Human-generated secret data

Joseph Bonneau

jcb82@cl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Security and Human Behaviour

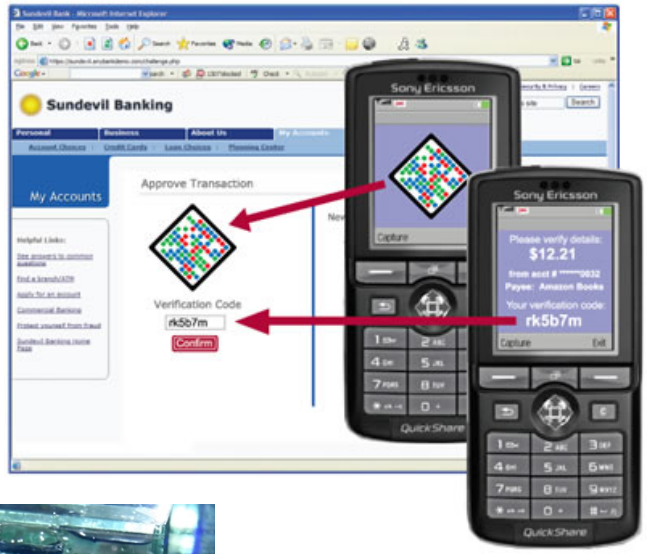
Cambridge, UK

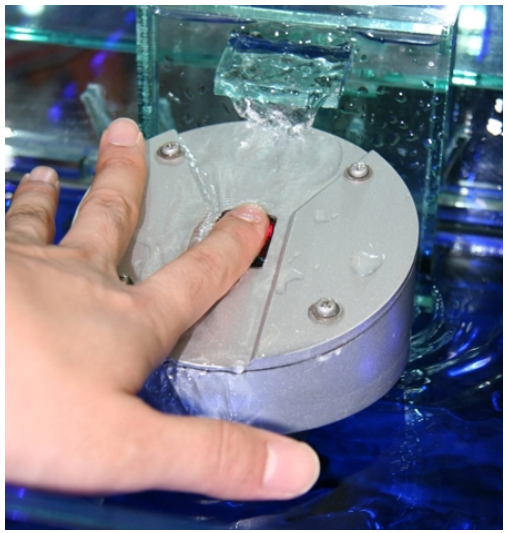
June 29, 2010

The Simple English guide to human-generated secrets

1. Computers try to tell humans apart by asking for secret memories. They can ask for other things, but those are very expensive.

Two-factor authentication remains far too expensive





The Simple English guide to human-generated secrets

1. Computers try to tell humans apart by asking for secret data. They can ask for other things, but these are very expensive.
2. Many computer scientists use something called “entropy” to measure security for this secret data, but there are a lot of mathematical equations which say this is a bad idea.

Measuring Security Against Guessing

Which is “harder” to guess:

- Surname of randomly chosen Internet user
- Randomly chosen 4-digit PIN

Measuring Security Against Guessing

Which is “harder” to guess:

- Surname of randomly chosen Internet user
  - H_1(surname) = **16.2 bits**
- Randomly chosen 4-digit PIN
  - H_1(PIN) = **13.3 bits**

Shannon Entropy

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

- H_1(surname) = **16.2 bits**
- H_1(PIN) = **13.3 bits**

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

Guessing Entropy

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

- $G(\text{surname}) \approx$ **137000 guesses**
- $G(\text{PIN}) \approx$ **5000 guesses**

- **Meaning:** Expected number of queries "Is $X = x_i$?" for $i = 1, 2, \ldots, N$ (optimal sequential guessing)

Alternate attack models not captured

What if we only want a 50% chance of breaking a given account?

- PIN: $\approx$ **5000 guesses**
- Surname: $\approx$ **8000 guesses**

Alternate attack models not captured

What if we only want a 10% chance of breaking a given account?

- PIN: $\approx$ **1000 guesses**
- Surname: $\approx$ **89 guesses**

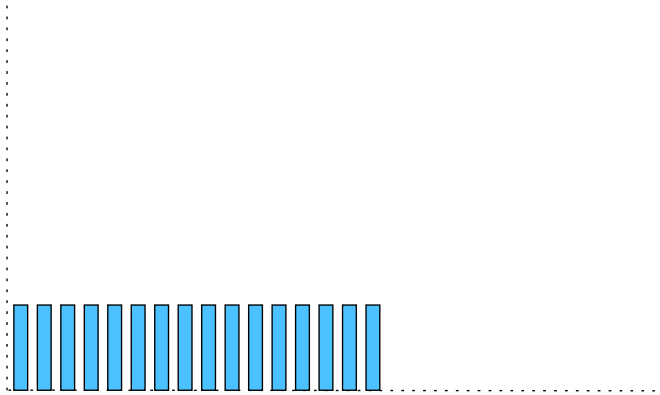
Need specific metrics for attackers who may give up

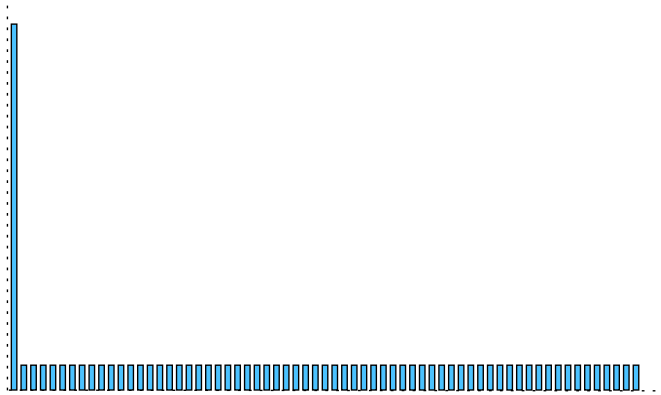
- **Marginal Guesswork**
  Give up after reaching probability α of success:

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

- Can convert to **bitstrength**: $\tilde{\mu}_\alpha(\mathcal{X}) = \lg\left(\frac{\mu_\alpha(\mathcal{X})}{\alpha}\right)$

