Data Protection

Jorge L. Villar

MCYBERS, UPC, Fall 2025

Introduction

Outline

Contents

Approximate timing:

- Introduction (1 hour)
- Symmetric Key Crypto (7 hours)
- Public Key Crypto (8 hours)
- Security Models (6 hours)
- Zero Knowledge (3 hours)
- Distributed Protocols (3 hours)

- Lab (11 sessions)

Organization

Classes (large group)

Face 2 face, two hours sessions with hourly break.

Lab sessions (small groups)

Lab work in groups of two students.

No final report or oral presentations this course!

Grading

Grade splitting:

- Lab Reports (60%)
- Final Exam (40%) January 16, 2026

More information at:

- `https://mat-web.upc.edu/people/jorge.villar/course_dataprot.html`
- `https://atenea.upc.edu/course/view.php?id=100913`

Outline

Introduction

The main goal:

Efficient and Reliable Data Transmission and Storage

Introduction

The main goal:
Efficient and Reliable Data Transmission and Storage

Data: digital data source = (infinite) sequence of binary or q-ary symbols

Introduction

The main goal:
Efficient and Reliable Data Transmission and Storage

Data: digital data source = (infinite) sequence of binary or q-ary symbols

Reliable: against

- imperfect environment $\longrightarrow$ **Information Coding**
- malicious users $\longrightarrow$ **Cryptography**

Introduction

The main goal:
Efficient and Reliable Data Transmission and Storage
e.g.: Binary Data Storage

Data: digital data source = (infinite) sequence of binary or q-ary symbols

Reliable: against

- imperfect environment $\longrightarrow$ **Information Coding**
- malicious users $\longrightarrow$ **Cryptography**

Efficient: w.r.t.

- read / write / modify / delete time (per bit / per access)
- physical space
- life of stored data

Example: Binary Data Storage (I)

Reliable against imperfect environment $\longrightarrow$ **Information Coding**

Add redundancy (w/o degrading efficiency too much)

- Use error correction codes: detect/correct a few errors
- Use data shuffling: protects against correlated error locations (Compact Disk)

Example: Binary Data Storage (I)

Reliable against imperfect environment $\longrightarrow$ **Information Coding**

Add redundancy (w/o degrading efficiency too much)

- Use error correction codes: detect/correct a few errors
- Use data shuffling: protects against correlated error locations (Compact Disk)

Efficient w.r.t. physical space $\longrightarrow$ **Information Coding**

Use a compression code (w/o degrading speed too much)

Example: Binary Data Storage (II)

Reliable against malicious users $\longrightarrow$ **Cryptography**

Confidentiality: Use symmetric encryption

Integrity: Use message authentication codes

Availability: Use secret sharing schemes

Outline

The Setting (I)

Perfect environment: No storage or communication errors or excessive message delivery delays.

The Setting (I)

Perfect environment: No storage or communication errors or excessive message delivery delays.

Users: divided into

- good guys (honest)
- bad guys (corrupted by an adversary)

The Setting (I)

Perfect environment: No storage or communication errors or excessive message delivery delays.

Users: divided into

- good guys (honest)
- bad guys (corrupted by an adversary)

Alternative model: Rational Cryptography (from Game Theory). Only selfish guys (not necessarily honest, can collude).

The Setting (I)

Perfect environment: No storage or communication errors or excessive message delivery delays.

Users: divided into

- good guys (honest)
- bad guys (corrupted by an adversary)

Alternative model: Rational Cryptography (from Game Theory). Only selfish guys (not necessarily honest, can collude).

Simplest case: One honest user, one bad user.
E.g.: Secure binary data storage.

The Setting (II)

Adversarial behavior:

The Setting (II)

Adversarial behavior:

- **static:** corrupted users are fixed before starting the actual attack
- **dynamic:** corrupted users are decided on-the-fly during the attack

The Setting (II)

Adversarial behavior:

- **static:** corrupted users are fixed before starting the actual attack
- **dynamic:** corrupted users are decided on-the-fly during the attack

- **passive:** corrupted users follow the protocol and try to learn more than they are allowed to
- **active:** corrupted users deviate from the protocol in any arbitrary way

The Setting (II)

Adversarial behavior:

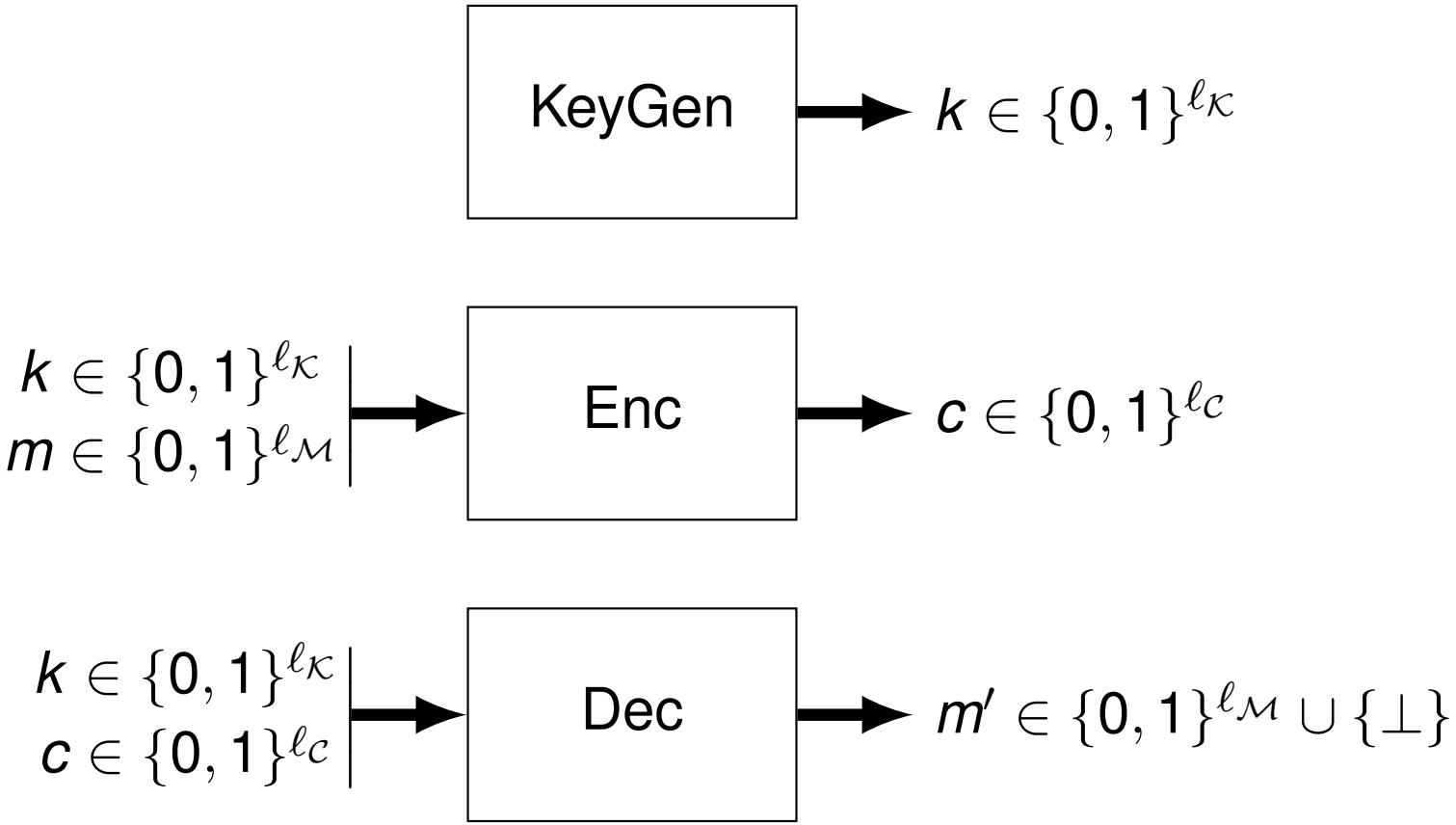
- **static:** corrupted users are fixed before starting the actual attack
- **dynamic:** corrupted users are decided on-the-fly during the attack

- **passive:** corrupted users follow the protocol and try to learn more than they are allowed to
- **active:** corrupted users deviate from the protocol in any arbitrary way

