワードが本来持つ情報量の正確な保証されているわけではない。パスワードの強度は、ある特徴をもつパスワード群がどれだけ推測攻撃に耐性をもつか、という視点で評価がされることが多い [6], [7], [11], [12], [13]。しかし、パスワードメータは与えられたパスワード単体がどれだけの強度を持つかを計算しなければならず、推測攻撃への耐性評価は計算に時間がかかるために直接的な適用はむずかしい。パスワード単体の強度計算についての議論は Castelluccia ら [14] や Dell'Amico ら [2] によって行われている。いずれも条件付確率をある程度間引いて行うことにより近似誤差が少ないまま計算量を少なくすることができるようになっており、ユーザがパスワードを入力している間に強度計算を行いメータ表示をすることが可能になっている。

3. パスワード強度メータの調査

パスワード強度メータに利用される強度計算手法については、前章においてその妥当性が議論されているものの現段階で広く利用されているサービスではどういったメータ表示とスコア計算が行われているかは明らかにされていない。本研究ではそこを明らかにすることを目的として、Alexa 社が提供する"The top 500 sites on the web[15]"にある上位 100 のサイトで利用されている強度メータをの実態を調査する。

パスワード強度メータは Web サイト上のユーザ登録だけでなく、スマートフォンや PC のクライアントアプリケーション上のユーザ登録でも利用されているケースがあるが、それらで利用されているメータのスコア計算手法を分析するとなった場合、アプリケーションのリバースエンジニアリングが必要となり、分析の正確性や利用規約の抵触などが考えられたために本研究では Web サイトに限定した調査を行った。

3.1 ユーザ登録におけるパスワード利用

Top 100 サイトのうち多くのサイトがユーザ登録の仕組みをもち、そこでパスワードの入力がされていた。パスワード入力に際しては、パスワードの構成ポリシが設定されているものとパスワード強度メータが配備されているものなど、サイトによる違いがいくつか現れた。また Alexa のサイトがドメインごとにランキングしていることから、Top100 サイトの中では同一サービスでありながら国ごとのサービス用ドメインが別に存在するサイトが上位にいくつも存在するケースがあった。最も多いものは Google 社のサイトであり、ランキングトップの Google.com をはじめ日本版の Goole.co.jp を含め Top100 のうち 18 が Google の各国のサービスドメインであった。Google 社は各国サービスドメインでのユーザ登録とともに関連サービスのユーザ登録を accounts.google.com で一元化している。Youtube.com は関連サービスに含まれている。同様なサービス展開は Amazon 社や Microsoft 社にも見られた。

調査の結果、各サービスにおいて集約されている分の重複を取り除いてパスワード強度メータの採用状況を数え上げると、計 13 種類のパスワード強度メータの利用が確認された。13 種類の分類を表 1 に示す。なお、Alexa Top 100 サイトのなかには中国のサービスが複数含まれており、それらはユーザ登録の画面を持っていたが、多くはまず最初に携帯電話の番号を入力させるものであり、本調査ではその先の登録まで調査することができなかった。携帯電話の SMS を通じて初期コードを入力し、その後にパスワード設定をさせ、そこにパスワードの構成ポリシや強度メータがある可能性が存在する。

3.2 外観による分類

各強度メータの表示はいくつかのタイプに分類ができる。それら分類を表 2 に示す。

Vol.2016-DPS-168 No.16
Vol.2016-SPT-21 No.16
Vol.2016-EIP-74 No.16
2016/11/18

分類名	利用ドメイン例
Google	Google.com, Google.co.jp 他
Twitter	Twitter.com
Yahoo!Japan	Yahoo.co.jp
VK	Vk.com
Yandex	Yandex.ru
eBay	Ebay.com
Reddit	Reddit.com
Mail.ru	Mail.ru
tumblr	Tumblr.com
Apple	Apple.com
NAVER	Naver.com
楽天	Rakuten.co.jp
Dropbox	Dropbox.com

表 1 パスワード強度メータの利用サービス

外観の分類	サービス名
連続メータ	Twitter, Yandex, Reddit, 楽天
離散メータ	Apple, Dropbox, eBay, Google, Mail.ru, NAVER, tumblr, VK, Yahoo!Japan