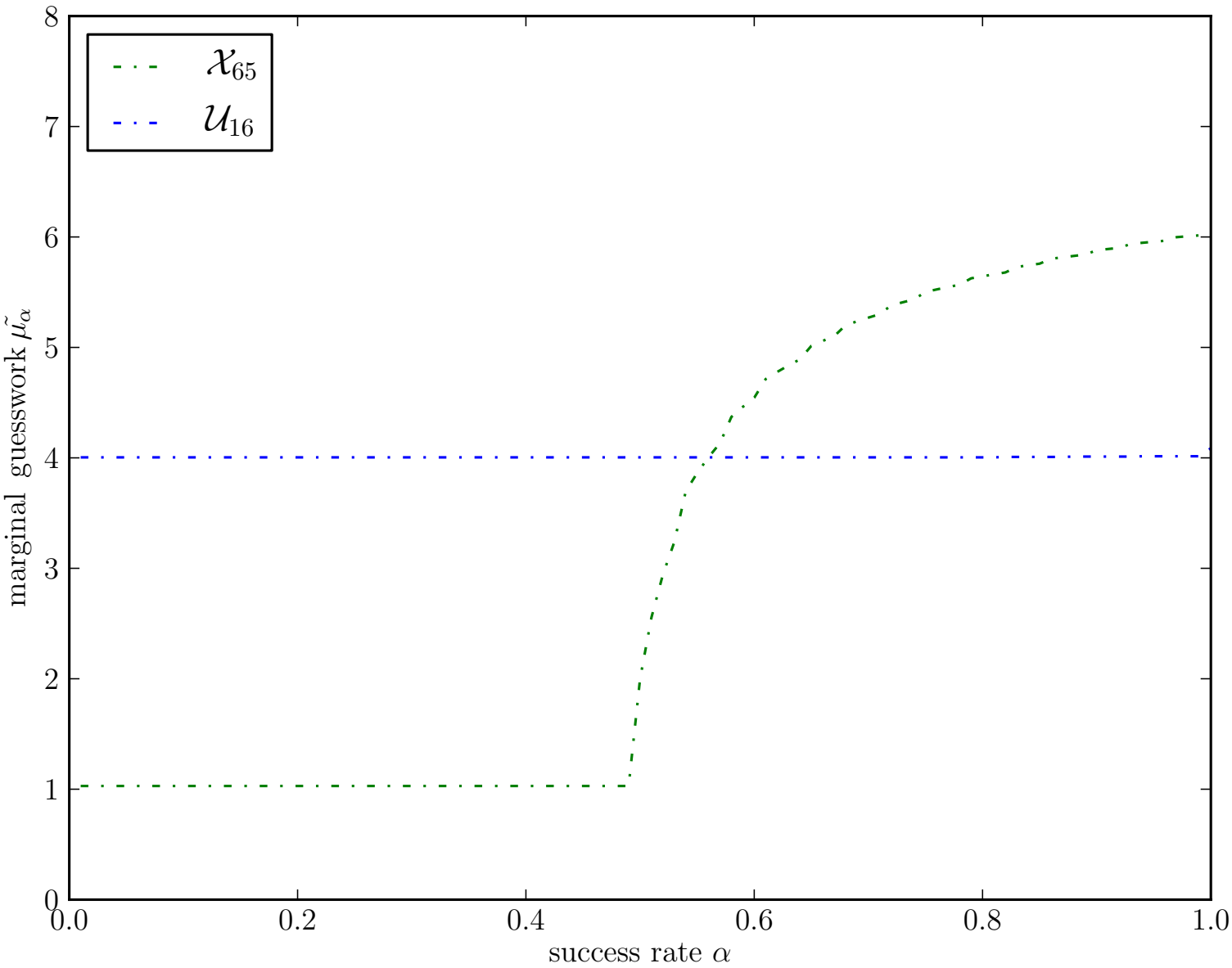
Example



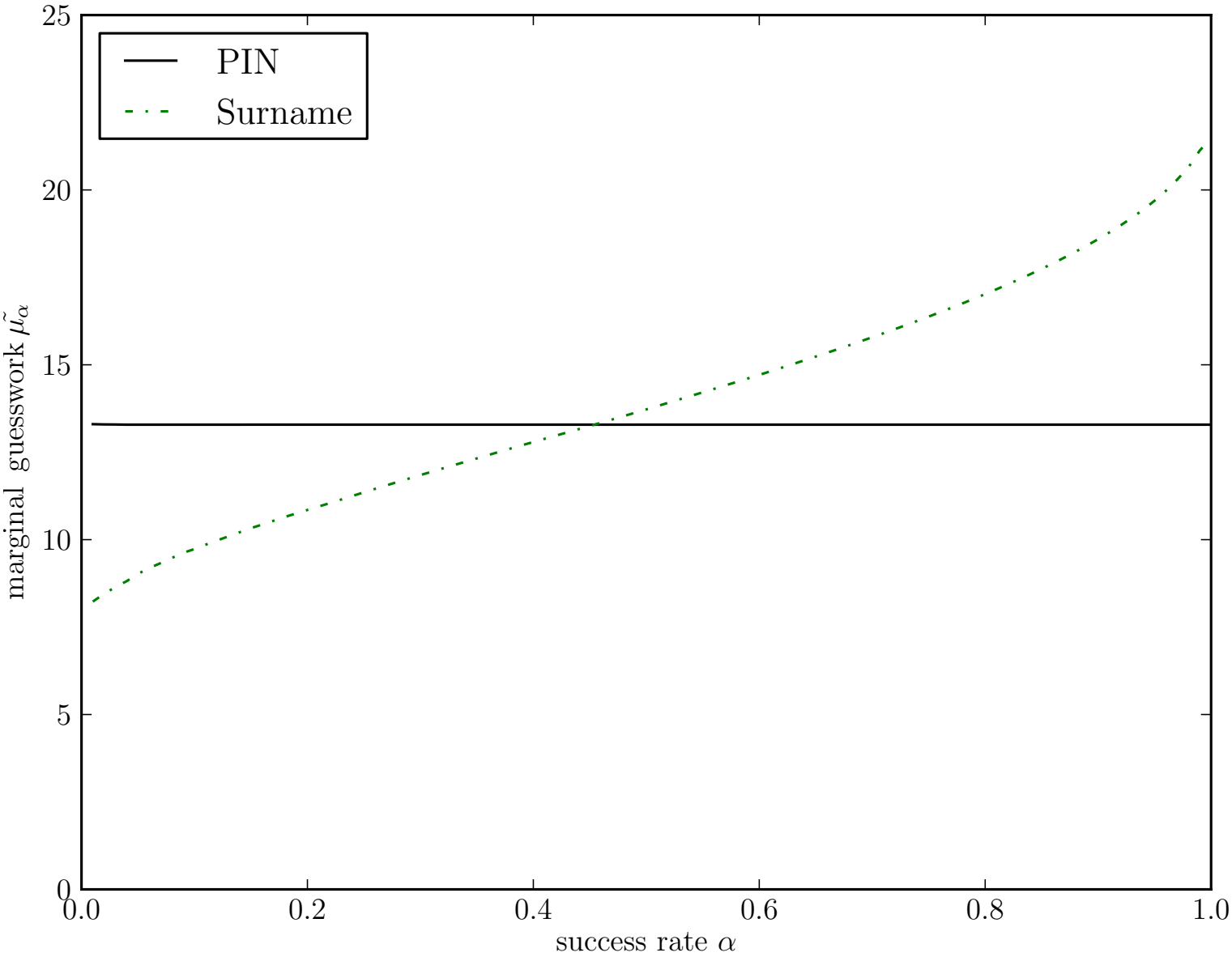


	$\mathcal{U}_{16}$	$\mathcal{X}_{65}$
H_1	4	4
$\tilde{G}$	4	5.1
$\tilde{\mu}_{\frac{1}{2}}$	4	1
$\tilde{\mu}_{\frac{3}{4}}$	4	5.46

The complete picture



The complete picture



Some theorems to wake you up in the morning

Theorem (adapted from Pliam)

Given any $m > 0$, $\beta > 0$ *and* $0 < \alpha < 1$, *there exists a distribution* $\mathcal{X}$ *such that* $\tilde{\mu}_\alpha(\mathcal{X}) < H_1(\mathcal{X}) - m$ *and* $\tilde{\lambda}_\beta(\mathcal{X}) < H_1(\mathcal{X}) - m$.

Theorem (adapted from Boztaş)

Given any $m > 0$, $\beta > 0$ *and* $0 < \alpha < 1$, *there exists a distribution* $\mathcal{X}$ *such that* $\tilde{\mu}_\alpha(\mathcal{X}) < \tilde{G}(\mathcal{X}) - m$ *and* $\tilde{\lambda}_\beta(\mathcal{X}) < \tilde{G}(\mathcal{X}) - m$.