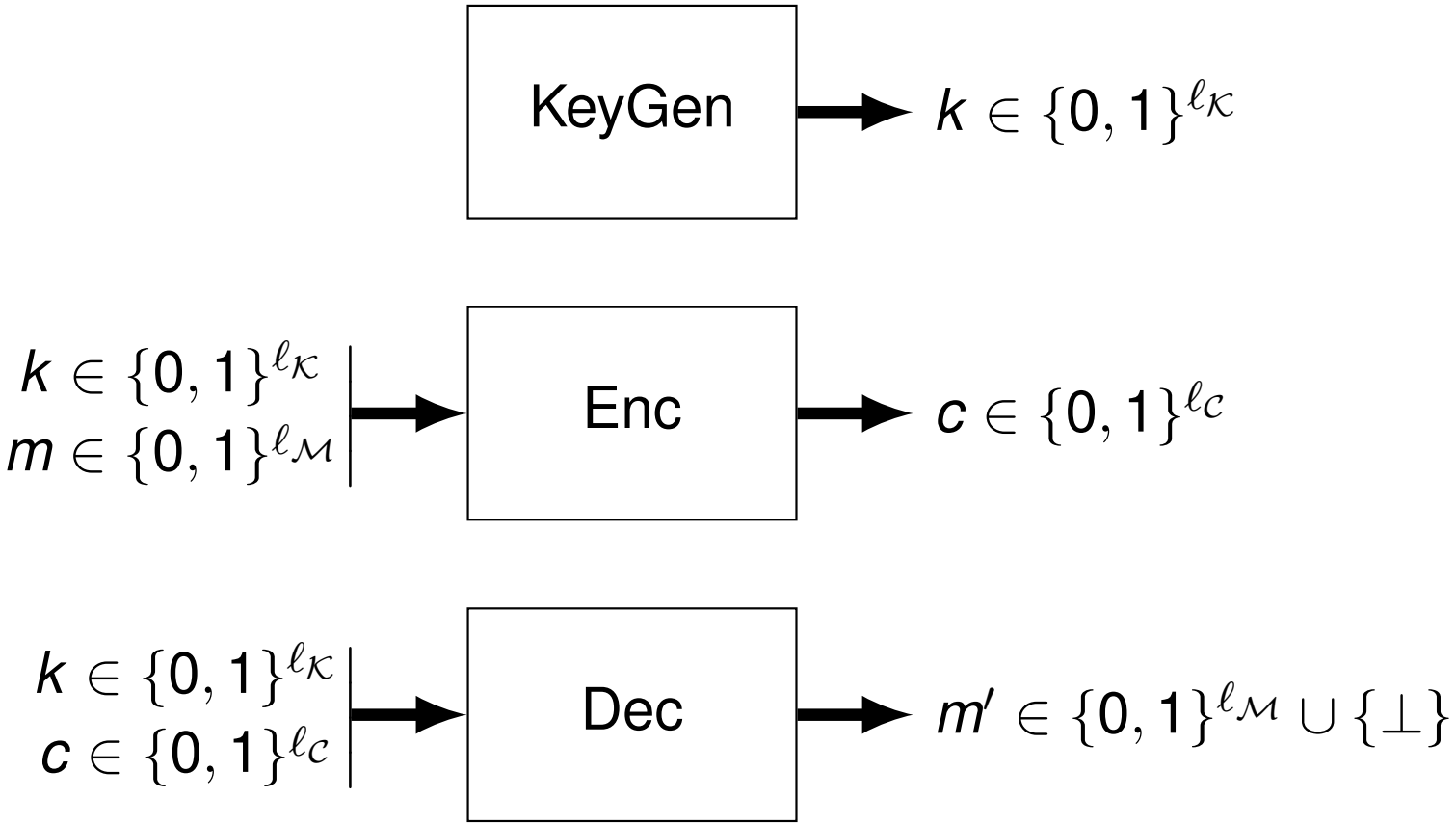
- **bounded:** the adversary has limited resources (computational power, memory)
- **unbounded:** the adversary has unlimited resources

Outline

Symmetric Encryption: Syntax



Symmetric Encryption: Correctness



$$\forall m \in \{0,1\}^{\ell_\mathcal{M}},\ \forall k \in \{0,1\}^{\ell_\mathcal{K}},\quad m = \mathsf{Dec}(k, \mathsf{Enc}(k, m))$$

Symmetric Encryption: Privacy

Informal definition:

"Impossible to find m from $c = \mathsf{Enc}(k, m)$ without k".

