

Applied Information Theory

Daniel Bosk

Department of Information and Communication Systems,
Mid Sweden University, Sundsvall.

14th May 2018

1 Introduction

- History

2 Shannon entropy

- Definition of Shannon Entropy
- Properties for Shannon entropy
- Conditional entropy
- Information density and redundancy
- Information gain

3 Application in security

- Passwords
- Research about human chosen passwords
- Identifying information

1 Introduction

■ History

2 Shannon entropy

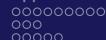
- Definition of Shannon Entropy
- Properties for Shannon entropy
- Conditional entropy
- Information density and redundancy
- Information gain

3 Application in security

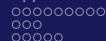
- Passwords
- Research about human chosen passwords
- Identifying information



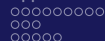
- Created 1948 by Shannon's paper 'A Mathematical Theory of Communication' [Sha48].
- He starts using the term 'entropy' as a measure for information.
 - In physics entropy measures the disorder of molecules.
 - Shannon's entropy measures disorder of information.
- He used this theory to analyse communication.
 - What are the theoretical limits for different channels?
 - How much redundancy is needed for certain noise?



- Created 1948 by Shannon's paper 'A Mathematical Theory of Communication' [Sha48].
- He starts using the term 'entropy' as a measure for information.
 - In physics entropy measures the disorder of molecules.
 - Shannon's entropy measures disorder of information.
- He used this theory to analyse communication.
 - What are the theoretical limits for different channels?
 - How much redundancy is needed for certain noise?



- Created 1948 by Shannon's paper 'A Mathematical Theory of Communication' [Sha48].
- He starts using the term 'entropy' as a measure for information.
 - In physics entropy measures the disorder of molecules.
 - Shannon's entropy measures disorder of information.
- He used this theory to analyse communication.
 - What are the theoretical limits for different channels?
 - How much redundancy is needed for certain noise?



- This theory is interesting on the physical layer of networking.
- It's also interesting for security.
 - Field of Information Theoretic Security
 - 'Efficiency' of passwords
 - Measure identifiability
 - ...



- This theory is interesting on the physical layer of networking.
- It's also interesting for security.
 - Field of Information Theoretic Security
 - 'Efficiency' of passwords
 - Measure identifiability
 - ...

1 Introduction

- History

2 Shannon entropy

- Definition of Shannon Entropy
- Properties for Shannon entropy
- Conditional entropy
- Information density and redundancy
- Information gain

3 Application in security

- Passwords
- Research about human chosen passwords
- Identifying information



Definition (Shannon entropy)

- Stochastic variable \mathbf{X} assumes values from X .
- Shannon entropy $H(\mathbf{X})$ defined as

$$H(\mathbf{X}) = -K \sum_{x \in X} \Pr(\mathbf{X} = x) \log \Pr(\mathbf{X} = x),$$

- Usually $K = \frac{1}{\log 2}$ to give entropy in unit bits (bit).



Shannon entropy can be seen as ...

- ... how much choice in each event.
- ... the uncertainty of each event.
- ... how many bits to store each event.
- ... how much information it produces.

Example (Toss a coin)

- Stochastic variable \mathbf{S} takes values from $S = \{h, t\}$.
- We have $\Pr(\mathbf{S} = h) = \Pr(\mathbf{S} = t) = \frac{1}{2}$.
- This gives $H(\mathbf{S})$ as follows:

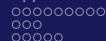
$$\begin{aligned} H(\mathbf{S}) &= -(\Pr(\mathbf{S} = h) \log \Pr(\mathbf{S} = h) + \Pr(\mathbf{S} = t) \log \Pr(\mathbf{S} = t)) \\ &= -2 \times \frac{1}{2} \log \frac{1}{2} = \log 2 = 1. \end{aligned}$$



Example (Roll a die)

- Stochastic variable \mathbf{D} takes values from $D = \{\square, \square\cdot, \square\cdot\cdot, \square\cdot\cdot\cdot, \square\cdot\cdot\cdot\cdot, \square\cdot\cdot\cdot\cdot\cdot\}$.
- We have $\Pr(\mathbf{D} = d) = \frac{1}{6}$ for all $d \in D$.
- The entropy $H(\mathbf{D})$ is as follows:

$$\begin{aligned}
 H(\mathbf{D}) &= - \sum_{d \in D} \Pr(\mathbf{D} = d) \log \Pr(\mathbf{D} = d) \\
 &= -6 \times \frac{1}{6} \log \frac{1}{6} = \log 6 \approx 2.585.
 \end{aligned}$$



Definition of Shannon Entropy

Remark

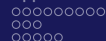
- If we didn't know already, we now know that a roll of a die ...
 - contains more 'choice' than a coin toss.
 - is more uncertain to predict than a coin toss.
 - requires more bits to store than a coin toss.
 - produces more information than a coin toss.
- What if we modify the die a bit?