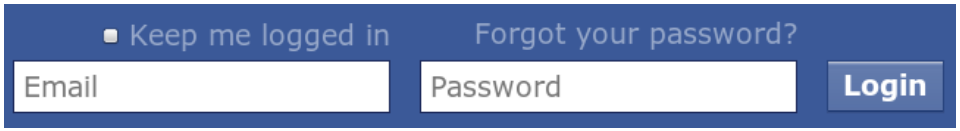
Theorem (from [BJM] FC 2010 paper)

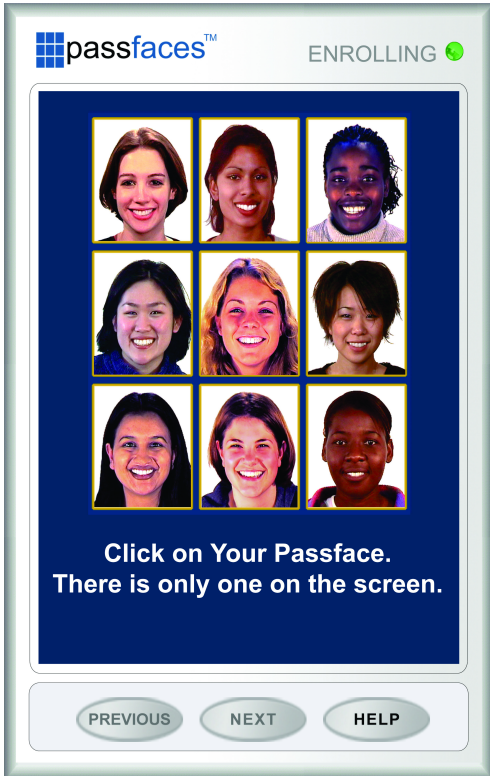
Given any $m > 0, \alpha_1 > 0$, *and* $\alpha_2 > 0$ *with* $0 < \alpha_1 < \alpha_2 < 1$, *there exists a distribution* $\mathcal{X}$ *such that* $\tilde{\mu}_{\alpha_1}(\mathcal{X}) < \tilde{\mu}_{\alpha_1}(\mathcal{X}) - m$.

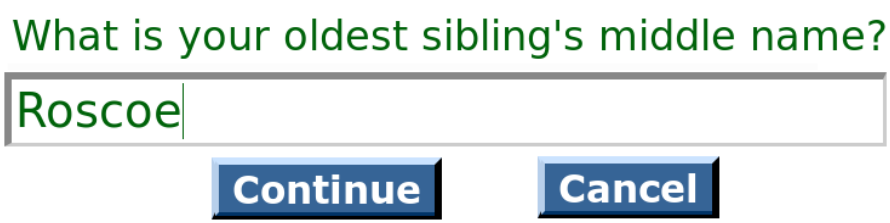
The Simple English guide to human-generated secrets

1. Computers try to tell humans apart by asking for secret data. They can ask for other things, but these are very expensive.
2. Many computer scientists use something called “entropy” to measure security for this secret data, but there are a lot of mathematical equations which say this is a bad idea.
3. Things that good people can remember aren’t unpredictable enough to prevent bad people from guessing them.