表 2 パスワード強度メータの外観分類

図 1 Dropbox メータ

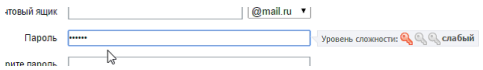

図 2 Mail.ru メータ

離散タイプのメータは、バーのブロックが分割されているもの（図 1）や記号利用（図 2）など離散値が明確なものと、一体化したバーの中で離散的にふるまうもの（図 3）などに分かれる。一体化したバーの中でふるまうものについては、おおむねその離散の分布が均等であるが、Apple と Ebay のメータに関しては均等の分布となっておらず、最高評価に至らない段階では均等だが最高評価の場合のバーの伸び率は下がっている。Apple のメータでは、バー全体の長さを 1 とした場合、おおよそ 0, 30%, 60%, 90%, 100% という段階（図 4）、eBay のメータではおおよそ 29%, 58%, 85%, 100% という段階となっている。いずれもメータ画像をキャプチャ氏、そのバーの長さを計測して得た値である。

3.3 構成ポリシ確認とスコア算出の確認ポイント

構成ポリシ確認の有無やスコア算出の有無はサービスにより異なるが、それらの確認や算出を JavaScript を用いてローカル（ブラウザ上）で行うものと、入力されたパスワードをサーバに送信しリモートで行うものとに大別される。

図 3 VK メータ

図 4 Apple メータの離散表示比較

利用される情報	サービス名
パスワード長	Twitter, Yahoo!Japan, VK, Yandex, Reddit, Mail.ru, Apple, 楽天,
同一文字の利用	Twitter,
文字の連続利用	Yahoo!Japan, Mail.ru, 楽天, Dropbox
フレーズの連続利用	Yahoo!Japan, Mail.ru
利用文字の種類数	Twitter, VK, Mail.ru, 楽天,
記号の有無	Twitter, Yahoo!Japan, Apple
数字の有無	Twitter, Yahoo!Japan, Reddit, Mail.ru, 楽天,
大文字利用の有無	Yahoo!Japan, Reddit, 楽天,
登録単語との一致	VK, 楽天,

表 3 スコア算出に用いられる情報

構成ポリシ確認をリモートで実施するものには、Twitter、Yandex、eBay、Reddit、tumblr、Apple がある。またスコア算出をリモートで実施するものには、Google、eBay、tumblr、NAVER がある。

3.4 スコア計算に利用される演算

ローカルでスコア算出を行うサービス 9 つについてスコア算出手法を分析した。利用される情報がスコアにどういった影響を及ぼすかはサービスにより変わる。表 4 は 9 つのサービスのうち Dropbox を除いた 8 つのサービスについて、利用している情報を表にまとめたものである。8 つのサービスはそれぞれ独自の算出方法を用いている一方で、Dropbox は公開されているライブラリ zxcvbn[16] を利用した算出をしている。

zxcvbn ではこれら 8 つのサービスとは大きくことなるスコア演算をしている [17]。その特徴を下に述べる。

- ブラックリスト（辞書）による一致確認
- キーボード配列上の近似文字の連続性確認

サービス名	スコア幅	平均スコア	分散
Mail.ru	0-3	1.00	0.43
Apple	0-4	1.01	0.02
楽天	0-100	33.84	382.03
Reddit	0-100	20.02	166.57
Twitter	0-100	27.54	228.72
VK	0-4	2.07	0.70
Yahoo!Japan	0-4	1.81	0.18
Yandex	0-100	38.55	14.69
tumblr	0-5	0.52	0.64
NAVER	0-4	1.34	0.97
Google	0-4	1.91	1.63

表 4 各メータの平均スコアと分散

サービス名	平均スコア	分散
Mail.ru	33.07	469.25
Apple	30.40	14.70
楽天	33.84	382.03
Reddit	20.02	166.57
Twitter	27.54	228.72
VK	51.84	435.20
Yahoo!Japan	45.29	113.09
Yandex	38.55	14.69
tumblr	10.44	254.27
NAVER	33.49	603.70
Google	47.81	1015.79

