

GUESSING HUMAN-CHOSEN SECRETS

Joseph Bonneau

jcb82@cl.cam.ac.uk



UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

MICROSOFT RESEARCH

CAMBRIDGE, UK

APRIL 5, 2012

Password research? In 2012??



Compatible Time-Sharing System, MIT 1961

What are passwords good for?

property	password
<i>Memorywise-Effortless</i>	
<i>Scalable-for-Users</i>	
<i>Nothing-to-Carry</i>	●
<i>Physically-Effortless</i>	
<i>Easy-to-Learn</i>	●
<i>Efficient-to-Use</i>	●
<i>Infrequent-Errors</i>	○
<i>Easy-Recovery-from-Loss</i>	●
<i>Accessible</i>	●
<i>Negligible-Cost-per-User</i>	●
<i>Server-Compatible</i>	●
<i>Browser-Compatible</i>	●
<i>Mature</i>	●
<i>Non-Proprietary</i>	●
<i>Resilient-to-Physical-Observation</i>	
<i>Resilient-to-Targeted-Impersonation</i>	○
<i>Resilient-to-Throttled-Guessing</i>	
<i>Resilient-to-Unthrottled-Guessing</i>	
<i>Resilient-to-Internal-Observation</i>	
<i>Resilient-to-Leaks-from-Other-Verifiers</i>	
<i>Resilient-to-Phishing</i>	
<i>Resilient-to-Theft</i>	●
<i>No-Trusted-Third-Party</i>	●
<i>Requiring-Explicit-Consent</i>	●
<i>Unlinkable</i>	●

Automated password managers?

property	password	Firefox
<i>Memorywise-Effortless</i>		○
<i>Scalable-for-Users</i>		●
<i>Nothing-to-Carry</i>	●	○
<i>Physically-Effortless</i>		○
<i>Easy-to-Learn</i>	●	●
<i>Efficient-to-Use</i>	●	●
<i>Infrequent-Errors</i>	○	●
<i>Easy-Recovery-from-Loss</i>	●	■
<i>Accessible</i>	●	●
<i>Negligible-Cost-per-User</i>	●	●
<i>Server-Compatible</i>	●	●
<i>Browser-Compatible</i>	●	■
<i>Mature</i>	●	●
<i>Non-Proprietary</i>	●	●
<i>Resilient-to-Physical-Observation</i>		○
<i>Resilient-to-Targeted-Impersonation</i>	○	○
<i>Resilient-to-Throttled-Guessing</i>		
<i>Resilient-to-Unthrottled-Guessing</i>		
<i>Resilient-to-Internal-Observation</i>		
<i>Resilient-to-Leaks-from-Other-Verifiers</i>		
<i>Resilient-to-Phishing</i>		●
<i>Resilient-to-Theft</i>	●	●
<i>No-Trusted-Third-Party</i>	●	●
<i>Requiring-Explicit-Consent</i>	●	●
<i>Unlinkable</i>	●	●

Graphical passwords?

property	password	Firefox	PCCP
<i>Memorywise-Effortless</i>			
<i>Scalable-for-Users</i>			
<i>Nothing-to-Carry</i>	●		●
<i>Physically-Effortless</i>			
<i>Easy-to-Learn</i>	●	●	●
<i>Efficient-to-Use</i>	●	●	
<i>Infrequent-Errors</i>	○		○
<i>Easy-Recovery-from-Loss</i>	●		●
<i>Accessible</i>	●	●	
<i>Negligible-Cost-per-User</i>	●	●	●
<i>Server-Compatible</i>	●	●	
<i>Browser-Compatible</i>	●		●
<i>Mature</i>	●	●	
<i>Non-Proprietary</i>	●	●	●
<i>Resilient-to-Physical-Observation</i>			
<i>Resilient-to-Targeted-Impersonation</i>	○	○	
<i>Resilient-to-Throttled-Guessing</i>			
<i>Resilient-to-Unthrottled-Guessing</i>			
<i>Resilient-to-Internal-Observation</i>			
<i>Resilient-to-Leaks-from-Other-Verifiers</i>			
<i>Resilient-to-Phishing</i>			
<i>Resilient-to-Theft</i>	●	●	●
<i>No-Trusted-Third-Party</i>	●	●	●
<i>Requiring-Explicit-Consent</i>	●	●	●
<i>Unlinkable</i>	●	●	●

Smartphones as a hardware token?

property	password	Firefox	PCCP	SMS
<i>Memorywise-Effortless</i>				
<i>Scalable-for-Users</i>				
<i>Nothing-to-Carry</i>				
<i>Physically-Effortless</i>				
<i>Easy-to-Learn</i>				
<i>Efficient-to-Use</i>				
<i>Infrequent-Errors</i>				
<i>Easy-Recovery-from-Loss</i>				
<i>Accessible</i>				
<i>Negligible-Cost-per-User</i>				
<i>Server-Compatible</i>				
<i>Browser-Compatible</i>				
<i>Mature</i>				
<i>Non-Proprietary</i>				
<i>Resilient-to-Physical-Observation</i>				
<i>Resilient-to-Targeted-Impersonation</i>				
<i>Resilient-to-Throttled-Guessing</i>				
<i>Resilient-to-Unthrottled-Guessing</i>				
<i>Resilient-to-Internal-Observation</i>				
<i>Resilient-to-Leaks-from-Other-Verifiers</i>				
<i>Resilient-to-Phishing</i>				
<i>Resilient-to-Theft</i>				
<i>No-Trusted-Third-Party</i>				
<i>Requiring-Explicit-Consent</i>				
<i>Unlinkable</i>				

Biometrics?

property	password	Firefox	PCCP	SMS	Iris
<i>Memorywise-Effortless</i>					
<i>Scalable-for-Users</i>					
<i>Nothing-to-Carry</i>					
<i>Physically-Effortless</i>					
<i>Easy-to-Learn</i>					
<i>Efficient-to-Use</i>					
<i>Infrequent-Errors</i>					
<i>Easy-Recovery-from-Loss</i>					
<i>Accessible</i>					
<i>Negligible-Cost-per-User</i>					
<i>Server-Compatible</i>					
<i>Browser-Compatible</i>					
<i>Mature</i>					
<i>Non-Proprietary</i>					
<i>Resilient-to-Physical-Observation</i>					
<i>Resilient-to-Targeted-Impersonation</i>					
<i>Resilient-to-Throttled-Guessing</i>					
<i>Resilient-to-Unthrottled-Guessing</i>					
<i>Resilient-to-Internal-Observation</i>					
<i>Resilient-to-Leaks-from-Other-Verifiers</i>					
<i>Resilient-to-Phishing</i>					
<i>Resilient-to-Theft</i>					
<i>No-Trusted-Third-Party</i>					
<i>Requiring-Explicit-Consent</i>					
<i>Unlinkable</i>					

Federated authentication?

property	password	Firefox	PCCP	SMS	Iris	OpenID
<i>Memorywise-Effortless</i>						
<i>Scalable-for-Users</i>						
<i>Nothing-to-Carry</i>	●		●		●	●
<i>Physically-Effortless</i>						
<i>Easy-to-Learn</i>	●	●	●	●	●	
<i>Efficient-to-Use</i>	●	●				●
<i>Infrequent-Errors</i>						
<i>Easy-Recovery-from-Loss</i>	●		●			●
<i>Accessible</i>	●	●				●
<i>Negligible-Cost-per-User</i>	●	●	●			●
<i>Server-Compatible</i>	●	●				
<i>Browser-Compatible</i>	●		●	●		●
<i>Mature</i>	●	●		●		●
<i>Non-Proprietary</i>	●	●	●	●		●
<i>Resilient-to-Physical-Observation</i>						
<i>Resilient-to-Targeted-Impersonation</i>						
<i>Resilient-to-Throttled-Guessing</i>						
<i>Resilient-to-Unthrottled-Guessing</i>						
<i>Resilient-to-Internal-Observation</i>						
<i>Resilient-to-Leaks-from-Other-Verifiers</i>						
<i>Resilient-to-Phishing</i>						
<i>Resilient-to-Theft</i>	●	●	●			●
<i>No-Trusted-Third-Party</i>	●	●	●		●	
<i>Requiring-Explicit-Consent</i>	●	●	●	●		●
<i>Unlinkable</i>	●	●	●	●		

There is no clear alternative to passwords...

[illegible]

Joseph Bonneau, Cormac Herley, Paul van Oorschot, and Frank Stajano.

"The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes", *IEEE Security & Privacy*, 2012.

The “last mile” of authentication



What's harder to guess?

- 1 A Cambridge student's password or an Oxford student's?
- 2 A human-chosen 4-digit PIN or a random 3-digit PIN?
- 3 A password or a stranger's surname?