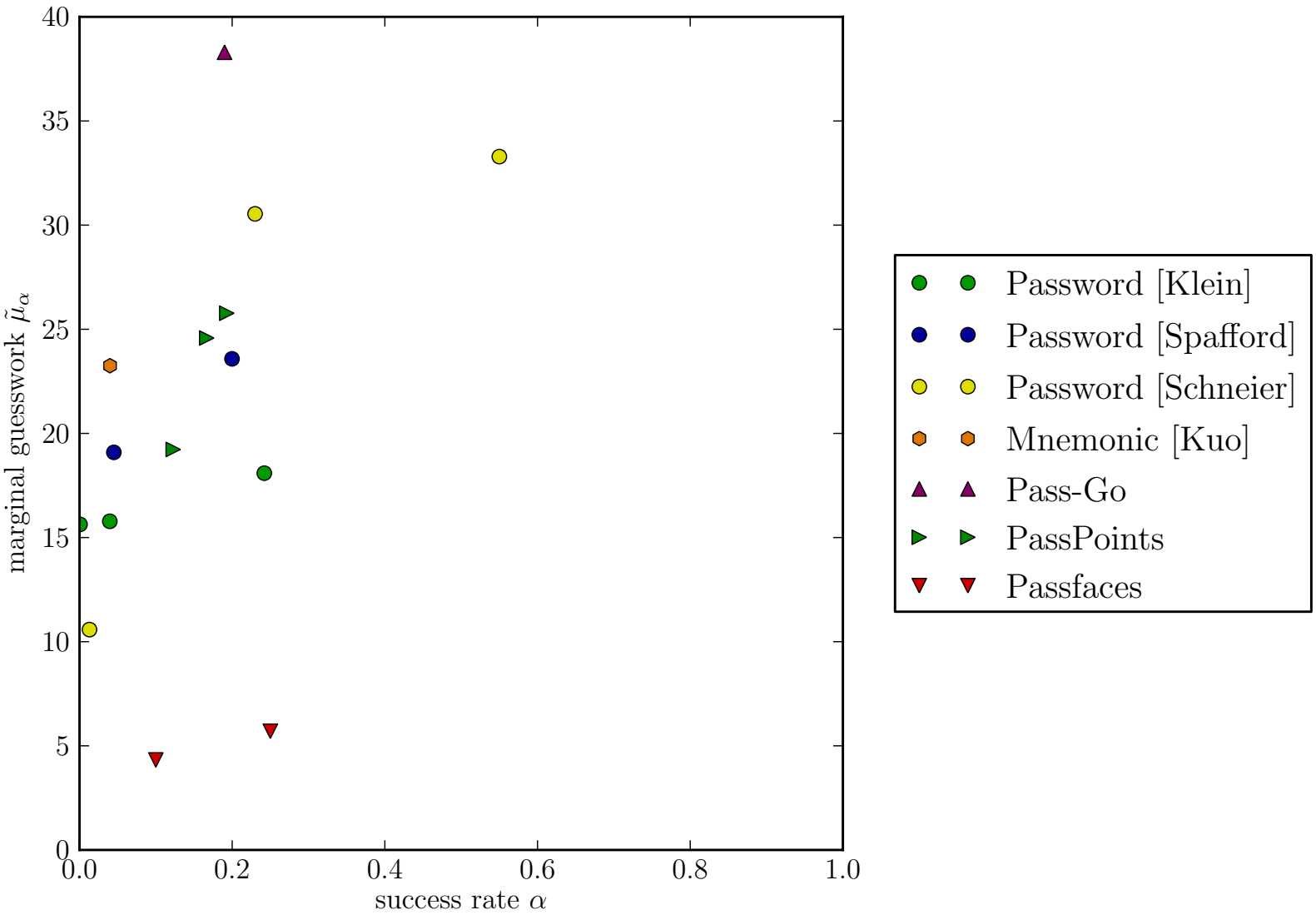
Comparing human-memorable secrets



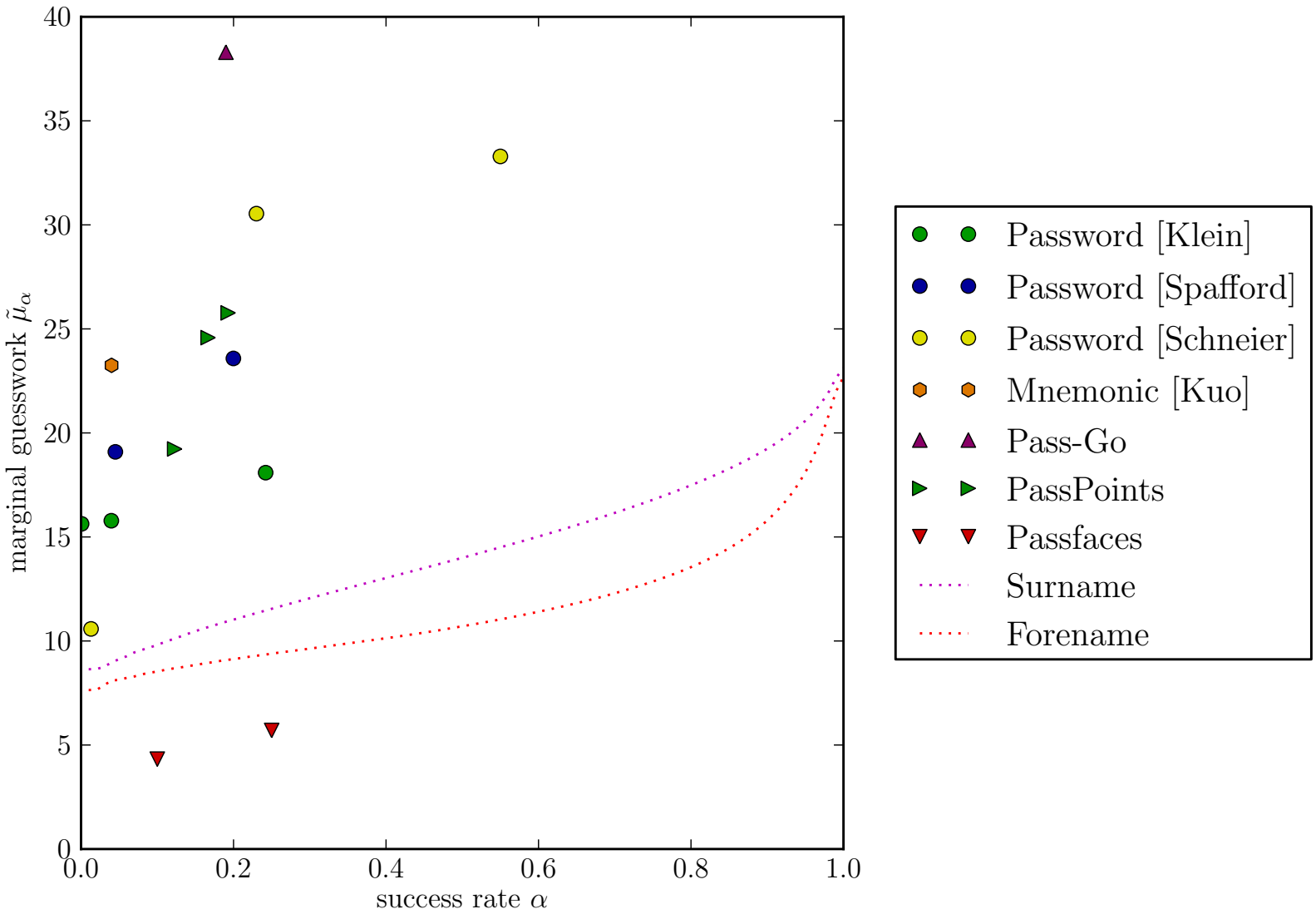




Comparing human-memorable secrets



Comparing human-memorable secrets



The Simple English guide to human-generated secrets

1. Computers try to tell humans apart by asking for secret data. They can ask for other things, but these are very expensive.
2. Many computer scientists use something called “entropy” to measure security for this secret data, but there are a lot of mathematical equations which say this is a bad idea.
3. Things that good people can remember aren’t unpredictable enough to prevent bad people from guessing them.
4. People at a gaming website called RockYou got pwned. Researchers now have many passwords to study.

RockYou loses a list of 32 M passwords

RockYou loses a list of 32 M passwords

290729	`123456`
79076	`12345`
76789	`123456789`
59462	`password`
49952	`iloveyou`
33291	`princess`
21725	`1234567`
20901	`rockyou`
20553	`12345678`
16648	`abc123`
16227	`nicole`
15308	`daniel`
15163	`babygirl`
14726	`monkey`
14331	`lovely`

RockYou loses a list of 32 M passwords

49952	`iloveyou`
13134	`iloveu`
5589	`iloveme`
3998	`iloveyou2`
3700	`iloveyou1`
2042	`iloveu2`
2007	`ilovehim`
1510	`ilovejesus`
1441	`ilovegod`
1358	`iloveyou!`
1096	`iloveu1`
1061	`iloveme1`
922	`ilovemyself`
908	`iloveboys`
894	`ilovechris`

RockYou loses a list of 32 M passwords

830	`lovesucks`
680	`lifesucks`
166	`schoolsucks`
101	`thissucks`
71	`luvsucks`
58	`sucks`
43	`mylifesucks`
33	`aolsucks`
30	`emosucks`
23	`bebosucks`
19	`l0vesucks`
18	`skoolsucks`
16	`love sucks`
16	`worksucks`
15	`lov3sucks`