Symmetric Encryption: Privacy

Informal definition:

> *"Impossible to find m from* $c = \mathrm{Enc}(k, m)$ *without k".*

More formally:

> *For any fixed* $c \in \{0,1\}^{\ell_C}$*, and for a uniformly distributed* $k \in \{0,1\}^{\ell_K}$*, the probability that* $c = \mathrm{Enc}(k, m)$ *is the same for all* $m \in \{0,1\}^{\ell_M}$*.*

Symmetric Encryption: Privacy

Informal definition:

> *"Impossible to find m from $c = \mathrm{Enc}(k, m)$ without k".*

More formally:

> *For any fixed $c \in \{0,1\}^{\ell_C}$, and for a uniformly distributed $k \in \{0,1\}^{\ell_K}$, the probability that $c = \mathrm{Enc}(k, m)$ is the same for all $m \in \{0,1\}^{\ell_M}$.*

Or better:

Definition (Perfect Privacy)

For any probability distribution (source) of $M \in \{0,1\}^{\ell_M}$ and for a uniformly distributed $K \in \{0,1\}^{\ell_K}$, **the random variables** M **and** $\mathrm{Enc}(K, M)$ **are independent.**

Bounds for Perfect Symmetric Encryption

Theorem

For any correct and perfectly private symmetric encryption scheme $\ell_{\mathcal{C}} \geq \ell_{\mathcal{M}}$ *and* $\ell_{\mathcal{K}} \geq \ell_{\mathcal{M}}$.

Proof: (A simple combinatorial argument.) details...

Bounds for Perfect Symmetric Encryption

Theorem

For any correct and perfectly private symmetric encryption scheme $\ell_{\mathcal{C}} \geq \ell_{\mathcal{M}}$ *and* $\ell_{\mathcal{K}} \geq \ell_{\mathcal{M}}$.

Proof: (A simple combinatorial argument.) details...

Caveat: In practice, not all binary strings in $\{0,1\}^{\ell_{\mathcal{M}}}$ are valid messages. (Use a compression code and then encrypt.)

Bounds for Perfect Symmetric Encryption

Theorem

For any correct and perfectly private symmetric encryption scheme $\ell_{\mathcal{C}} \geq \ell_{\mathcal{M}}$ *and* $\ell_{\mathcal{K}} \geq \ell_{\mathcal{M}}$.

Proof: (A simple combinatorial argument.) details...

Caveat: In practice, not all binary strings in $\{0,1\}^{\ell_{\mathcal{M}}}$ are valid messages. (Use a compression code and then encrypt.)

The key cannot be reused for further encryptions!
Enc'$(k, m_1 \| m_2) = \text{Enc}(k, m_1) \| \text{Enc}(k, m_2)$ leaks information on $m_1 \| m_2$, unless $\ell_{\mathcal{K}} \geq 2\ell_{\mathcal{M}}$.

There is no perfect solution for binary private storage!
In practice, we need $\ell_{\mathcal{K}} \ll \ell_{\mathcal{M}}$.

A Generalization for Redundant Sources

Replace the sets $\{0,1\}^{\ell_{\mathcal{M}}}$, $\{0,1\}^{\ell_{\mathcal{K}}}$, $\{0,1\}^{\ell_{\mathcal{C}}}$ by probability distributions M, K, C on some finite sets $\mathcal{M}$, $\mathcal{K}$, $\mathcal{C}$.

Replace binary length by a measure of the average information given by a random variable (Shannon's entropy).

A Generalization for Redundant Sources

Replace the sets $\{0,1\}^{\ell_{\mathcal{M}}}$, $\{0,1\}^{\ell_{\mathcal{K}}}$, $\{0,1\}^{\ell_{\mathcal{C}}}$ by probability distributions M, K, C on some finite sets $\mathcal{M}$, $\mathcal{K}$, $\mathcal{C}$.

Replace binary length by a measure of the average information given by a random variable (Shannon's entropy).

Definition (Perfect Privacy)

For any probability distribution (source) of $M \in \mathcal{M}$ and for a uniformly distributed $K \in \mathcal{K}$, **the random variables** M **and** $\mathrm{Enc}(K, M)$ **are independent.**

A Generalization for Redundant Sources

Replace the sets $\{0,1\}^{\ell_{\mathcal{M}}}$, $\{0,1\}^{\ell_{\mathcal{K}}}$, $\{0,1\}^{\ell_{\mathcal{C}}}$ by probability distributions M, K, C on some finite sets $\mathcal{M}$, $\mathcal{K}$, $\mathcal{C}$.

Replace binary length by a measure of the average information given by a random variable (Shannon's entropy).