表 5 各メータの平均スコアと分散（正規化後）

- 一部文字の数字入れ替え確認
- 単語の反転確認

4. 各強度メータの特徴分析

3 章の分析の結果、各サービスのメータはスコア算出方法などサービスごとに大きく異なっていることが判明した。そこで本章では、パスワードデータセットを用いてそれぞれのメータがどういったスコアの特徴を持つかを分析する。

パスワードデータセットとして RockYou データセットを用いた。RockYou データセットは、2009 年に RockYou.com の利用者パスワードが漏えいした際の漏えいデータである。32,603,388 ユーザのパスワードデータが含まれている。

RockYou データセットの全パスワードに対し各メータのスコア計算を行い、その結果を分析する。ローカルでスコア算出しているメータに対しては、それぞれの算出手法が判明しているため、分析用の PC に演算をシミュレートしたプログラムを作成し、各パスワードのスコアを算出した。リモートでスコア算出しているメータに対しては、算出手法は判明していないため、算出を行う URL へのアクセスを行うクローラを作成し、各パスワードでのスコアを算出した。Google に対しては 25,858,997 のパスワード、NAVER に対しては 16,103,811 のパスワード、tumblr に対しては 17,513,097 でスコアを算出した。eBay に関しては、同一 IP アドレスからの 1 日の最大アクセス数が設定されており、クローラを用いた分析が行えなかった。

各メータの平均スコアと分散を表 4 に示す。メータによりスコア幅が異なるために直接の評価はできないものの、同じスコア幅をもつメータ同士でもサービスの違いによる平均スコアと分散が大きく異なることがわかる。

スコアの幅は各サービスのメータによって異なるため、直接の比較が難しい。そこで、スコアの幅を正規化することによる比較を行う。正規化は、それぞれのメータの外観を用いて、すべてのメータのスコア幅を 0-100 にする。スコア幅が 0-100 でないメータに対して、3.2 節で述べたようにメータ外観においてそれぞれの離散値の幅を測定し、それを用いて正規化をした。たとえば 0-4 のスコア幅においてそのメータ外観が各スコアで等間隔で表示される場合、それぞれのスコアを 25 倍した。一方で、0-4 のスコア幅においてそのメータ外観が各スコアで等間隔でない場合は、その割合にしたがい 0-100 になるように倍率を設定して正規化を行った。

正規化を行った後の平均スコアと分散を表 5 に示す。サービスごとで平均スコアに大きな差が出ていることがわかる。また正規化を行った状態でのスコア分布を図 5 に示す。各メータの特徴が良く表れており、Apple は高いピークを持つ以外は他のスコアへの分布が少ない一方で、Reddit は各スコアにまんべんなく分布が広がり、ピークの高さが低いことが見て取れる。またそれぞれのピーク位置もメータにより異なり、各サービスにおいて統一ないし類似したスコア分布をしているものが存在していないことがわかる。

5. 議論

5.1 パスワードが持つ情報量との比較

本研究での調査は、各パスワードメータの挙動分析をして、かつパスワードデータセットを用いたスコア分析をしたものであった。さらに深めるためには、それぞれのパスワードがどれほどの情報量を持つかを算出したのち、それぞれのメータのスコアと比較をすることで、情報量との乖離を分析可能になるだろう。パスワードデータセットの中での出現頻度の逆数を用いて情報量とするやり方と、Dell'Amico らが提案した近似手法による算出との 2 通りが考えられる。