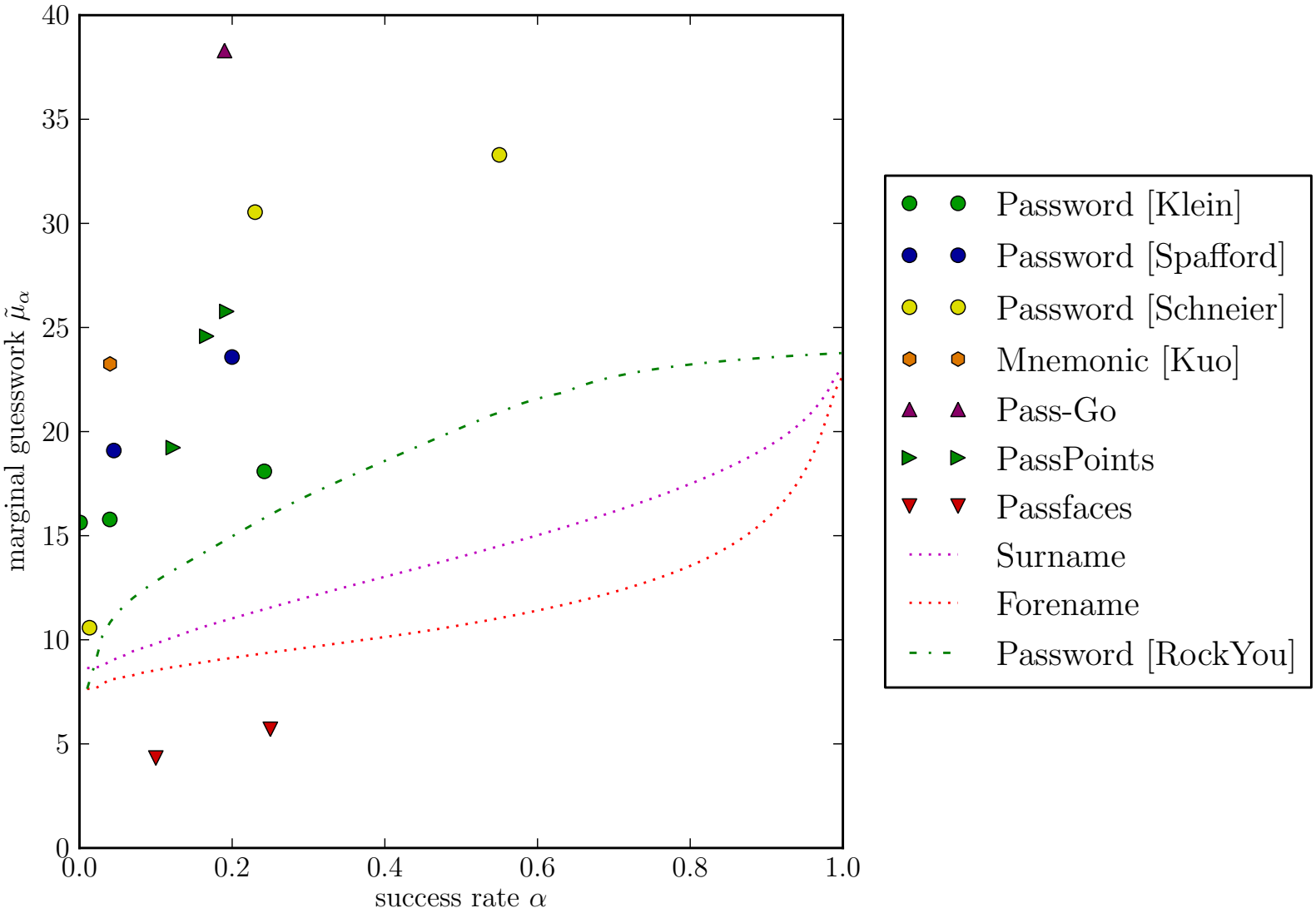
RockYou loses a list of 32 M passwords

28	`joeishot`
11	`joeismine`
10	`joeisfit`
9	`joeissexy`
8	`joeiscool`
6	`joeisgay`
6	`joeishot1`
4	`joeis#1`
3	`joeis1`
3	`joeisa`
3	`joeisastud`
3	`joeiscool1`
3	`joeissexy1`
3	`joeissohot`
3	`joeisthebest`

RockYou loses a list of 32 M passwords

1023	fresita
1023	mookie
1022	leelee
1021	tequieromucho
1020	giovanni
1020	harry
1018	celticfc
1018	ranger
1017	austin1
1017	newcastle
1017	preston
1017	snuggles
1017	tagged
1016	erica
1016	sniper

RockYou loses a list of 32 M passwords



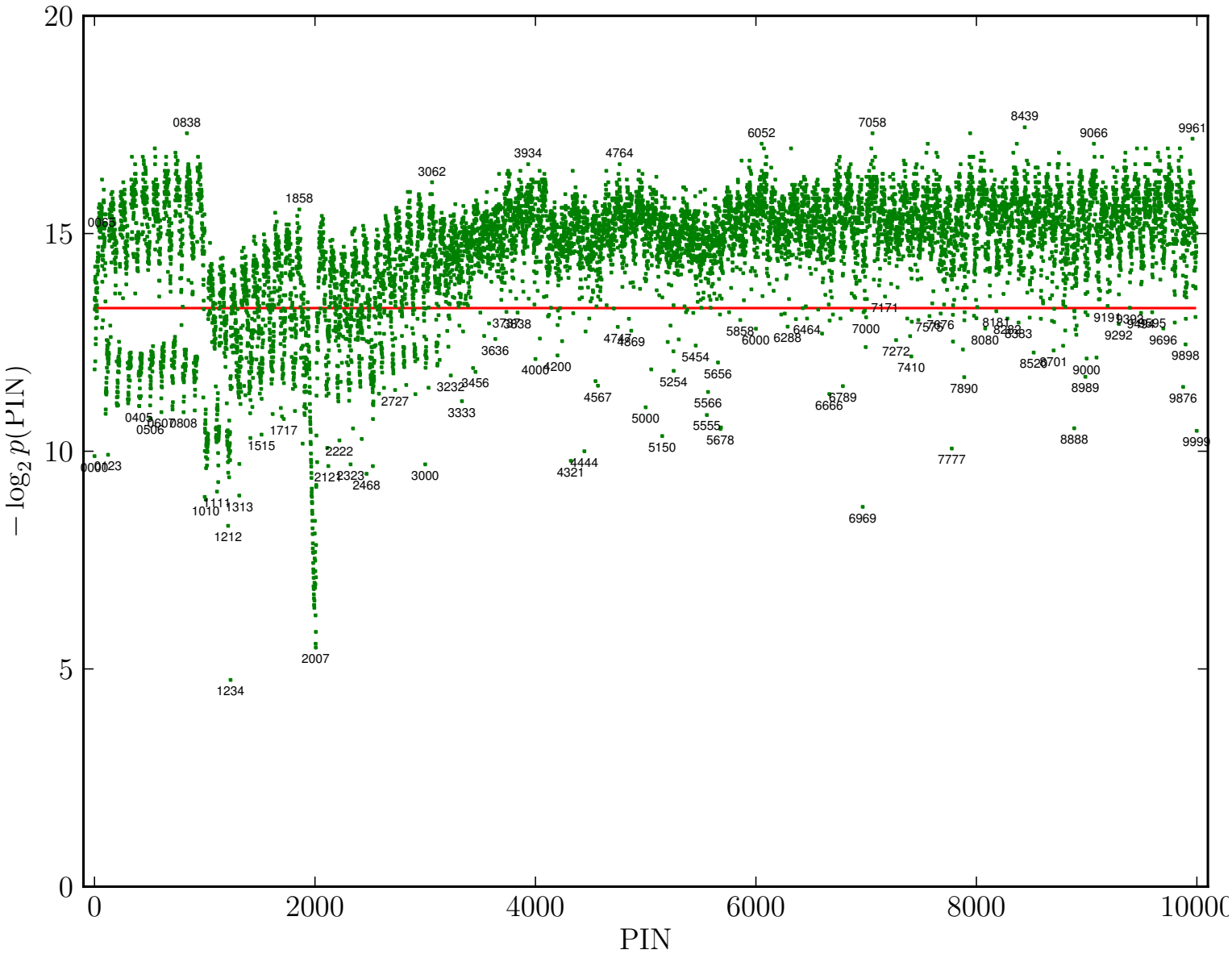
The Simple English guide to human-generated secrets

1. Computers try to tell humans apart by asking for secret data. They can ask for other things, but these are very expensive.
2. Many computer scientists use something called “entropy” to measure security for this secret data, but there are a lot of mathematical equations which say this is a bad idea.
3. Things that good people can remember aren’t unpredictable enough to prevent bad people from guessing them.
4. People at a gaming website called RockYou got pwned. Researchers now have many passwords to study.
5. Computer scientists have never studied how people pick banking PINs, but people are very bad at picking 4-digit numbers for other things, and so they might be bad at picking banking PINs too.