Semantic password evaluation

year	study	length	% digits	% special
1989	Riddle et al.	4.4	3.5	—
1992	Spafford	6.8	31.7	14.8
1999	Wu	7.5	25.7	4.1
1999	Zviran and Haga	5.7	19.2	0.7
2004	Campbell and Bryant	8.0	78.0	3.0
2006	Cazier and Medlin	7.4	35.0	1.3
2010	Devillers et al.	7.9	54.0	3.4
2009	RockYou data	7.9	44.4	2.6

Semantic password evaluation

Length Char.	94 Character Alphabet			10 char. alphabet		94 char alphabet
	No Checks	Dictionary Rule	Dict. & Comp. Rule			
1	4	-	-	3	3.3	6.6
2	6	-	-	5	6.7	13.2
3	8	-	-	7	10.0	19.8
4	10	14	16	9	13.3	26.3
5	12	17	20	10	16.7	32.9
6	14	20	23	11	20.0	39.5
7	16	22	27	12	23.3	46.1
8	18	24	30	13	26.6	52.7
10	21	26	32	15	33.3	65.9
12	24	28	34	17	40.0	79.0
14	27	30	36	19	46.6	92.2
16	30	32	38	21	53.3	105.4
18	33	34	40	23	59.9	118.5
20	36	36	42	25	66.6	131.7
22	38	38	44	27	73.3	144.7
24	40	40	46	29	79.9	158.0
30	46	46	52	35	99.9	197.2
40	56	56	62	45	133.2	263.4

NIST “entropy” formula

Evaluating passwords via cracking tools



Secure Purchase Accent Password Recovery Software



Purchase Accent software securely **online** and **offline**. To purchase, choose a product title and a type of license you'd like to pay for, then just follow the instructions.

Pay with credit cards, checks, money orders, PayPal, or bank transfers.

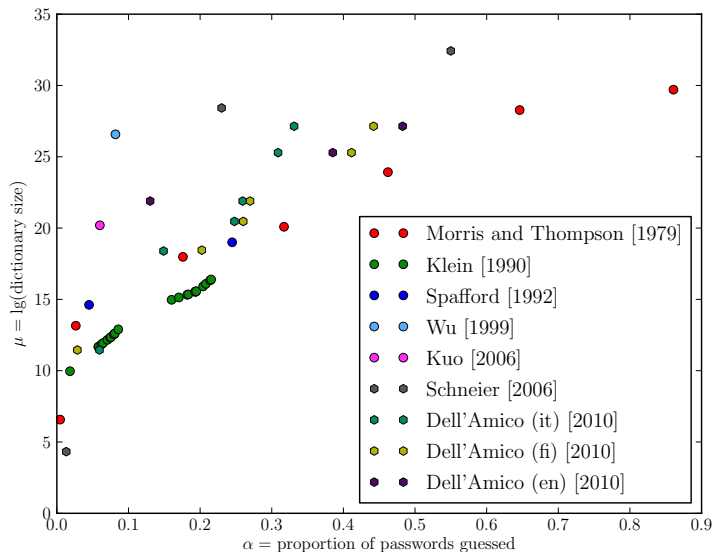
Read [Terms and Conditions](#).



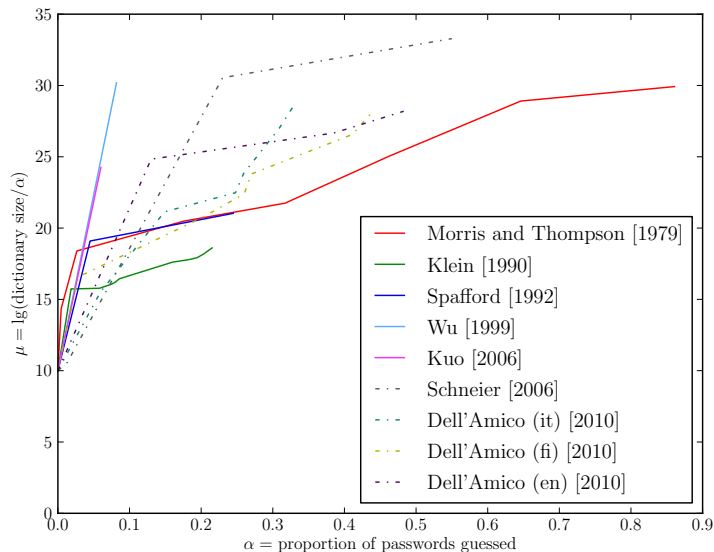
Prices:

- Standard Edition - **£1,399**
(you save £396)
- Forensic Edition - **£3,995**
(you save £1,895)
- Business Edition - **£13,995**
(you save £5,177)

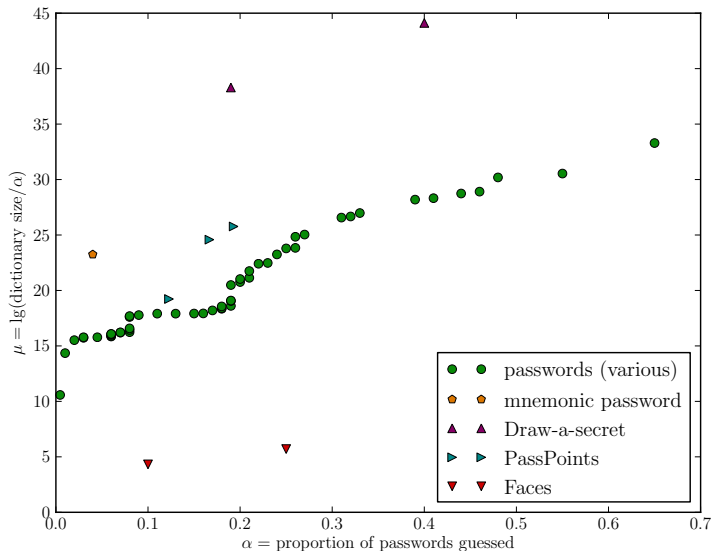
Evaluating passwords via cracking tools



Evaluating passwords via cracking tools



Evaluating passwords via cracking tools



Desiderata for evaluating human-chosen secrets

- sound
- repeatable
- no operator bias
- no demographic bias
- practical

Massive password data sets available for the first time

290729	123456
79076	12345
76789	123456789
59462	password
49952	iloveyou
33291	princess
21725	1234567
20901	rockyou
20553	12345678
16648	abc123

Dissertation: Guessing human-chosen secrets

- 1 Develop statistical guessing metrics
- 2 Collect large-scale data in a privacy-preserving manner
- 3 Approximate metrics using a random sample
- 4 Analyse data!

Develop statistical guessing metrics

- Assume a completely-known distribution $\mathcal{X}$
- $\mathcal{X}$ has N events (passwords) $x_1, x_2, \dots$
- Events have probability $p_1 \geq p_2 \geq \dots \geq p_N \geq 0$
- Each user chooses at random $X \stackrel{\text{R}}{\leftarrow} \mathcal{X}$

Question: How hard is it to guess X ?

$$H_1(\mathcal{X}) = - \sum_{i=1}^N p_i \lg p_i$$

Interpretation: Expected number of queries “Is $X \in \mathcal{S}$?” for arbitrary subsets $\mathcal{S} \subseteq \mathcal{X}$ needed to guess X . (Source-Coding Theorem)

Guesswork (guessing entropy)

$$G_1(\mathcal{X}) = E \left[\#_{\text{guesses}}(X \stackrel{R}{\leftarrow} \mathcal{X}) \right] = \sum_{i=1}^N p_i \cdot i$$

Intepretation: Expected number of queries “Is $X = x_i$?” for $i = 1, 2, \dots, N$ (optimal sequential guessing)

G_1 fails badly for real password distributions

```
1,ed65e09b98bdc70576d6c5f5e2ee38a9
1,e54d409c55499851aeb25713c1358484
1,dee489981220f2646eb8b3f412c456d9
1,c4df8d8e225232227c84d0ed8439428a
1,bd9059497b4af2bb913a8522747af2de
1,b25d6118ffc44b12b014feb81ea68e49
1,aac71eb7307f4c54b12c92d9bd45575f
1,9475d62e1f8b13676deab3824492367a
1,92965710534a9ec4b30f27b1e7f6062a
1,80f5a0267920942a73693596fe181fb7
1,76882fb85a1a8c6a83486aba03c031c9
1,6a60e0e51a3eb2e9fed6a546705de1bf
1,6843b9efec36f428deabce22c0fc1805
1,5dfbcd6390b77d06df9027c8d7d4fe84
1,4374968e935a9a0f8d56785ea682eb5f
1,0753929001291091112580111091991a
1,0410412e7194adc1b419a87a47979c36
...
```

Random 128-bit passwords in the wild at RockYou ($\sim 2^{-20}$)