Definition (Perfect Privacy)

For any probability distribution (source) of $M \in \mathcal{M}$ and for a uniformly distributed $K \in \mathcal{K}$, **the random variables M and** $\mathrm{Enc}(K, M)$ **are independent.**

Theorem (Shannon)

For any correct and perfectly private symmetric encryption scheme $H(C) \geq H(M)$ and $H(K) \geq H(M)$.

Shannon's Entropy of a Discrete Random Variable (I)

It is a measure of the average amount of information in a random variable:

Definition (Shannon's Entropy)

Let X be a random variable defined on a finite set $\mathcal{X}$.
$H(X) = -\sum_{x \in \mathcal{X}} \Pr[X = x] \log_2 \Pr[X = x]$.

Shannon's Entropy of a Discrete Random Variable (I)

It is a measure of the average amount of information in a random variable:

Definition (Shannon's Entropy)

Let X be a random variable defined on a finite set $\mathcal{X}$.
$H(X) = -\sum_{x \in \mathcal{X}} \Pr[X = x] \log_2 \Pr[X = x]$.

Main properties:

- **Submodularity**: For any joint distribution (X, Y, Z) $H(X, Y, Z) + H(Z) \leq H(X, Z) + H(Y, Z)$. ▸ details...
- For a trivial (deterministic) random variable, x, $H(x) = 0$.
- For a uniform distribution $H(U_{\mathcal{X}}) = \log_2 |\mathcal{X}|$.

Shannon's Entropy of a Discrete Random Variable (II)

Other properties:

- $0 \leq H(X) \leq \log_2 |\mathcal{X}|$ (with r.h.s. equality for the uniform distribution).
- $H(X) \leq H(X, Y) \leq H(X) + H(Y)$ (with r.h.s. equality for independent variables).
- $H(g(X)) \leq H(X)$ (with equality for injective maps).
- $H(X, g(X)) = H(X)$.

Proof of Shannon's Theorem

Assume that $C = \text{Enc}(K, M)$, and K and M are independent

Correctness implies
$H(M, K, C) = H(\text{Dec}(K, C), K, C) = H(K, C) \leq H(K) + H(C)$

Perfect Privacy implies
$H(M, C) = H(M) + H(C)$

Therefore,
$H(K) + H(C) \geq H(M, K, C) \geq H(M, C) = H(M) + H(C) \quad \Rightarrow$
$H(K) \geq H(M)$

Moreover,
$H(K) + H(M) = H(K, M) = H(K, \text{Dec}(K, C)) \leq H(K, C) \leq$
$H(K) + H(C) \quad \Rightarrow \quad H(M) \leq H(C)$

Proof of Shannon's Theorem

Assume that $C = \text{Enc}(K, M)$, and K and M are independent

Correctness implies
$H(M, K, C) = H(\text{Dec}(K, C), K, C) = H(K, C) \leq H(K) + H(C)$

Perfect Privacy implies
$H(M, C) = H(M) + H(C)$

Therefore,
$H(K) + H(C) \geq H(M, K, C) \geq H(M, C) = H(M) + H(C) \quad \Rightarrow$
$H(K) \geq H(M)$

Moreover,
$H(K) + H(M) = H(K, M) = H(K, \text{Dec}(K, C)) \leq H(K, C) \leq$
$H(K) + H(C) \quad \Rightarrow \quad H(M) \leq H(C)$

Proof of Shannon's Theorem

Assume that $C = \text{Enc}(K, M)$, and K and M are independent

Correctness implies
$H(M, K, C) = H(\text{Dec}(K, C), K, C) = H(K, C) \leq H(K) + H(C)$

Perfect Privacy implies
$H(M, C) = H(M) + H(C)$

Therefore,
$H(K) + H(C) \geq H(M, K, C) \geq H(M, C) = H(M) + H(C) \quad \Rightarrow$
$H(K) \geq H(M)$

Moreover,
$H(K) + H(M) = H(K, M) = H(K, \text{Dec}(K, C)) \leq H(K, C) \leq$
$H(K) + H(C) \quad \Rightarrow \quad H(M) \leq H(C)$

The One-Time Pad

For fixed length binary strings, $\ell_{\mathcal{M}} = \ell_{\mathcal{K}} = \ell_{\mathcal{C}} = \ell$,
$\mathsf{Enc}(k, m) = k \oplus m$ and $\mathsf{Dec}(k, c) = k \oplus c$