Example (Roll of a modified die)

- Stochastic variable D' taking values from D .
- We now have $\Pr(D' = \text{Ⓜ}) = \frac{9}{10}$ and $\Pr(D' = d) = \frac{1}{10} \times \frac{1}{5}$ for $d \neq \text{Ⓜ}$.
- This yields

$$\begin{aligned}
 H(D') &= - \left(\frac{9}{10} \log \frac{9}{10} + \sum_{d \neq \text{Ⓜ}} \frac{1}{50} \log \frac{1}{50} \right) \\
 &= - \frac{9}{10} \log \frac{9}{10} - 5 \times \frac{1}{50} \log \frac{1}{50} \\
 &= - \frac{9}{10} \log \frac{9}{10} - \frac{1}{10} \log \frac{1}{50} \approx 0.701.
 \end{aligned}$$

- Note that the log function is the logarithm in base 2 (i.e. \log_2).



Remark

- This die is much easier to predict.
- It produces much less information — less than a coin toss!
- Requires less data for storage etc.

Definition

- Function $f: \mathbb{R} \rightarrow \mathbb{R}$ such that

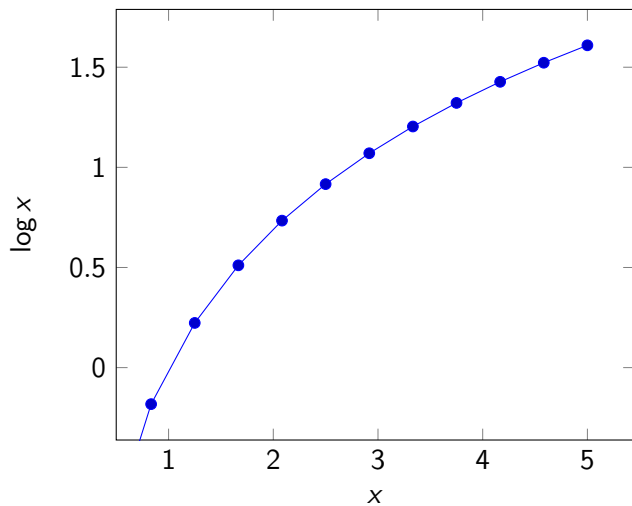
$$tf(x) + (1 - t)f(y) \leq f(tx + (1 - t)y),$$

- Then f is *concave*.
- With strict inequality for $x \neq y$ we say that f is *strictly concave*.

Example

$\log: \mathbb{R} \rightarrow \mathbb{R}$ is strictly concave.

Properties for Shannon entropy



Theorem (Jensen's inequality)

- *Strictly concave function* $f: \mathbb{R} \rightarrow \mathbb{R}$.
- *Real numbers* $a_1, a_2, \dots, a_n > 0$ *such that* $\sum_{i=1}^n a_i = 1$.
- *Then we have*

$$\sum_{i=1}^n a_i f(x_i) \leq f\left(\sum_{i=1}^n a_i x_i\right).$$

- *We have equality iff* $x_1 = x_2 = \dots = x_n$.

Theorem

- *Stochastic variable \mathbf{X} with probability distribution*

$$p_1, p_2, \dots, p_n, \text{ where } p_i > 0 \text{ for } 1 \leq i \leq n.$$

- *Then $H(\mathbf{X}) \leq \log n$.*
- *Equality iff $p_1 = p_2 = \dots = p_n = 1/n$.*

Proof.

The theorem follows directly from Jensen's inequality:

$$\begin{aligned} H(\mathbf{X}) &= -\sum_{i=1}^n p_i \log p_i = \sum_{i=1}^n p_i \log \frac{1}{p_i} \\ &\leq \log \sum_{i=1}^n p_i \frac{1}{p_i} = \log n. \end{aligned}$$

With equality iff $p_1 = p_2 = \dots = p_n$.

Q.E.D.

Corollary

$H(\mathbf{X}) = 0$ iff $\Pr(\mathbf{X} = x) = 1$ for some $x \in X$ and $\Pr(\mathbf{X} = x') = 0$ for all $x \neq x' \in X$.