Lemma: For a mixture distribution

$$\mathcal{Z} = p \cdot \mathcal{X} + q \cdot \mathcal{Y}$$

we have

$$G_1(\mathcal{Z}) \geq p \cdot G_1(\mathcal{X}) + q \cdot G_1(\mathcal{Y})$$

G_1 fails badly for real password distributions

Lemma: For a mixture distribution

$$\mathcal{Z} = p \cdot \mathcal{X} + q \cdot \mathcal{Y}$$

we have

$$G_1(\mathcal{Z}) \geq p \cdot G_1(\mathcal{X}) + q \cdot G_1(\mathcal{Y})$$

Therefore, for real passwords $\mathcal{P}$, we have:

$$G_1(\mathcal{P}) \geq p \cdot G_1(\mathcal{X}) + 2^{-20} \cdot G_1(\mathcal{U}_{2^{128}})$$

$$\boxed{G_1(\mathcal{P}) > 2^{107}}$$

Optimal attacks will give up on the hardest values

New goal: Correctly guess k of m values $X_1, X_2 \dots X_M \stackrel{R}{\leftarrow} \mathcal{X}$

Optimal attacks will give up on the hardest values

Name #1	Name #2	Name #3	...	Name # m
Smith	Smith	Smith	...	Smith
Jones	Jones	Jones	...	Jones
Johnson	Johnson	Johnson	...	Johnson
...	...	...	...	...
Ytterrock	Ytterrock	Ytterrock	...	Ytterrock
Zdrzynski	Zdrzynski	Zdrzynski	...	Zdrzynski

Optimal attacks will give up on the hardest values

Name #1	Name #2	Name #3	...	Name # m
Smith	Smith	Smith	...	Smith
Jones	Jones	Jones	...	Jones
Johnson	Johnson	Johnson	...	Johnson
...	...	...	...	...
Ytterrock	Ytterrock	Ytterrock	...	Ytterrock
Zdrzynski	Zdrzynski	Zdrzynski	...	Zdrzynski

Optimal attacks will give up on the hardest values

Name #1	Name #2	Name #3	...	Name # m
Smith	Smith	Smith	...	Smith
Jones	Jones	Jones	...	Jones
Johnson	Johnson	Johnson	...	Johnson
...	...	...	...	...
Ytterrock	Ytterrock	Ytterrock	...	Ytterrock
Zdrzynski	Zdrzynski	Zdrzynski	...	Zdrzynski

(obvious optimal strategy)

Partial guessing—model attackers who give up

- **β -success-rate**

Success rate after β guesses:

$$\lambda_{\beta}(\mathcal{X}) = \sum_{i=1}^{\beta} p_i$$

- **α -work-factor**

Dictionary size needed for α success probability:

$$\mu_{\alpha}(\mathcal{X}) = \min \left\{ \mu \in [1, N] \mid \sum_{i=1}^{\mu} p_i \geq \alpha \right\}$$

Partial guessing—model attackers who give up

- **β -success-rate**

Success rate after β guesses:

$$\lambda_{\beta}(\mathcal{X}) = \sum_{i=1}^{\beta} p_i$$

- **α -work-factor**

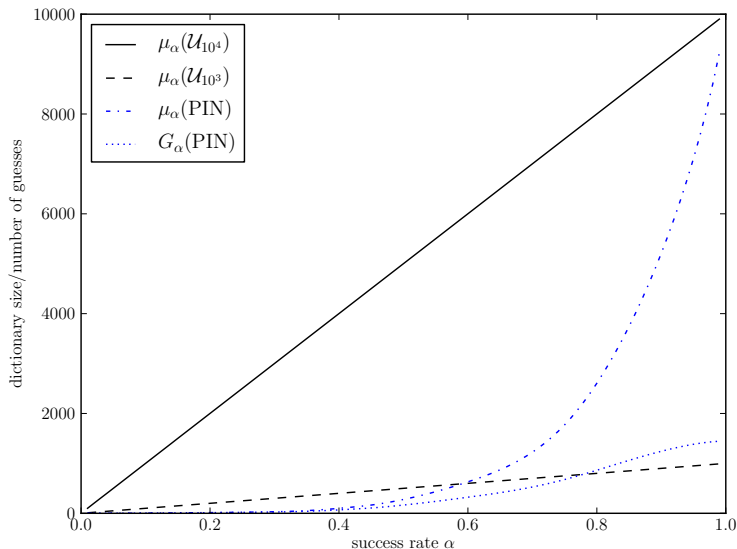
Dictionary size needed for α success probability:

$$\mu_{\alpha}(\mathcal{X}) = \min \left\{ \mu \in [1, N] \left| \sum_{i=1}^{\mu} p_i \geq \alpha \right. \right\}$$

$$G_{\alpha}(\mathcal{X}) = (1 - \lceil \alpha \rceil) \cdot \mu_{\alpha}(\mathcal{X}) + \sum_{i=1}^{\mu_{\alpha}(\mathcal{X})} p_i \cdot i$$

Interpretation: Expected number of guesses to succeed with probability α

Guessing curves visualise all possible attacks



For more on PIN statistics

Joseph Bonneau, Sören Preibusch and Ross Anderson.

A birthday present every eleven wallets? The security of customer-chosen banking PINs.

Financial Crypto 2012.

All metrics convertible to bits

- Find the size of a uniform distribution $\mathcal{U}_N$ with equivalent security
- Easy cases:

$$\tilde{\lambda}_\beta(\mathcal{X}) = \lg \left(\frac{\beta}{\lambda_\beta(\mathcal{X})} \right)$$

$$\tilde{\mu}_\alpha(\mathcal{X}) = \lg \left(\frac{\mu_\alpha(\mathcal{X})}{\lceil \alpha \rceil} \right)$$

- More complicated:

$$\tilde{G}_\alpha(\mathcal{X}) = \lg \left[\frac{2 \cdot G_\alpha(\mathcal{X})}{\lceil \alpha \rceil} - 1 \right] - \lg(2 - \lceil \alpha \rceil)$$

- Sanity check:

$$\tilde{\lambda}_\beta(\mathcal{U}_N) = \tilde{\mu}_\alpha(\mathcal{U}_N) = \tilde{G}_\alpha(\mathcal{U}_N) = \lg N$$

All metrics convertible to bits

- Find the size of a uniform distribution $\mathcal{U}_N$ with equivalent security
- Easy cases:

$$\tilde{\lambda}_\beta(\mathcal{X}) = \lg \left(\frac{\beta}{\lambda_\beta(\mathcal{X})} \right)$$

$$\tilde{\mu}_\alpha(\mathcal{X}) = \lg \left(\frac{\mu_\alpha(\mathcal{X})}{\lceil \alpha \rceil} \right)$$

- More complicated:

$$\tilde{G}_\alpha(\mathcal{X}) = \lg \left[\frac{2 \cdot G_\alpha(\mathcal{X})}{\lceil \alpha \rceil} - 1 \right] - \lg(2 - \lceil \alpha \rceil)$$

- Sanity check:

$$\tilde{\lambda}_\beta(\mathcal{U}_N) = \tilde{\mu}_\alpha(\mathcal{U}_N) = \tilde{G}_\alpha(\mathcal{U}_N) = \lg N$$

All metrics convertible to bits

- Find the size of a uniform distribution $\mathcal{U}_N$ with equivalent security
- Easy cases:

$$\tilde{\lambda}_\beta(\mathcal{X}) = \lg \left(\frac{\beta}{\lambda_\beta(\mathcal{X})} \right)$$

$$\tilde{\mu}_\alpha(\mathcal{X}) = \lg \left(\frac{\mu_\alpha(\mathcal{X})}{\lceil \alpha \rceil} \right)$$

- More complicated:

$$\tilde{G}_\alpha(\mathcal{X}) = \lg \left[\frac{2 \cdot G_\alpha(\mathcal{X})}{\lceil \alpha \rceil} - 1 \right] - \lg(2 - \lceil \alpha \rceil)$$

- Sanity check:

$$\tilde{\lambda}_\beta(\mathcal{U}_N) = \tilde{\mu}_\alpha(\mathcal{U}_N) = \tilde{G}_\alpha(\mathcal{U}_N) = \lg N$$

All metrics convertible to bits

- Find the size of a uniform distribution $\mathcal{U}_N$ with equivalent security
- Easy cases:

$$\tilde{\lambda}_\beta(\mathcal{X}) = \lg \left(\frac{\beta}{\lambda_\beta(\mathcal{X})} \right)$$

$$\tilde{\mu}_\alpha(\mathcal{X}) = \lg \left(\frac{\mu_\alpha(\mathcal{X})}{\lceil \alpha \rceil} \right)$$