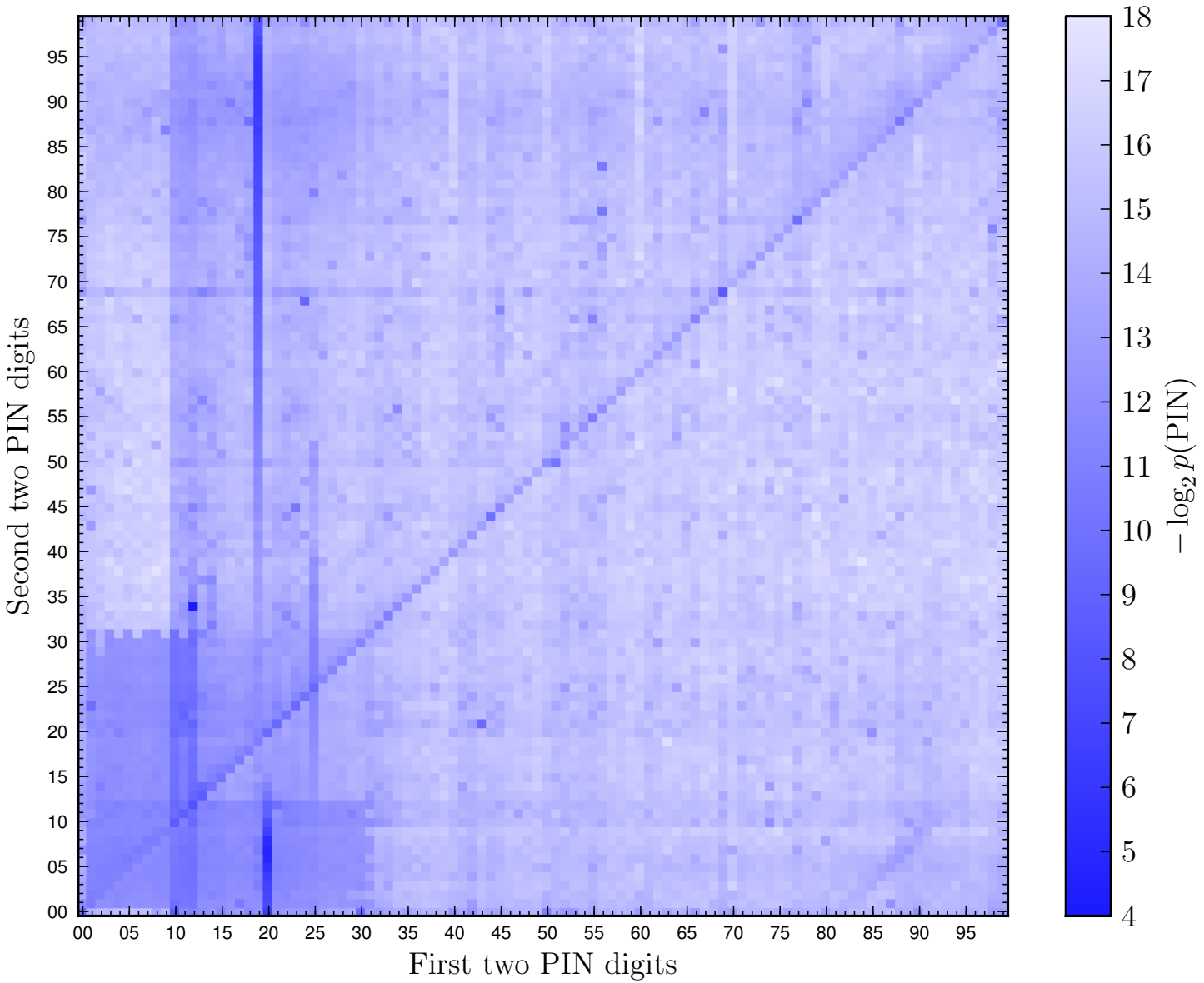
How bad might user-chosen PINs be?

```
grep -E "([^0-9]|^)[0-9]{4}([^0-9]|$)" < rockyou.txt
```

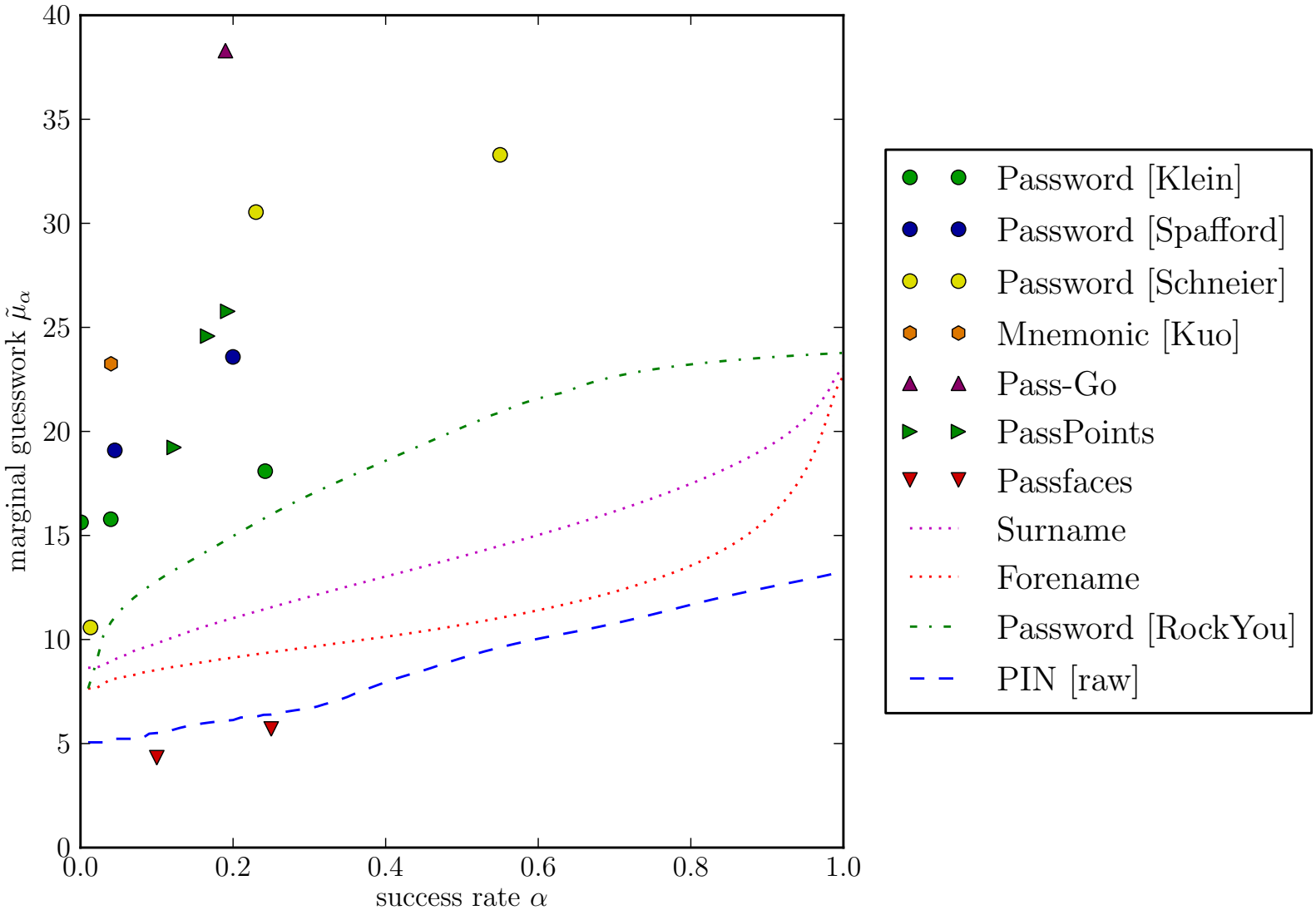
How bad might user-chosen PINs be?