Proof.

- If $\Pr(\mathbf{X} = x) = 1$, then $n = 1$ and thus $H(\mathbf{X}) = \log n = 0$.
- If $H(\mathbf{X}) = 0$, then $H(\mathbf{X}) \leq \log n = 0$. Thus $n = 1$.

Q.E.D.

Lemma

- *Stochastic variables \mathbf{X} and \mathbf{Y} .*
- *Then we have*

$$H(\mathbf{X}, \mathbf{Y}) \leq H(\mathbf{X}) + H(\mathbf{Y}).$$

- *Equality iff \mathbf{X} and \mathbf{Y} are independent.*

Definition (Conditional entropy)

- Define *conditional entropy* $H(\mathbf{Y} \mid \mathbf{X})$ as

$$H(\mathbf{Y} \mid \mathbf{X}) = - \sum_y \sum_x \Pr(\mathbf{Y} = y) \Pr(\mathbf{X} = x \mid y) \log \Pr(\mathbf{X} = x \mid y).$$

Remark

This is the uncertainty in \mathbf{Y} which is not revealed by \mathbf{X} .

Definition (Conditional entropy)

- Define *conditional entropy* $H(\mathbf{Y} \mid \mathbf{X})$ as

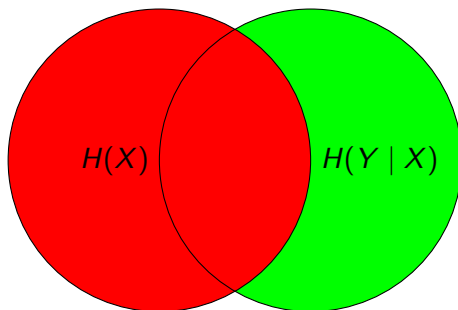
$$H(\mathbf{Y} \mid \mathbf{X}) = - \sum_y \sum_x \Pr(\mathbf{Y} = y) \Pr(\mathbf{X} = x \mid y) \log \Pr(\mathbf{X} = x \mid y).$$

Remark

This is the uncertainty in \mathbf{Y} which is not revealed by \mathbf{X} .

Theorem

$$H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y} | \mathbf{X})$$



Corollary

$$H(\mathbf{X} \mid \mathbf{Y}) \leq H(\mathbf{X}).$$

Corollary

$$H(\mathbf{X} \mid \mathbf{Y}) = H(\mathbf{X}) \text{ iff } \mathbf{X} \text{ and } \mathbf{Y} \text{ independent.}$$

Definition

- Natural language L .
- Stochastic variable \mathbf{P}_L^n of strings of length n .
- (Alphabet P_L .)
- Entropy of L defined as

$$H_L = \lim_{n \rightarrow \infty} \frac{H(\mathbf{P}_L^n)}{n}.$$

- Redundancy in L is

$$R_L = 1 - \frac{H_L}{\log |P_L|}.$$

Remark

Meaning we have H_L bits per character in L .

Example ([Sha48])

- Entropy of 1–1.5 bits per character in English.
- Redundancy of approximately $1 - \frac{1.25}{\log 26} \approx 0.73$.

Example ([Sha48])

Two-dimensional cross-word puzzles requires redundancy of approximately 0.5.

Example

- Redundancy of 'SMS languages' is lower than for 'non-SMS languages'.
- Compare 'också' and 'oxå'.

Remark

- Lower redundancy is more space-efficient.
- Incurs more errors.

Definition

- Set U of possible outcomes.
- Probability of outcome $u \in U$ denoted p_u .
- We learn that some *unknown* outcome is in $A \subset U$.
- Then the *information gain* $G(A | U)$ is defined as

$$G(A | U) = \log \frac{1}{\Pr(A)} = -\log \Pr(A),$$

where $\Pr(A) = \sum_{i \in A} p_i$.

Example (Roll of dice again)

- Someone rolls and we should guess the result, $\frac{1}{6}$ chance.
- We learn that it was an even number, we gain

$$-\log\left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) = -\log\frac{3}{6} = \log\frac{6}{3} = \log 2 = 1.$$

- The remaining uncertainty is 1.58 bit.

Remark

- $X' = \{\square, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square \\ \square & \square & \square \end{smallmatrix}\}$
- $H(X') = -\sum_{x \in X'} \Pr(X' = x) \log \Pr(X' = x)$
- I.e. $-3 \times \frac{1}{3} \log \frac{1}{3} \approx 1.58$.