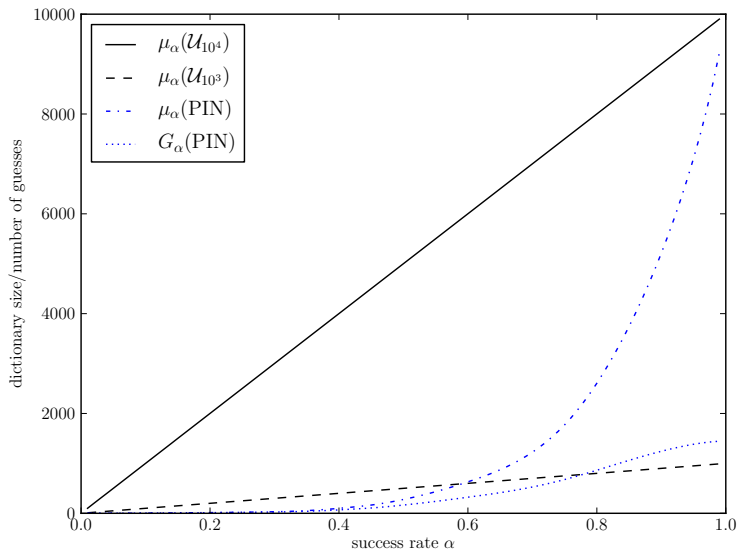
- More complicated:

$$\tilde{G}_\alpha(\mathcal{X}) = \lg \left[\frac{2 \cdot G_\alpha(\mathcal{X})}{\lceil \alpha \rceil} - 1 \right] - \lg(2 - \lceil \alpha \rceil)$$

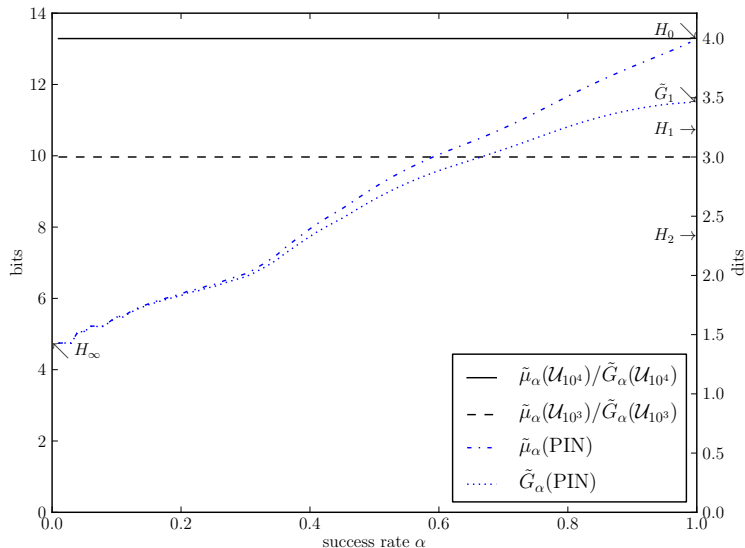
- **Sanity check:**

$$\tilde{\lambda}_\beta(\mathcal{U}_N) = \tilde{\mu}_\alpha(\mathcal{U}_N) = \tilde{G}_\alpha(\mathcal{U}_N) = \lg N$$

Guessing curves visualise all possible attacks



Guessing curves visualise all possible attacks



No single metric can represent all attacks

Non-comparability with H_1

Given any δ, α, β there exists a distribution $\mathcal{X}$ such that:

- $\tilde{\lambda}_\beta(\mathcal{X}) < H_1(\mathcal{X}) - \delta$
- $\tilde{\mu}_\alpha(\mathcal{X}) < H_1(\mathcal{X}) - \delta$
- $\tilde{G}_\alpha(\mathcal{X}) < H_1(\mathcal{X}) - \delta$

Non-comparability between different α

Given any $\delta, \alpha_1, \alpha_2$ there exists a distribution $\mathcal{X}$ such that:

- $\tilde{\mu}_{\alpha_1}(\mathcal{X}) < \tilde{\mu}_{\alpha_2}(\mathcal{X})$
- $\tilde{G}_{\alpha_1}(\mathcal{X}) < \tilde{G}_{\alpha_2}(\mathcal{X})$

Summary of guessing metrics

- H_1 doesn't apply to guessing
- G_1 is unrealistic
- $\tilde{G}_\alpha$ for $\alpha < 1$ is the best metric
- But...
 - $\tilde{\mu}_\alpha$ when comparing to cracking results
 - $\tilde{\lambda}_\beta$ for online attacks

Summary of guessing metrics

- H_1 doesn't apply to guessing
- G_1 is unrealistic
- $\tilde{G}_\alpha$ for $\alpha < 1$ is the best metric
- But...
 - $\tilde{\mu}_\alpha$ when comparing to cracking results
 - $\tilde{\lambda}_\beta$ for online attacks

Summary of guessing metrics

- H_1 doesn't apply to guessing
- G_1 is unrealistic
- $\tilde{G}_\alpha$ for $\alpha < 1$ is the best metric
- But...
 - $\tilde{\mu}_\alpha$ when comparing to cracking results
 - $\tilde{\lambda}_\beta$ for online attacks

Summary of guessing metrics

- H_1 doesn't apply to guessing
- G_1 is unrealistic
- $\tilde{G}_\alpha$ for $\alpha < 1$ is the best metric
- But...
 - $\tilde{\mu}_\alpha$ when comparing to cracking results
 - $\tilde{\lambda}_\beta$ for online attacks

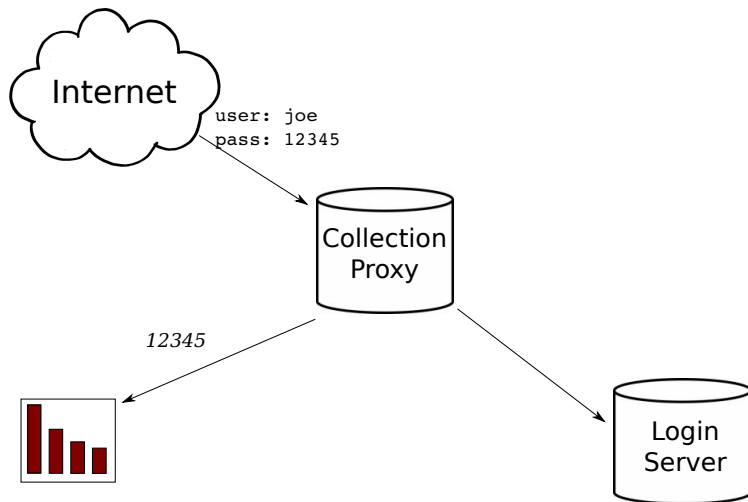
Collect large-scale data in a privacy-preserving manner

Whither the RockYou data?

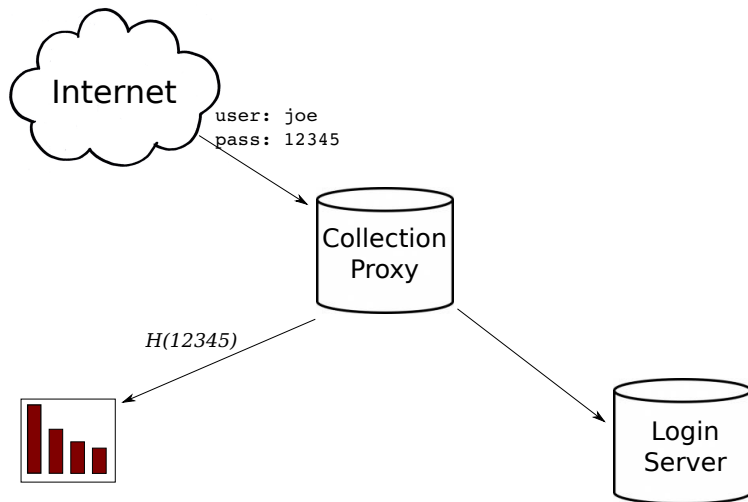
- No demographic info
- “Just” a photo-sharing website
- Ethical concerns:

Serge Egelman, Joseph Bonneau, Sonia Chiasson, David Dittrich and Stuart Schechter.
Its Not Stealing If You Need It: A Panel on The Ethics of Performing Research Using Public
Data of Illicit Origin. *2012 Workshop on Ethics in Computer Security Research*.