5.2 リモートサーバでのパスワード判断の脅威

今回の研究ではパスワード強度メータのスコア算出に焦点を当てて調査と分析を行ったが、その中でスコア算出の前に構成ポリシを適用し、ポリシに合致するものが初めて

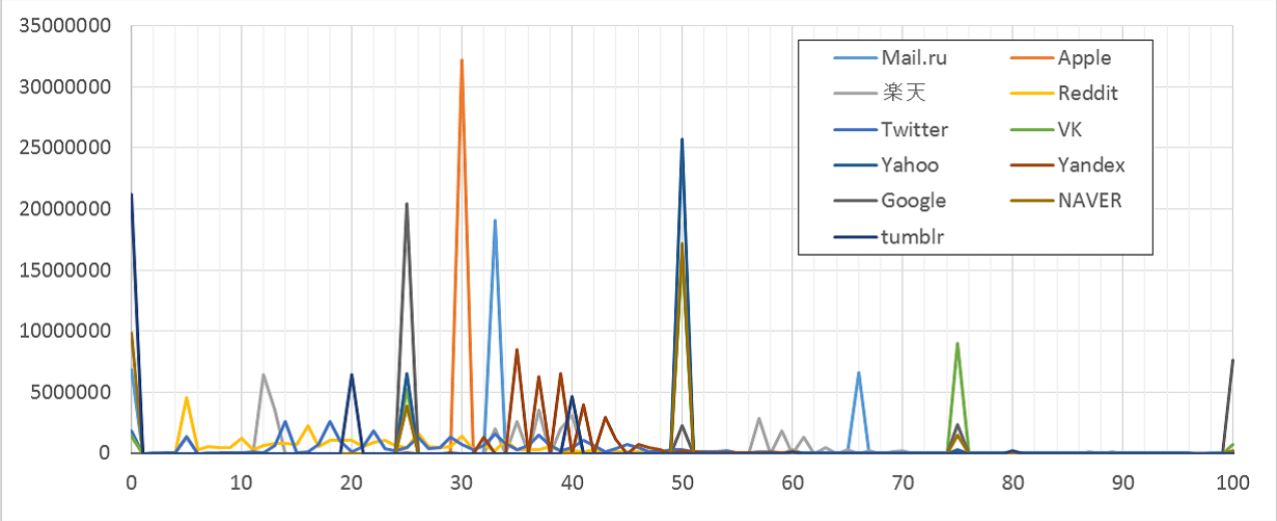

図 5 各メータによるスコア分布（正規化後）

スコア算出されるという合わせた利用法があることが判明した。そしてそれらの中に、構成ポリシとの適合チェックやスコアの計算をローカル（ブラウザ）で行うのではなく、サーバ側で行うものがいくつも存在することが分かった。3.3 節で述べたように、構成ポリシの確認は 6 社のメータで、またスコア算出は 4 社のメータで確認された。これらのリモートサーバの URL を確認すると、いずれもメータの URL と同じドメインのサーバに対して送られていた。利用者にしてみれば、構成ポリシの確認とスコア算出がローカルで行われているかリモートで行われているかについてを判断することは難しい。多くの場合、利用者はパスワード 1 文字 1 文字の入力に対してリモートのサーバに情報が送られているとは思っていないことも考えられる。

またリモートへの通信は脅威も含むことに注意が必要である。ローカルの作業でポリシ確認とスコア算出がされるという前提が敷ければ、通信を観測するおとでパスワードのデータが（サービス提供者含む）外部に不要に送られていることが監視・検知が可能であるが、リモートの通信を含むとなると、それが正しい挙動かどうかの判断が難しくなる。実際、現代のブラウザにはさまざまな拡張機能をユーザ側で導入することができ、用途や悪意のある開発者の存在を考えると、拡張機能によりパスワード入力がフックされそのデータがサービス提供者と関係ない第 3 者に漏えいすることも考えられる。この場合、その通信が正しい挙動なのかどうかを利用者側が判断することは非常に難しいだろう。今回の調査では同一ドメインに対しデータを行っていたため、判断することは可能であるが、同一ドメインへの送信は強制されているわけではなく、今後のサービス提供者の動向によては第 3 者ドメインとも思えるサーバとも通信することは十分に考えられる。パスワードが利用者から判断の難しいリモートへ送られることに関しては、十分な注意が必要となるだろう。