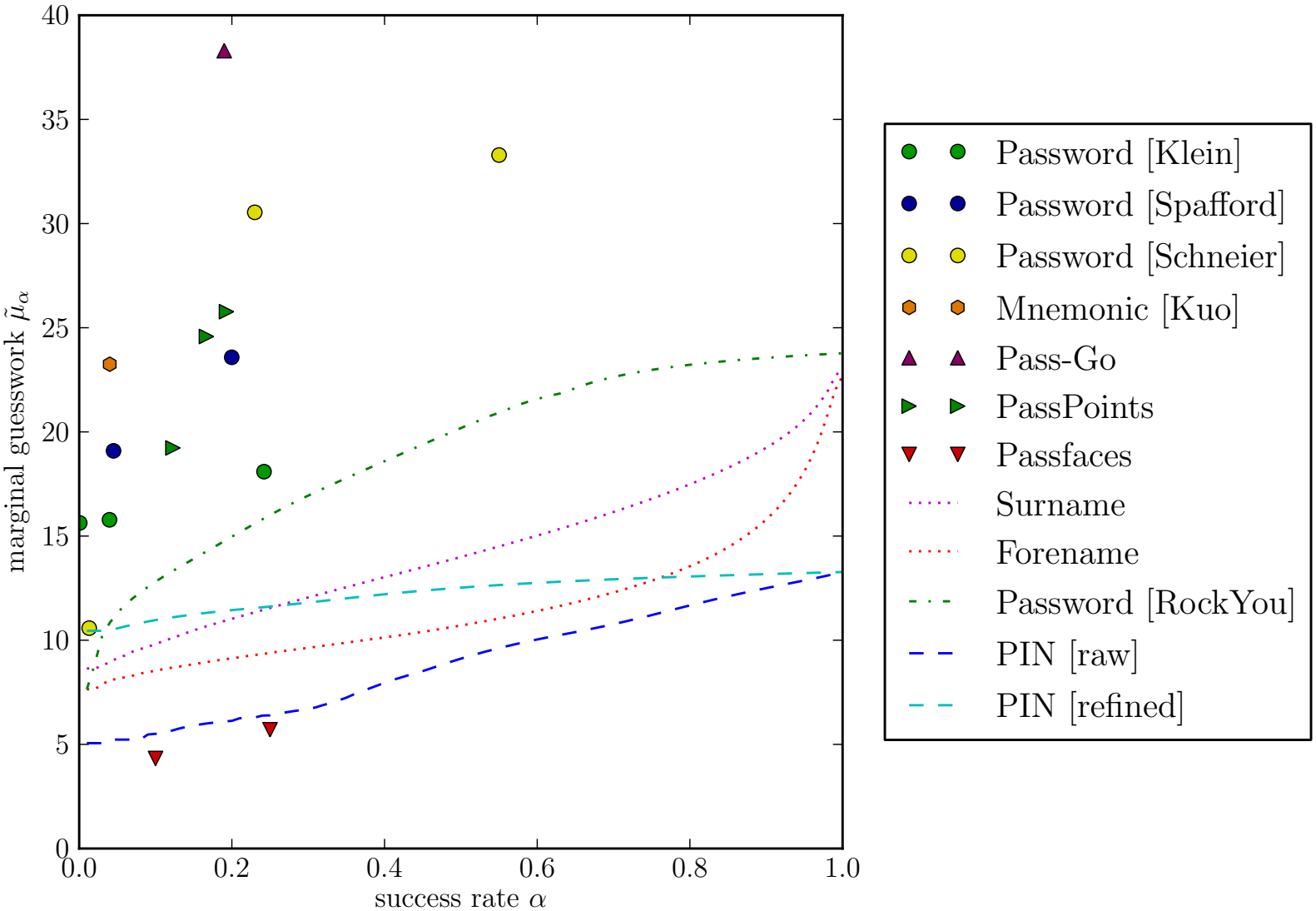
How bad might user-chosen PINs be?



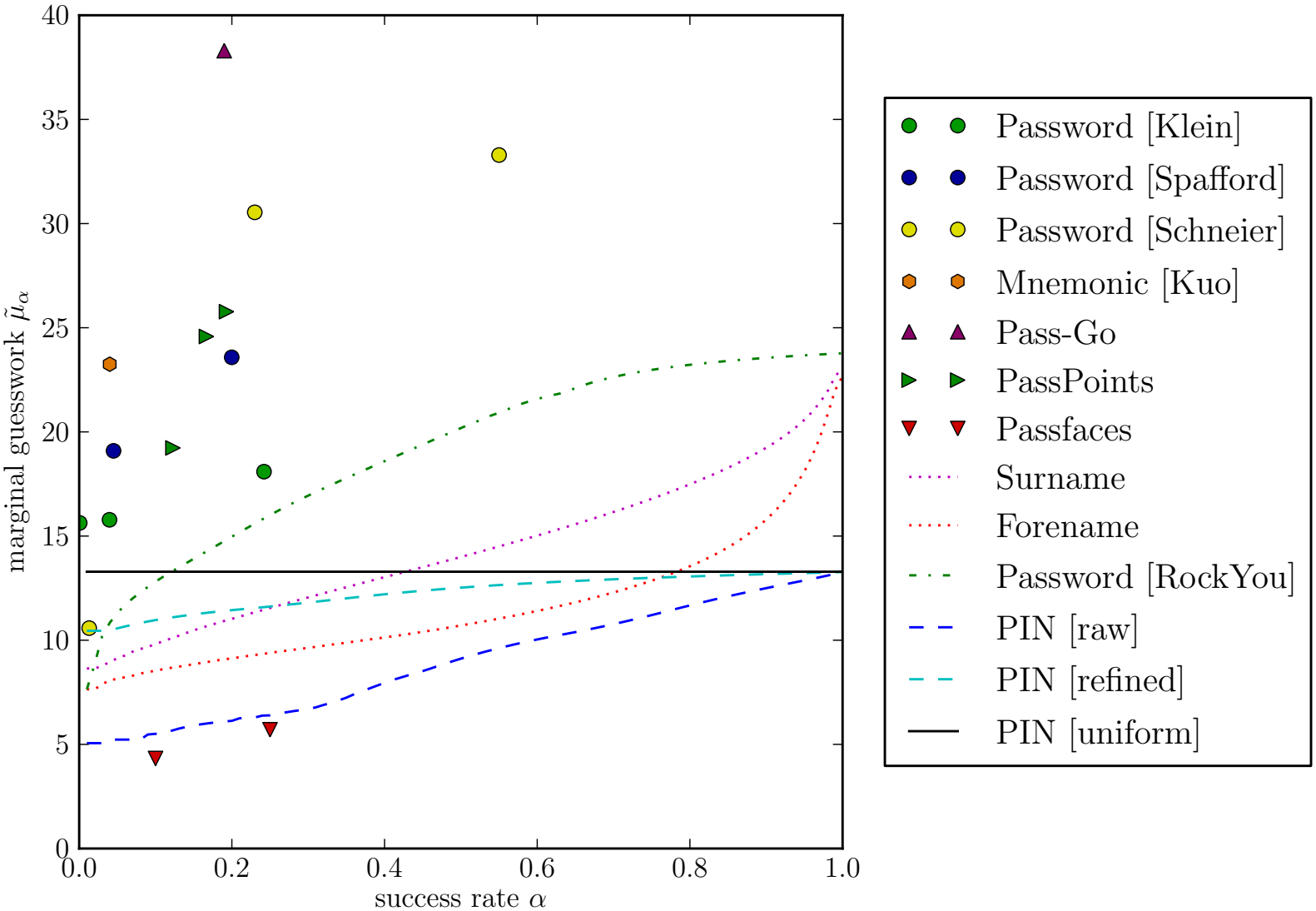
How bad might user-chosen PINs be?