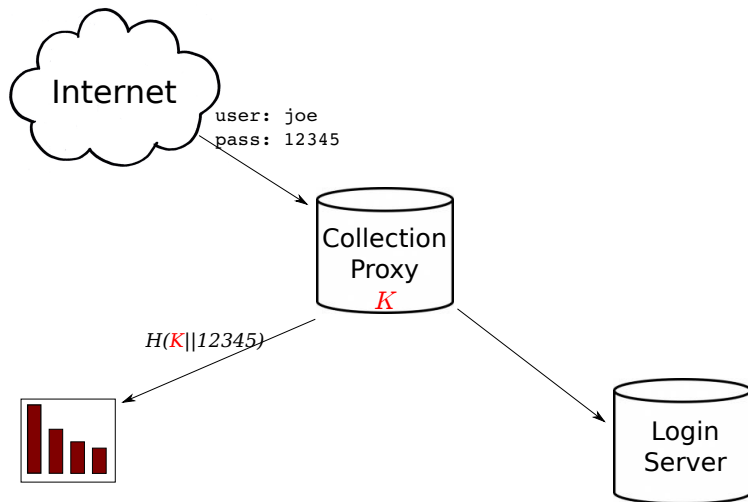
Collecting large-scale data at Yahoo!



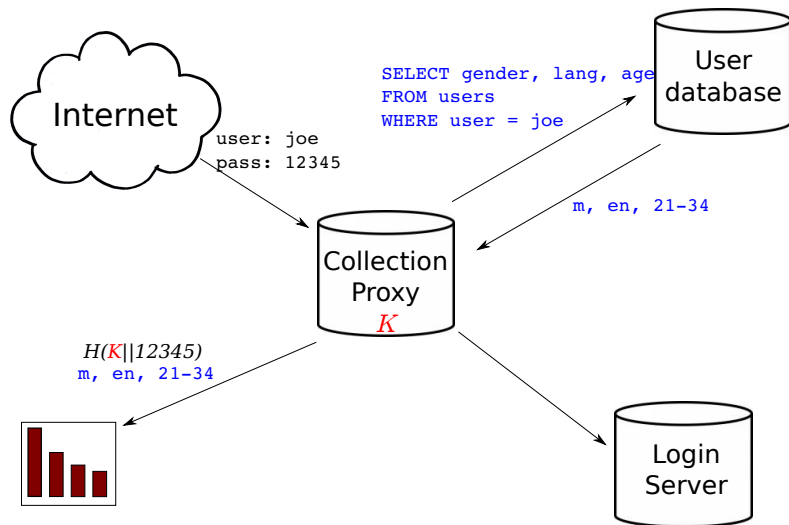
Collecting large-scale data at Yahoo!



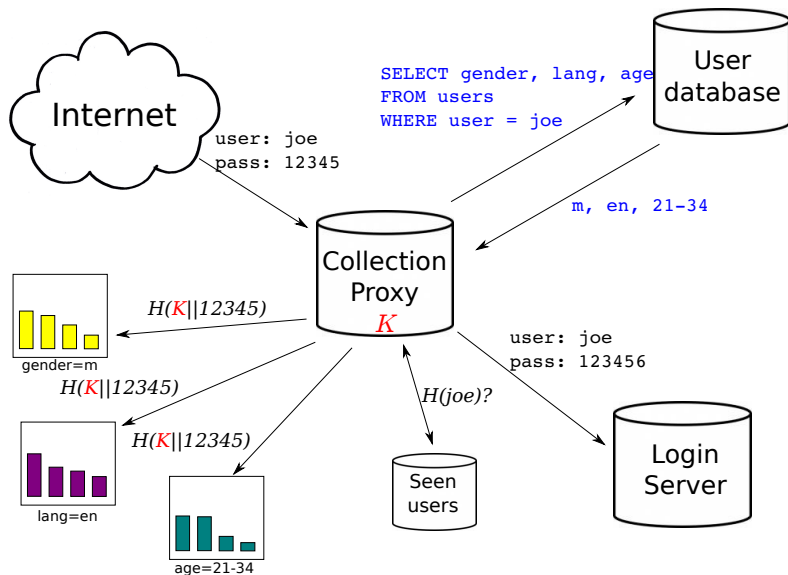
Collecting large-scale data at Yahoo!



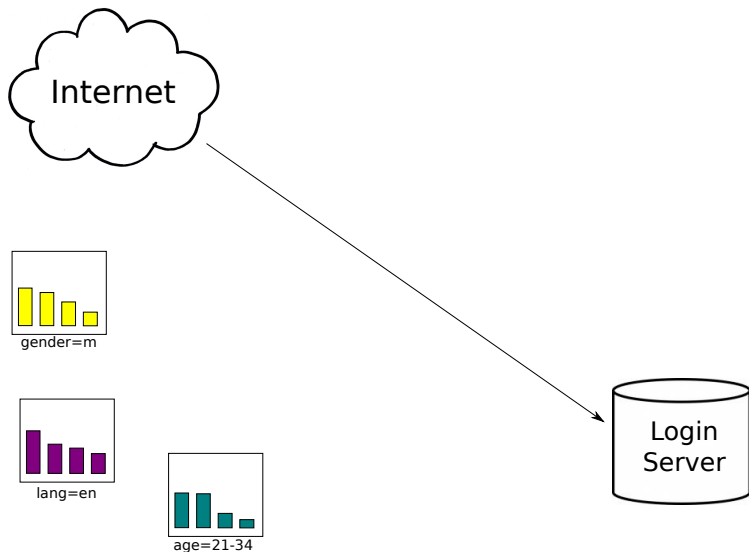
Collecting large-scale data at Yahoo!



Collecting large-scale data at Yahoo!



Collecting large-scale data at Yahoo!

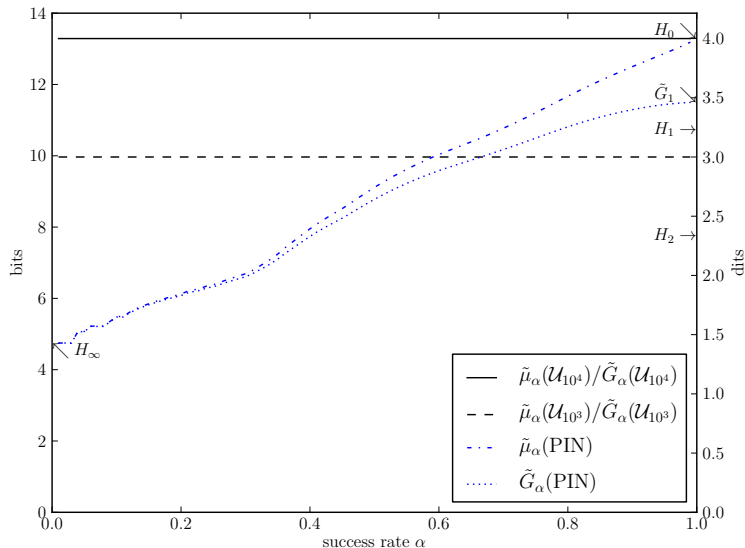


Collecting large-scale data at Yahoo!

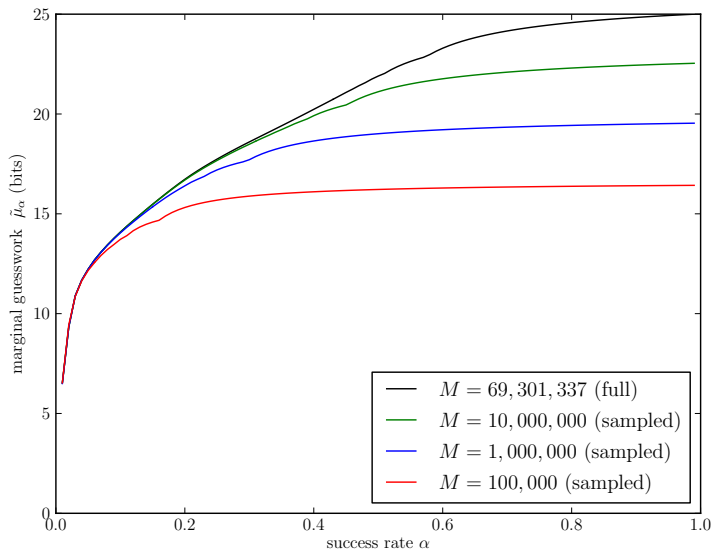
- Experiment run May 23–25, 2011
- 69,301,337 unique users
- 42.5% unique
- 328 different predicate functions

Approximate metrics using a random sample

Reminder: guessing curves



Sample size is a major problem for passwords...