Example (Roll of dice again)

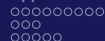
- Someone rolls and we should guess the result, $\frac{1}{6}$ chance.
- We learn that it was an even number, we gain

$$-\log\left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) = -\log\frac{3}{6} = \log\frac{6}{3} = \log 2 = 1.$$

- The remaining uncertainty is 1.58 bit.

Remark

- $X' = \{\square, \blacksquare, \blacksquare\}$
- $H(\mathbf{X}') = -\sum_{x \in X'} \Pr(\mathbf{X}' = x) \log \Pr(\mathbf{X}' = x)$
- i.e. $-3 \times \frac{1}{3} \log \frac{1}{3} \approx 1.58$.



Example (Dice yet again)

- We learn the die show less than five, i.e. not 🎲 nor 🎲.
- This yields

$$-\log\left(4 \times \frac{1}{6}\right) = \log\frac{6}{4} \approx 0.58$$

1 Introduction

- History

2 Shannon entropy

- Definition of Shannon Entropy
- Properties for Shannon entropy
- Conditional entropy
- Information density and redundancy
- Information gain

3 Application in security

- Passwords
- Research about human chosen passwords
- Identifying information

Idea [Kom+11]

- Look at different aspects of passwords individually, then summarize.
- Can use $H(x_1, x_2, \dots, x_n) \leq H(x_1) + H(x_2) + \dots + H(x_n)$.
- This allows us to reason about bounds.

Example

- We can look at properties such as:
 - length,
 - number of and placement of character classes,
 - the actual characters,
 - ...

Remark

- These are *not independent*.
- The sum will be an *upper bound*.

Example

- We can look at properties such as:
 - length,
 - number of and placement of character classes,
 - the actual characters,
 - ...

Remark

- These are *not independent*.
- The sum will be an *upper bound*.

Remark

- With an upper bound we know it's not possible to do better.
- With an average we know how well most users will do.
- With a lower bound we have a guarantee — not possible!

Remark

- If a password policy yields low entropy, it implies it's bad.
- If a password policy yields high entropy, it *doesn't* imply that it's good.

Exercise

Why?

Remark

- If a password policy yields low entropy, it implies it's bad.
- If a password policy yields high entropy, it *doesn't* imply that it's good.

Exercise

Why?

Passwords

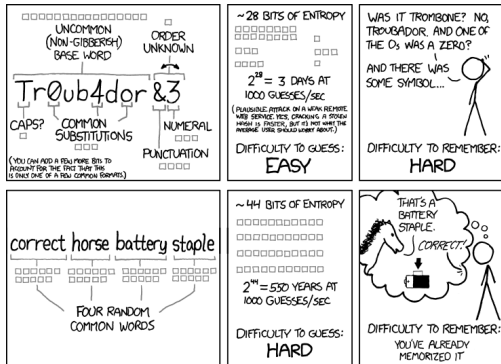


Figure: xkcd:s serie om lösenordsstyrka. Bild: xkcd [xkc].

Example (Standard password)

- We have
 - 26 alphabetic characters,
 - 10 numbers,
 - 10 special characters (approximately).
- This yields $\log(2 \times 26 + 10 + 10) = \log 72 \approx 6$ bit per password character.
- A 10-character *uniformly randomly* generated password contains 60 bit.

Remark

What happens when we require two upper and two lower-case characters, two numbers must be included?

Example (Standard password)

- We have
 - 26 alphabetic characters,
 - 10 numbers,
 - 10 special characters (approximately).
- This yields $\log(2 \times 26 + 10 + 10) = \log 72 \approx 6$ bit per password character.
- A 10-character *uniformly randomly* generated password contains 60 bit.

Remark

What happens when we require two upper and two lower-case characters, two numbers must be included?

Example (Four-word passphrase)

- We have 125 000 words in the standard Swedish dictionary.
- This yields $\log 125\,000 \approx 17$ bit per word.
- A four-word *uniformly randomly* generated passphrase contains 68 bit.

Example (Random sentence)

- We estimated the entropy per character in a language.
- It was approximately 1.25 bit for English.
- A 20-character *uniformly randomly* generated sentence would yield 25 bit.

Remark

- All these require uniform randomness.
- Humans are bad at remembering random things.
- Thus they will choose non-randomly.
- The entropy will thus be (possibly much) lower.

Example ('Linguistic properties of multi-word passwords' [BS12])

- Investigates how linguistics affect the choice of multi-word passphrases.
- Users don't choose them randomly, prefer adapted to natural language.
- 'correct horse battery staple' is preferred to 'horse correct battery staple' since the first is more grammatically correct.