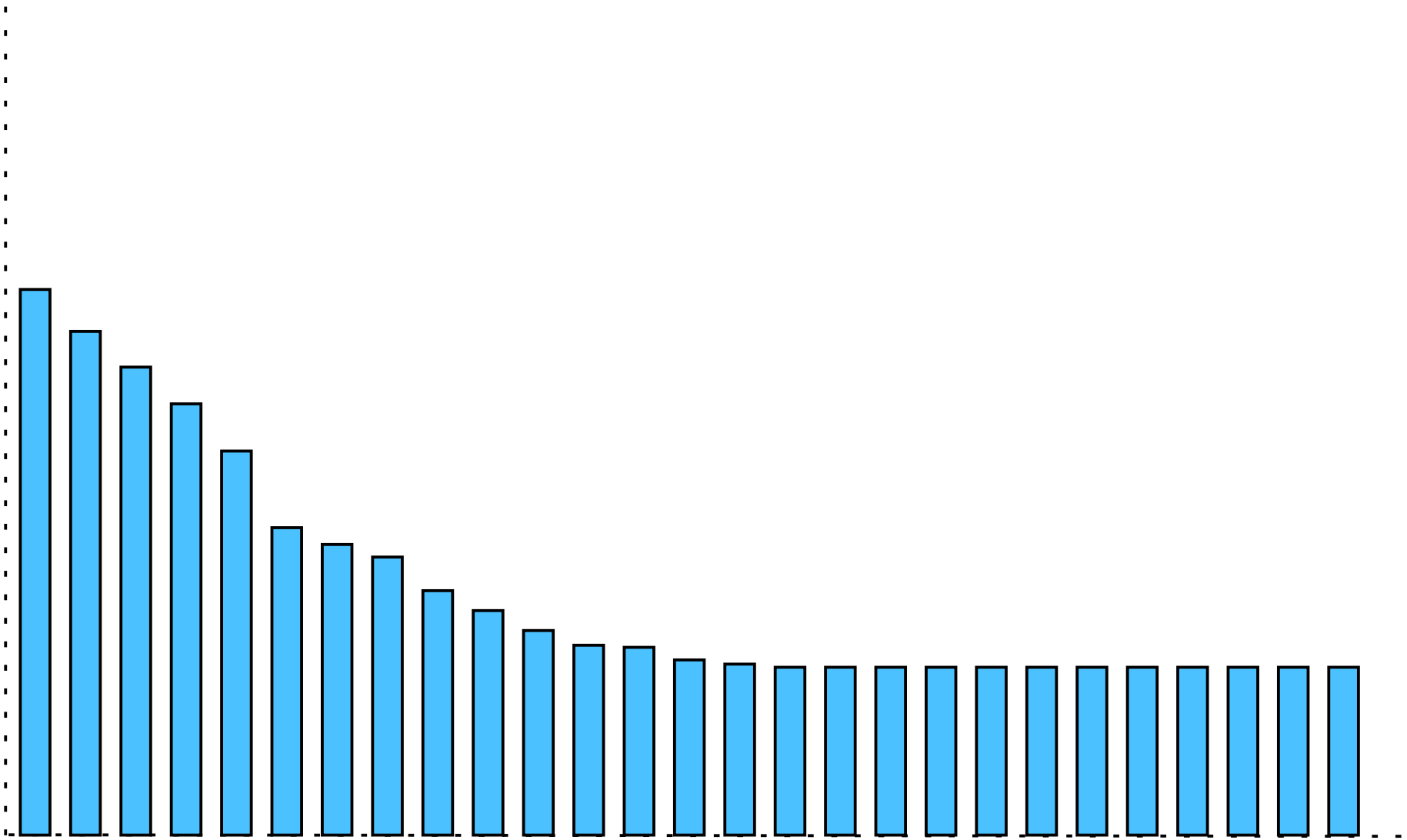
How bad might user-chosen PINs be?



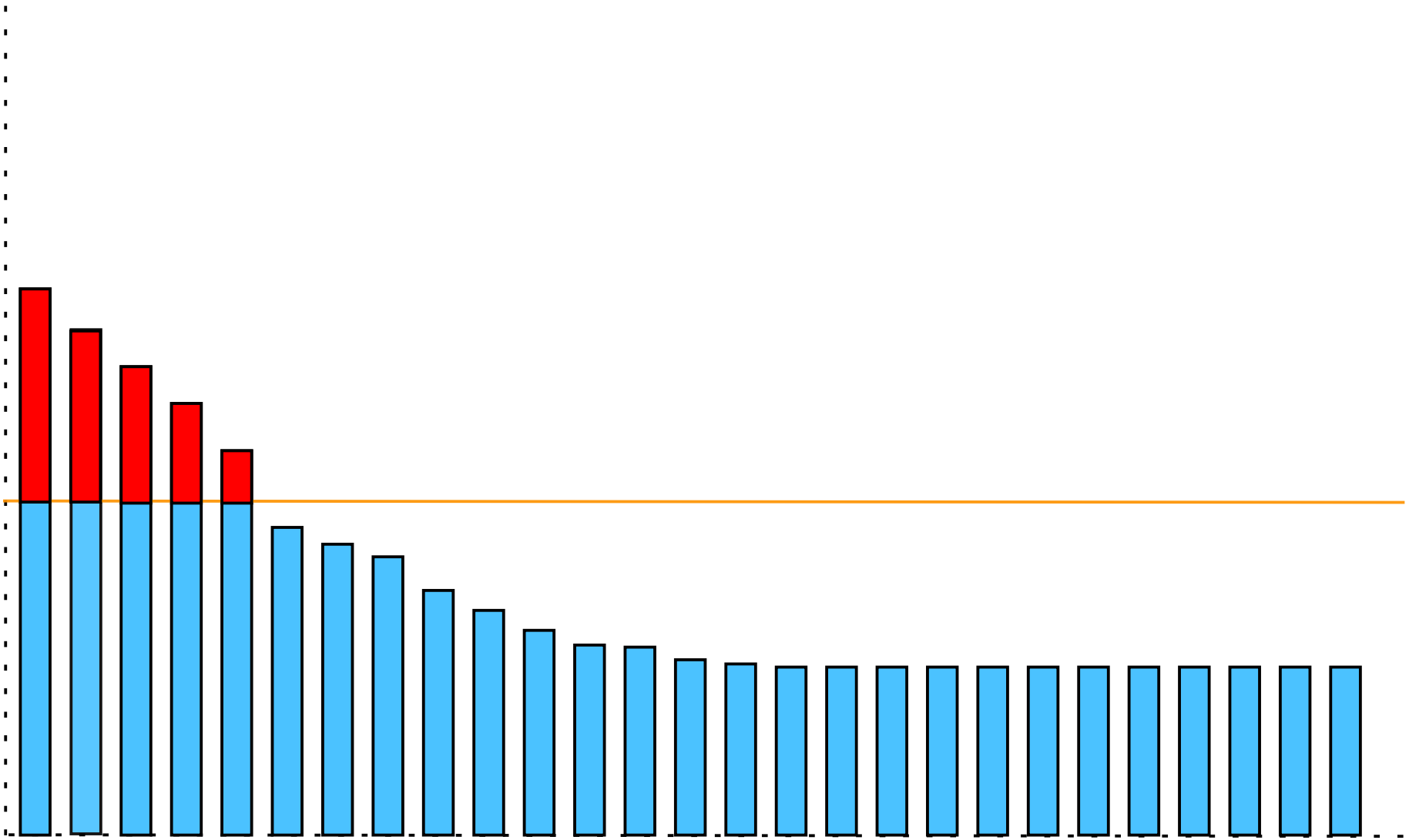
How bad might user-chosen PINs be?



Steering users away from the easiest choices



Steering users away from the easiest choices



Steering users away from the easiest choices