Low-frequency blindness theorem (Valiant)

- Given a sample of size M
- Want to approximate f
- If there exist distributions $\mathcal{X}, \mathcal{Y}$ such that $f(\mathcal{X}) = f(\mathcal{Y}) = \delta$ and $\mathcal{X}, \mathcal{Y}$ are isomorphic for all events with $p \geq \frac{1}{M}$, then we can't compute f to within any accuracy δ using M samples
- Related: Canonical Testing Theorem

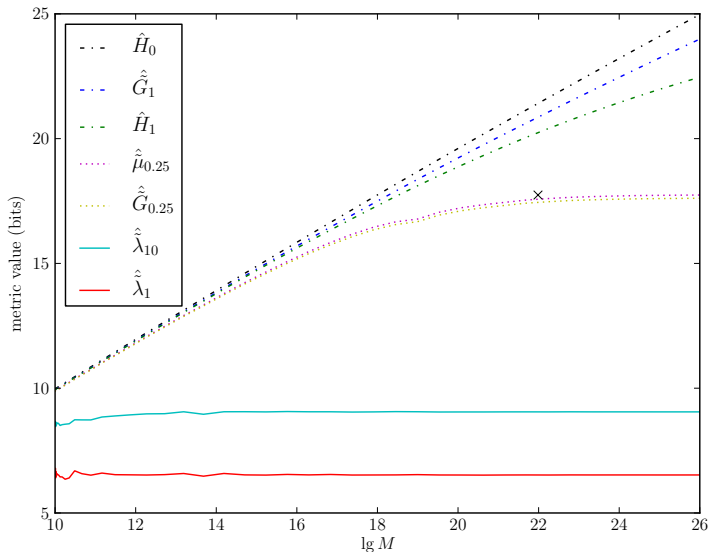
Low-frequency blindness theorem (Valiant)

- Given a sample of size M
- Want to approximate f
- If there exist distributions $\mathcal{X}, \mathcal{Y}$ such that $f(\mathcal{X}) = f(\mathcal{Y}) = \delta$ and $\mathcal{X}, \mathcal{Y}$ are isomorphic for all events with $p \geq \frac{1}{M}$, then we can't compute f to within any accuracy δ using M samples
- Related: Canonical Testing Theorem

Low-frequency blindness theorem (Valiant)

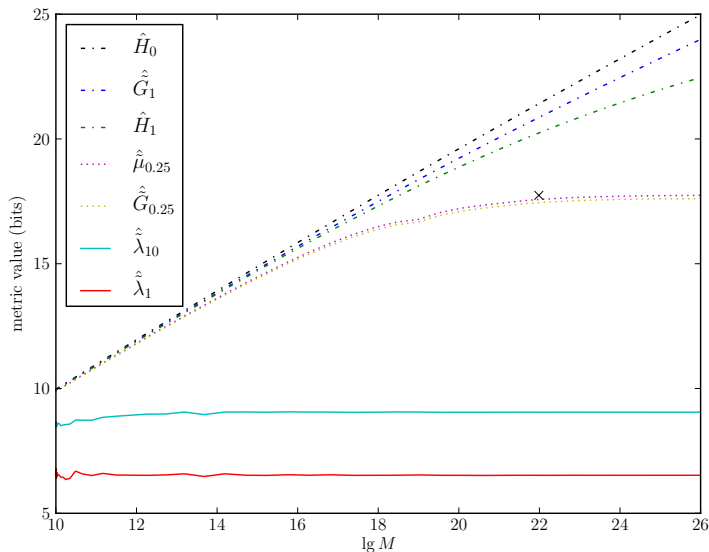
- Given a sample of size M
- Want to approximate f
- If there exist distributions $\mathcal{X}, \mathcal{Y}$ such that $f(\mathcal{X}) = f(\mathcal{Y}) = \delta$ and $\mathcal{X}, \mathcal{Y}$ are isomorphic for all events with $p \geq \frac{1}{M}$, then we can't compute f to within any accuracy δ using M samples
- Related: Canonical Testing Theorem

Theoretical results

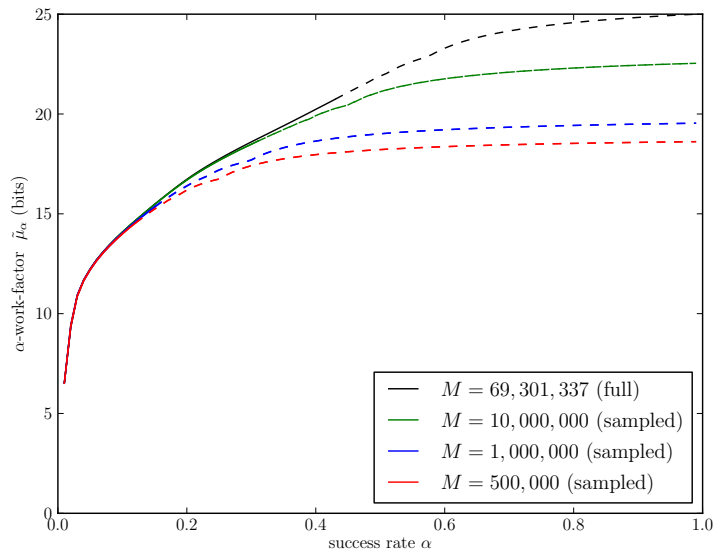


- No non-parametric method of estimating N
- Estimating H_1 requires $O(\frac{N}{\ln N})$ samples (Valiant & Valiant, 2011)

We can predict our confidence range by bootstrapping



We can predict our confidence range by bootstrapping



We can parametrically extend our approximation

Simple power-law distribution: $\Pr[p(x) > y] \propto y^{1-a}$

Unfortunately, a varies strongly with sample size:

M	69M	10M	1M	100k
$\hat{a}$	2.99	3.23	3.70	4.21

We can parametrically extend our approximation

- Generalized inverse-Gaussian Poisson (Sichel) distribution^a:

$$\psi(p|b, c, g) = \frac{2^{g-1} p^{g-1} e^{\frac{p}{c} - \frac{b^2 c}{4p}}}{(bc)^g \cdot K_g(b)}$$

- Probability of k observations given pdf ψ :

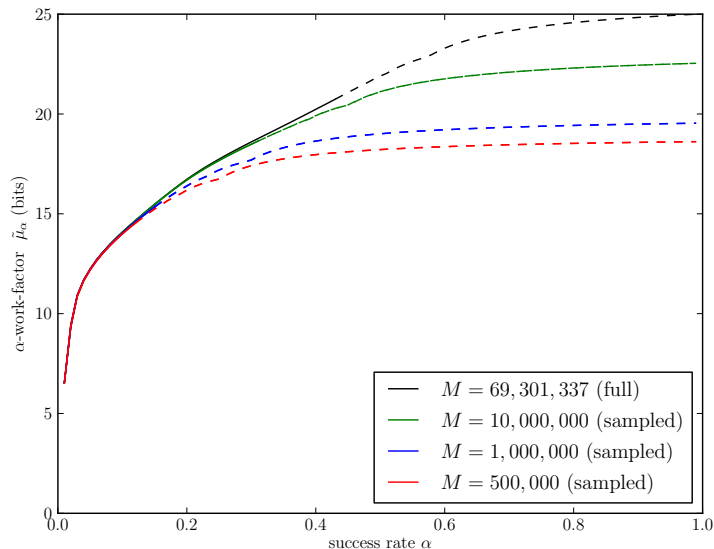
$$\Pr[k \text{ obs.}] = \frac{\int_0^1 \frac{(p \cdot M)^k \cdot e^{-p \cdot M}}{k!} \psi(p) dp}{1 - \int_0^1 e^{-p \cdot M} \psi(p) dp}$$

- optimize b, c, g by maximizing:

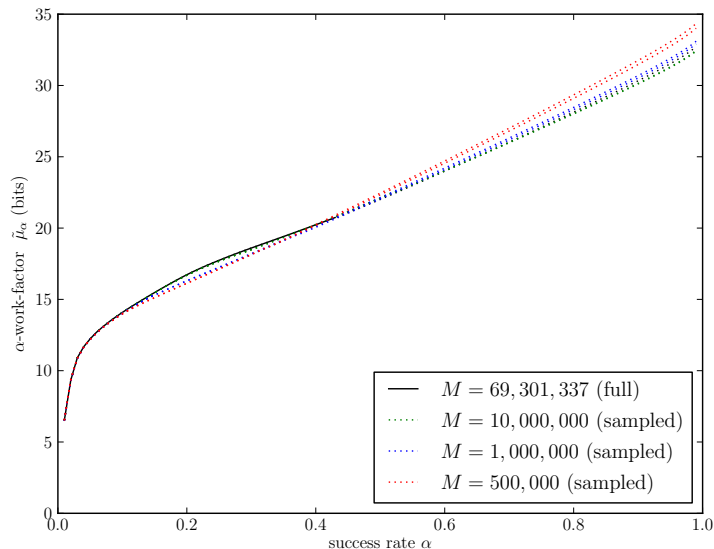
$$\text{Likelihood} = \prod_{i=1}^{\hat{N}} \Pr[f_i \text{ obs.}]$$

^a K_g is the modified Bessel function of the second kind.