Example (*Human Selection of Mnemonic Phrase-based Passwords* [KRC06])

- Studied how users creates easy-to-remember passwords.
- Also investigated the strength of phrase-based passwords.
- E.g. Google's example 'To be or not to be, that is the question'¹ which results in '2bon2btitq'.
- This particular password has apparently been used by many ...

¹URL: <http://www.lightbluetouchpaper.org/2011/11/08/want-to-create-a-really-strong-password-dont-ask-google/>

Remark

- There is a PhD thesis on the topic of guessing passwords: Joseph Bonneau. *Guessing human-chosen secrets*. Tech. rep. UCAM-CL-TR-819. University of Cambridge, Computer Laboratory, May 2012. URL: <http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-819.pdf>.
- There is even a conference dedicated to passwords: PasswordsCon.

Example

Do we get more information from zodiac signs or birthdays?

$$-\sum_{\text{zodiacs}} \frac{1}{12} \log \frac{1}{12} = \log 12 \approx 3.58$$

$$< -\sum_{\text{days of year}} \frac{1}{365} \log \frac{1}{365} = \log 365 \approx 8.51.$$

Exercise

How much information do we need to uniquely identify an individual?

Example

- Sometime during 2011 there were $n = 6\,973\,738\,433^2$ people on earth.
- To give everyone a unique identifier we need $\log n \approx 32.7 \approx 33$ bits of information.

²According to the World Bank.

Identifying information in browsers

- Electronic Frontier Foundation (EFF) studied [Eck10] how much information a web-browser shares.
- You can try your browser in <http://panopticlick.eff.org/>.

Example (My browser)

- My Firefox-browser with all addons gave 21.45 bits of entropy.
- Then the number of tested users were 2 860 696.

Identifying information in browsers

- Electronic Frontier Foundation (EFF) studied [Eck10] how much information a web-browser shares.
- You can try your browser in <http://panopticlick.eff.org/>.

Example (My browser)

- My Firefox-browser with all addons gave 21.45 bits of entropy.
- Then the number of tested users were 2 860 696.

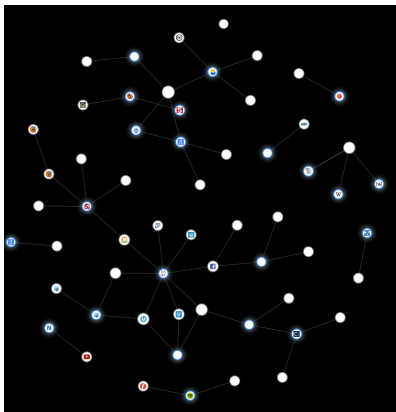


Figure: Screenshot from Collusion (now Lightbeam) for Firefox. Map over all pages that track me using this information.

- [Bon12] Joseph Bonneau. *Guessing human-chosen secrets*. Tech. rep. UCAM-CL-TR-819. University of Cambridge, Computer Laboratory, May 2012. URL: <http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-819.pdf>.
- [BS12] Joseph Bonneau and Ekaterina Shutova. ‘Linguistic properties of multi-word passwords’. In: *USEC. 2012*. URL: http://www.cl.cam.ac.uk/~jcb82/doc/BS12-USEC-passphrase_linguistics.pdf.

- [Eck10] Peter Eckersley. ‘How Unique Is Your Browser?’ In: *Privacy Enhancing Technologies*. Springer. 2010, pp. 1–18. URL: <https://panopticklick.eff.org/static/browser-uniqueness.pdf>.
- [Kom+11] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L. Mazurek, Lujo Bauer, Christin Nicolas, Lorrie Faith Cranor and Serge Egelman. ‘Of passwords and people: Measuring the effect of password-composition policies’. In: *CHI*. 2011. URL: http://cups.cs.cmu.edu/rshay/pubs/passwords_and_people2011.pdf.

- [KRC06] Cynthia Kuo, Sasha Romanosky and Lorrie Faith Cranor. *Human Selection of Mnemonic Phrase-based Passwords*. Tech. rep. 36. Institute of Software Research, 2006. URL: <http://repository.cmu.edu/isr/36/>.
- [Sha48] C. E. Shannon. 'A Mathematical Theory of Communication'. In: *The Bell System Technical Journal* 27 (July 1948), pp. 379–423, 623–656.
- [xkc] xkcd. *Password Strength*. URL: <https://xkcd.com/936/>.