6. まとめ

本稿では、Alexa Top 100 のサイトで利用されているパスワード強度メータについて、その挙動の分析と、実際のパスワードデータセットを用いた強度のスコア分布の調査と分析を行った。調査の結果、多くのサイトで強度メータが利用されていることが判明した一方で、そのスコア計算手法は統一ないし類似した手法ではなくそれぞれが独自の手法で計算しており、同じパスワードデータセットを用いてもそのスコア平均や分散、分布が大きくことなることがわかった。これは Dell'Amico らが指摘していた、多くのサイトが採用しているスコア計算方法はパスワードが持つ本来の情報量を反映したものになっていないことの裏付けと言えよう。

今後は、それらメータが実際の情報量との乖離がどれほどあるかのさらなる調査がされるとともに、その是正が求められていくと考えられる。また副次的に得られた結果として、パスワードのデータを逐次リモートに送信するサービスが複数観測された。これにより将来的な脅威の可能性が判明されたことも今後議論されていく点であると考えられる。

謝辞 本研究の一部は JSPS 科研費 JP16H02813 の助成を受けたものです。

参考文献

[1] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In Proceedings of the 2012 IEEE Symposium on Security and Privacy, SP '12, pages 553-567, Washington, DC, USA, 2012. IEEE Computer Society.
[2] M. Dell'Amico and M. Filippone. Monte carlo strength

evaluation: Fast and reliable password checking. In Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15, pages 158-169, New York, NY, USA, 2015. ACM.

[3] W. E. Burr, D. F. Dodson, E. M. Newton, R. A. Perlner, W. T. Polk, S. Gupta, E. A. Nabbus, U. D. of Commerce, N. I. of Standards, and Technology. Electronic Authentication Guideline: Recommendations of the National Institute of Standards and Technology - Special Publication 800-63-1. CreateSpace Independent Publishing Platform, USA, 2012.

[4] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In Proceedings of the 2012 IEEE Symposium on Security and Privacy, SP '12, pages 538-552, Washington, DC, USA, 2012. IEEE Computer Society.

[5] M. Weir, S. Aggarwal, B. d. Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. In Proceedings of the 2009 30th IEEE Symposium on Security and Privacy, SP '09, pages 391-405, Washington, DC, USA, 2009. IEEE

[6] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In Security and Privacy (SP), 2012 IEEE Symposium on, pages 523-537, May 2012.

[7] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How does your password measure up? the effect of strength meters on password creation. In Proceedings of the 21st USENIX Conference on Security Symposium, Security'12, pages 5-5, Berkeley, CA, USA, 2012. USENIX Association.

[8] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS '10, pages 162-175, New York, NY, USA, 2010. ACM.

[9] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. Does my password go up to eleven?: The impact of password meters on password selection. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, pages 2379-2388, New York, NY, USA, 2013. ACM.

[10] X. de Carn de Carnavalet and M. Mannan. From very weak to very strong: Analyzing password-strength meters. In Network and Distributed System Security Symposium (NDSS 2014). Internet Society, 2014.

[11] Z. Li, W. Han, and W. Xu. A large-scale empirical analysis of chinese web passwords. In 23rd USENIX Security Symposium (USENIX Security 14), pages 559-574, San Diego, CA, Aug. 2014. USENIX Association.

[12] B. Ur, S. M. Segreti, L. Bauer, N. Christin, L. F. Cranor, S. Komanduri, D. Kurilova, M. L. Mazurek, W. Melicher, and R. Shay. Measuring real-world accuracies and biases in modeling password guessability. In 24th USENIX Security Symposium (USENIX Security 15), pages 463-481, Washington, D.C., Aug. 2015. USENIX Association.

[13] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS '10, pages 162-175, New York, NY, USA, 2010. ACM.

[14] C. Castelluccia, M. Drmuth, and D. Perito. Adaptive password-strength meters from markov models. In NDSS. The Internet Society, 2012.

[15] Alexa, The top 500 sites on the web, `http://www.alexa.com/topsites`

[16] Dropbox, "dropbox/zxcvbn: A realistic password strength estimaor", github, `https://github.com/dropbox/zxcvbn`

[17] Dan Wheeler, "zxcvbn: realistic password strength estimation", Dropbox TechBlog, `https://blogs.dropbox.com/tech/2012/04/zxcvbn-realistic-password-strength-estimation/`, 2012