We can parametrically extend our approximation

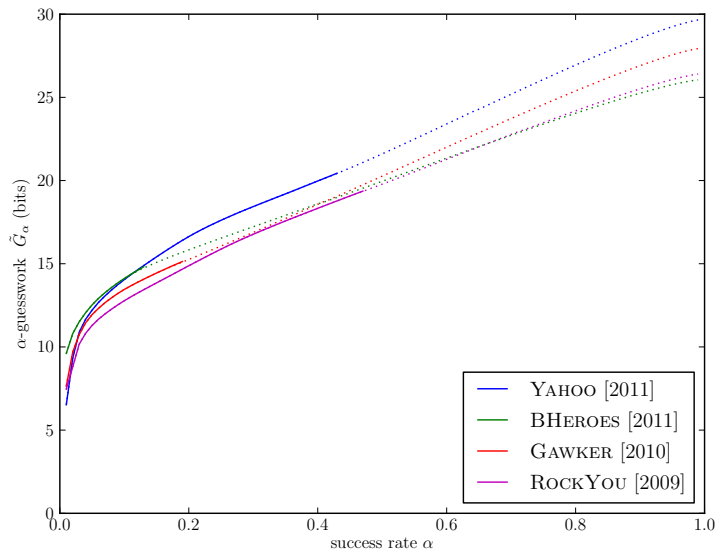


We can parametrically extend our approximation

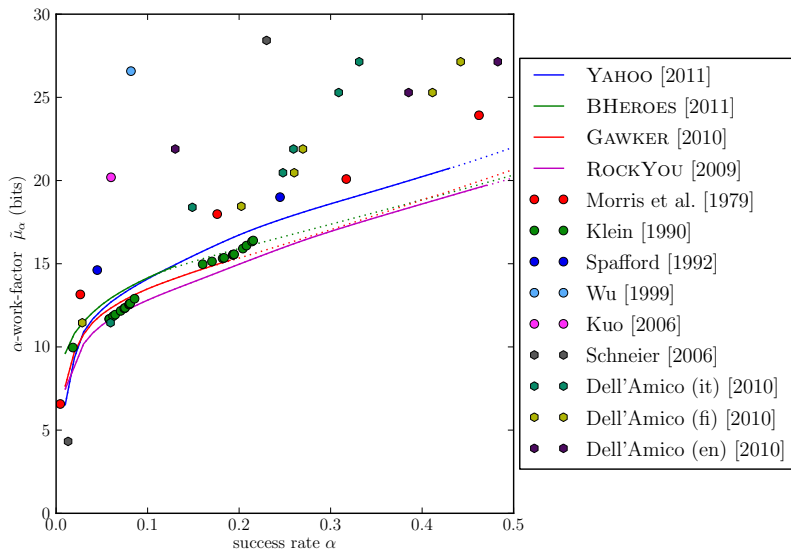


Analyse data

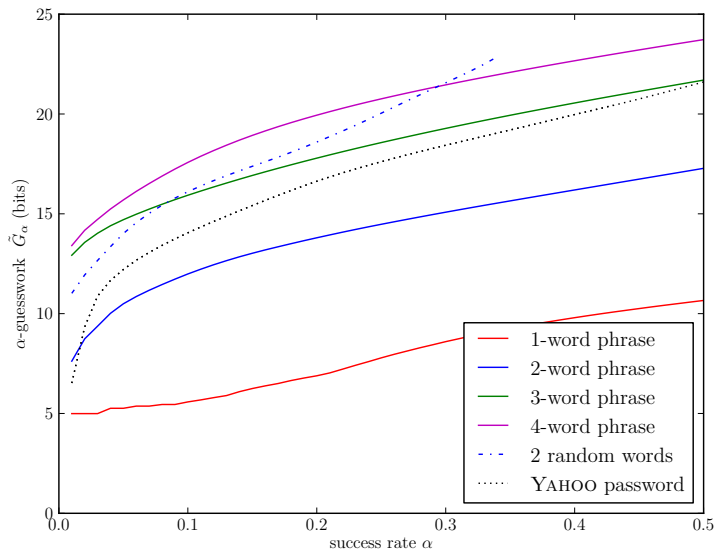
How secure are Yahoo! passwords



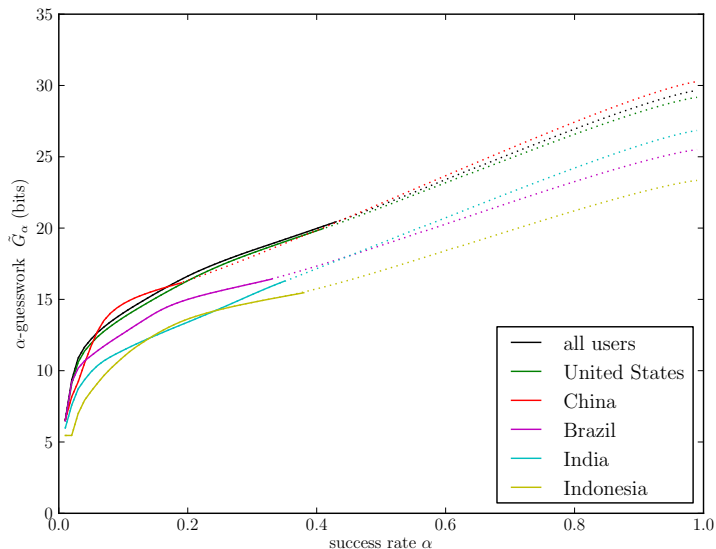
How secure are Yahoo! passwords



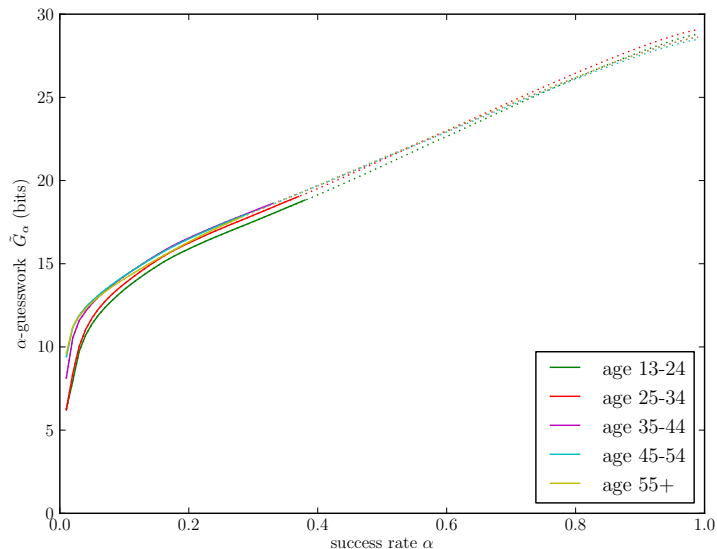
How secure are Yahoo! passwords



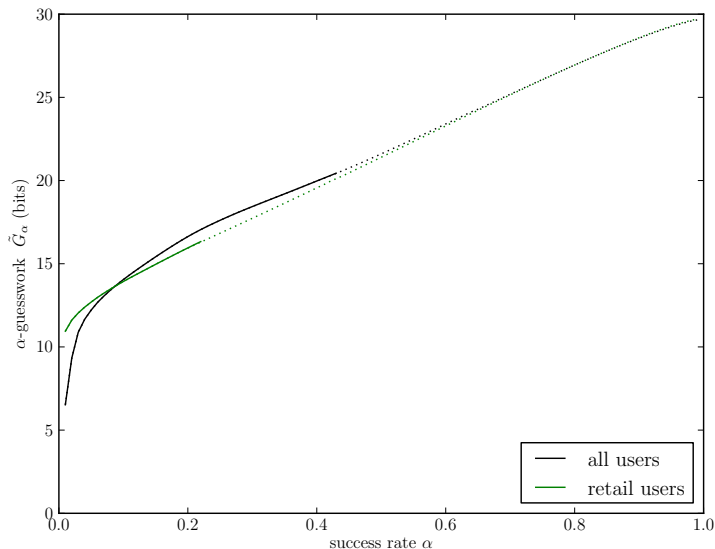
Demographic trends



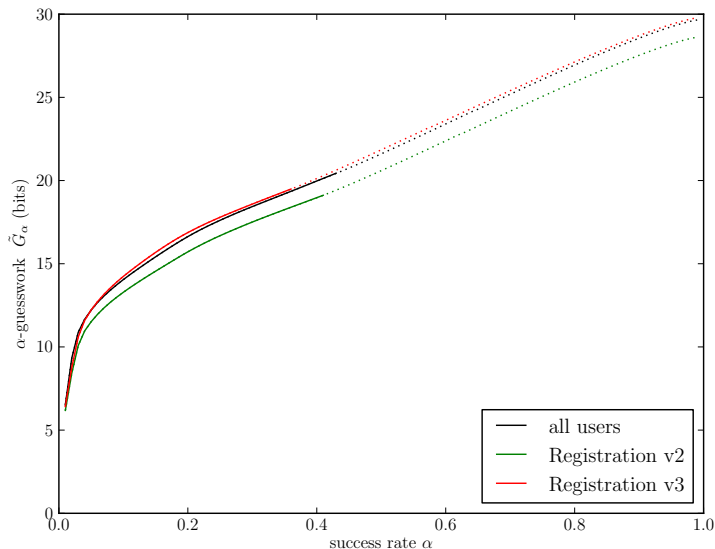
Demographic trends



Demographic trends



Demographic trends



Demographic summary

- Differences small but detectable
- Online attack $\tilde{\lambda}_{10} = 6\text{--}9$ bits
- Offline attack $\tilde{G}_{0.5} = 15\text{--}25$ bits
- There is no “good group” of users

