comes uniform, which is actually desirable for passwords. These observations may be of use to algorithm designers, for dimensioning data structures or even taking advantage of the relatively heavy-tailed nature of users' choices.

We also see that fitting a distribution provides relatively good approximations of the Shannon Entropy, guesswork and other statistics that are of interest when assessing a password distribution. Using a uniform model, where all passwords are equally likely, provides reasonable approximations for the data sets with smaller s, but provides a poor estimate of min-Entropy.

We have seen that demography of the userbase choosing the passwords can be evident in the most popular passwords, and even the name of the website is a likely password. Some sites, for example Twitter, have noticed this and implement banned password lists [16], which includes many of the more common passwords, including the name of the site. This gives weight to the advice that site administrators checking the security using password cracking software should include custom dictionaries including locally used terms.

The Zipf distribution decays relatively slowly, so we expect there to be a large number of relatively commonly chosen passwords. We investigated if these passwords vary significantly from site to site. We see that the lists of passwords from each site have quite a lot in common. While they do not provide the optimal order for guessing, the larger lists provide good guidance about the ranking of passwords in other lists. We've demonstrated that this can provide a significant speedup in guessing or cracking passwords using moderate numbers of guesses, particularly over simple dictionary attack, but also over a range of the guess-generating techniques described in [4].

An attacker can gain a useful starting point for cracking passwords if they collect leaked passwords. If a hashed password is exposed, the time for an attacker to hash, say, 20 million passwords is relatively small, even on a single CPU. We note that this adds extra weight to the advice that reusing passwords between websites is a risk, even if there is no way for an attacker to identify which pairs of users are common to the websites. This is because if just one site stores the password in plaintext format and that password is leaked, then it facilitates the subsequent cracking of that password on systems where the passwords are hashed.

Banning more commonly chosen passwords may result in a more even spread of password in use. Interestingly, we saw that most English dictionary words are not necessarily common passwords: out of more than 220,000 dictionary words, less than 15,000 appeared as passwords in the Gawker data set. We proposed a scheme based on the Metropolis-Hastings algorithm that aims to generate more uniform password choices, without having to know a list of common passwords in advance. A basic implementation of this is relatively straight forward, and could be easily incorporated into a password management system or PAM module [14].

8. CONCLUSION

We have seen that a Zipf distribution is a relatively good match for the frequencies with which users choose passwords. We have also seen that the passwords found in the lists that we have studied have relatively similar orderings. Consequently, passwords from one list provide good candidates when guessing or cracking passwords from another list. Finally, we presented a scheme that can guide users to distribute their passwords more uniformly.

Acknowledgment: We thank Ken Duffy and Niall Murphy for thought-provoking comments and Dermot Frost for an offer of CPU cycles. The authors were supported by NUIM SPUR, SFI 07/SK/I1216a and 08/SRC/I1403.

9. REFERENCES

[1] L. A. Adamic. Zipf, power-laws, and pareto-a ranking tutorial. Xerox Palo Alto Research Center, `http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html`, 2000.

[2] E. Arikan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42, Janurary 1996.

[3] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[4] M. Dell'Amico, P. Michiardi, and Y. Roudier. Password strength: An empirical analysis. In *INFOCOM*, pages 1–9, 2010.

[5] D. E. Eastlake, J. I. Schiller, and S. Crocker. RFC 4086: Randomness requirements for security. pages 1–47, 2005.

[6] W. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97, 1970.

[7] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms*, pages 133–144. ACM, 2009.

[8] D. Malone and W. Sullivan. Guesswork is not a substitute for entropy. In *Proceedings of the Information Technology and Telecommunications Conference*, 2005.

[9] D. Malone and W. G. Sullivan. Guesswork and entropy. *IEEE Transactions on Information Theory*, 50(3):525–526, 2004.

[10] J. L. Massey. Guessing and entropy. In *In Proceedings of the 1994 IEEE International Symposium on Information Theory*, page 204, 1994.

[11] M. McDowell, J. Rafail, and S. Hernan. Cyber security tip. `http://www.us-cert.gov/cas/tips/ST04-002.html`, 2009.

[12] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087, 1953.

[13] J. O. Pliam. The disparity between work and entropy in cryptology. *Theory of Cryptography Library: Record*, pages 98–24, 1998.

[14] V. Samar. Unified login with pluggable authentication modules (PAM). In *Proc. CCS'96*, pages 1–10, 1996.

[15] S. Schechter, C. Herley, and M. Mitzenmacher. Popularity is everything: a new approach to protecting passwords from statistical-guessing attacks. In *Proc. HotSec'10*, pages 1–6, 2010.

[16] twitter.com. Source code from twitter registration page. `view-source:https://twitter.com/signup` (search for twttr.BANNED_PASSWORDS), 2010.