Surprisingly little language variation

		dictionary										global
		de	en	es	fr	id	it	ko	pt	zh	vi	
target	de	6.5%	3.3%	2.6%	2.9%	2.2%	2.8%	1.6%	2.1%	2.0%	1.6%	3.5%
	en	4.6%	8.0%	4.2%	4.3%	4.5%	4.3%	3.4%	3.5%	4.4%	3.5%	7.9%
	es	5.0%	5.6%	12.1%	4.6%	4.1%	6.1%	3.1%	6.3%	3.6%	2.9%	6.9%
	fr	4.0%	4.2%	3.4%	10.0%	2.9%	3.2%	2.2%	3.1%	2.7%	2.1%	5.0%
	id	6.3%	8.7%	6.2%	6.3%	14.9%	6.2%	5.8%	6.0%	6.7%	5.9%	9.3%
	it	6.0%	6.3%	6.8%	5.3%	4.6%	14.6%	3.3%	5.7%	4.0%	3.2%	7.2%
	ko	2.0%	2.6%	1.9%	1.8%	2.3%	2.0%	5.8%	2.4%	3.7%	2.2%	2.8%
	pt	3.9%	4.3%	5.8%	3.8%	3.9%	4.4%	3.5%	11.1%	3.9%	2.9%	5.1%
	zh	1.9%	2.4%	1.7%	1.7%	2.0%	2.0%	2.9%	1.8%	4.4%	2.0%	2.9%
	vi	5.7%	7.7%	5.5%	5.8%	6.3%	5.7%	6.0%	5.8%	7.0%	14.3%	7.8%

For $\beta = 1000$, greatest efficiency loss is only 4.8 (fr/vi)

For more on Yahoo! passwords

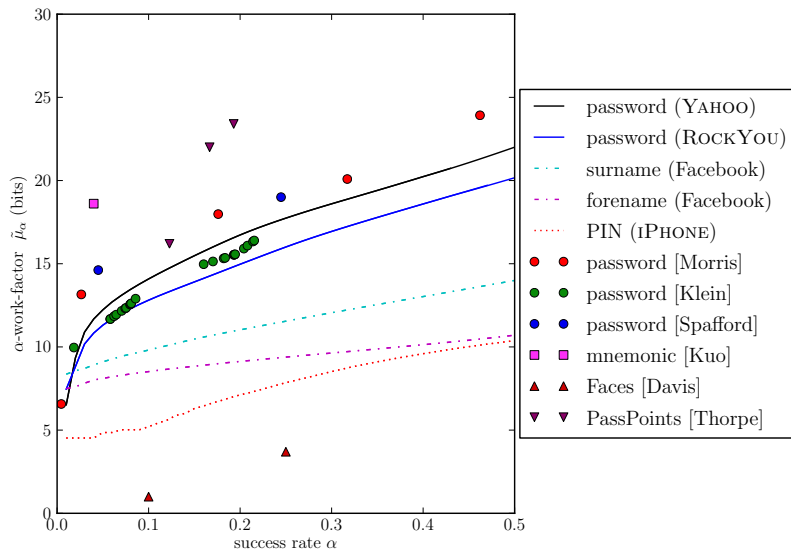
Joseph Bonneau.

The science of guessing: analyzing an anonymized corpus of 70 million passwords. *IEEE Security & Privacy*, 2012.

Statistical analysis of passwords

property	semantic	cracking	statistical
sound		●	○
no operator bias	●		●
no demographic bias			●
repeatable	●	○	●
practical	●	○	●
works with small data	●	●	

Putting it all together



The future of passwords

- Retire passwords from most current roles
- Avoid aggregating large databases of credentials
- Figure out what humans can actually remember...

A long road ahead

Results from a study of password authentication in the wild:

- 29–40% of websites don't hash passwords during storage
- 41% of websites don't use any encryption for password submission
 - 22% do so incompletely
- 84% of websites don't rate-limit against guessing attacks
- 97% of websites leak usernames to simple

Joseph Bonneau and Sören Preibusch.

The password thicket: technical and market failures in human authentication on the web.

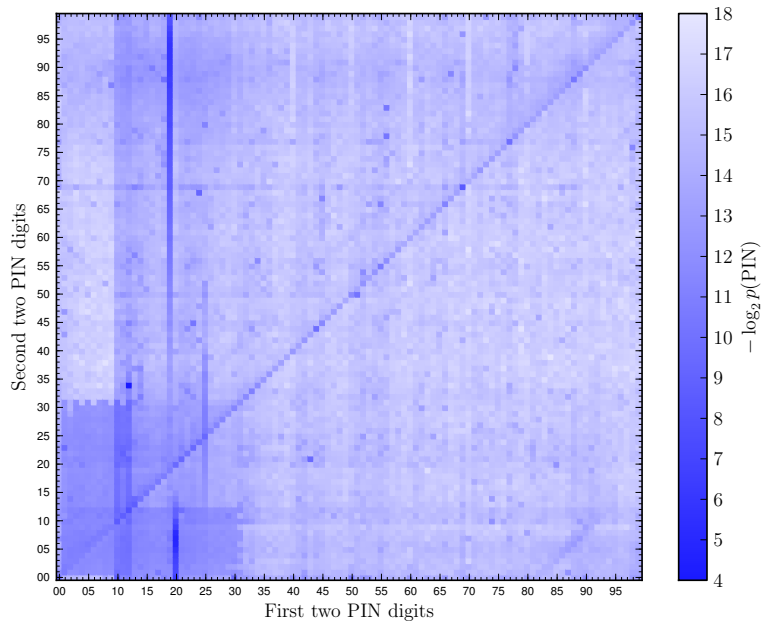
Workshop on the Economics of Information Security, 2010.

Thank you

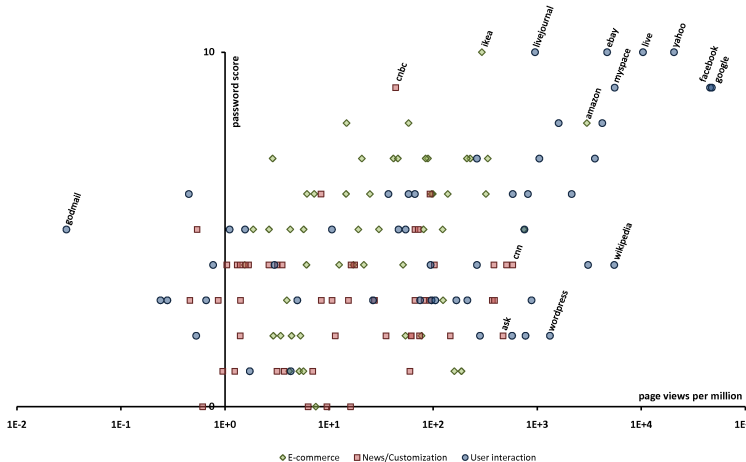
jcb82@cl.cam.ac.uk

Other projects on authentication

Human choice of PINs



The economics of password collection



пароль

=

пароль

=

%26%231087%3B%26%231072%3B%26%231088%3B
%26%231086%3B%26%231083%3B%26%231100%3B

6 $\rightarrow$ 42 $\rightarrow$ 78

Passwords and character encoding

пароль

=

пароль

=

%26%231087%3B%26%231072%3B%26%231088%3B
%26%231086%3B%26%231083%3B%26%231100%3B

6 → 42 → 78

The screenshot shows the IMDbPro website's 'Forgotten Password' page. At the top, there is a search bar and navigation links for 'Careers', 'Industry Directory', and 'In Production'. Below the navigation bar, the page title is 'Forgotten Password' with a subtitle 'Please choose a new password.' A red error message states 'Password too long (max. 64 characters)'. There are two input fields: 'Password:' and 'Confirm Password:', both containing six dots. A 'Go' button is visible in the top right corner of the page header.