The One-Time Pad

For fixed length binary strings, $\ell_{\mathcal{M}} = \ell_{\mathcal{K}} = \ell_{\mathcal{C}} = \ell$,
$\mathsf{Enc}(k, m) = k \oplus m$ and $\mathsf{Dec}(k, c) = k \oplus c$

For an abelian (additive) group $\mathcal{G}$, let $\mathcal{M} = \mathcal{K} = \mathcal{C} = \mathcal{G}$,
$\mathsf{Enc}(k, m) = m + k$ and $\mathsf{Dec}(k, c) = c - k$

Perfect secrecy is guaranteed if k is uniformly distributed in $\mathcal{K}$

The One-Time Pad

For fixed length binary strings, $\ell_{\mathcal{M}} = \ell_{\mathcal{K}} = \ell_{\mathcal{C}} = \ell$,
$\mathsf{Enc}(k, m) = k \oplus m$ and $\mathsf{Dec}(k, c) = k \oplus c$

For an abelian (additive) group $\mathcal{G}$, let $\mathcal{M} = \mathcal{K} = \mathcal{C} = \mathcal{G}$,
$\mathsf{Enc}(k, m) = m + k$ and $\mathsf{Dec}(k, c) = c - k$

Perfect secrecy is guaranteed if k is uniformly distributed in $\mathcal{K}$

It is normally used as an "information theoretical" piece in more complex protocols

Weakening Secrecy

To overcome the previous limitations, consider only **computationally bounded adversaries**:

Weakening Secrecy

To overcome the previous limitations, consider only **computationally bounded adversaries**:

Definition (Perfect Privacy)

For any probability distribution (source) of $M \in \mathcal{M}$ and for a uniformly distributed $K \in \mathcal{K}$, **the random variables** M **and** $\mathrm{Enc}(K, M)$ **are independent.**

Weakening Secrecy

To overcome the previous limitations, consider only **computationally bounded adversaries**:

Definition (Perfect Privacy)

For any probability distribution (source) of $M \in \mathcal{M}$ and for a uniformly distributed $K \in \mathcal{K}$, **the random variables** M **and** $\mathrm{Enc}(K, M)$ **are independent.**

Definition (Informal Computational Privacy)

For any probability distribution (source) of $M \in \mathcal{M}$ and for a uniformly distributed $K \in \mathcal{K}$, **the random variables** M **and** $\mathrm{Enc}(K, M)$ **behave for a bounded adversary as if they were independent.**

Weakening Secrecy

Definition (Informal Computational Privacy)

For any probability distribution (source) of $M \in \mathcal{M}$ and for a uniformly distributed $K \in \mathcal{K}$, **the random variables M and** $\text{Enc}(K, M)$ **behave for a bounded adversary as if they were independent.**

Based on efficient statistical tests a computationally bounded adversary can run.

Needs some extra assumptions from Complexity Theory.

Data Protection

Jorge L. Villar

MCYBERS, UPC, Fall 2025

END

Bounds for Perfect Symmetric Encryption: Details

Plot of Enc function:
$G = \{(m, k, c = \mathsf{Enc}(k, m))\}_{m \in \{0,1\}^{\ell_{\mathcal{M}}}, k \in \{0,1\}^{\ell_{\mathcal{K}}}}$

Correctness implies
$(m, k, c), (m', k, c) \in G \quad \Rightarrow \quad m' = m$
Then, fixing k, we obtain $\ell_{\mathcal{C}} \geq \ell_{\mathcal{M}}$

In addition, Perfect Privacy implies
$(m, k, c) \in G \quad \Rightarrow \quad \forall m' \; \exists k' \quad (m', k', c) \in G$
Then, fixing c, we obtain $\ell_{\mathcal{K}} \geq \ell_{\mathcal{M}}$

Bounds for Perfect Symmetric Encryption: Details

Plot of Enc function:
$G = \{(m, k, c = \mathrm{Enc}(k, m))\}_{m \in \{0,1\}^{\ell_{\mathcal{M}}}, k \in \{0,1\}^{\ell_{\mathcal{K}}}}$

Correctness implies
$(m, k, c), (m', k, c) \in G \quad \Rightarrow \quad m' = m$
Then, fixing k, we obtain $\ell_{\mathcal{C}} \geq \ell_{\mathcal{M}}$

In addition, Perfect Privacy implies
$(m, k, c) \in G \quad \Rightarrow \quad \forall m' \; \exists k' \quad (m', k', c) \in G$
Then, fixing c, we obtain $\ell_{\mathcal{K}} \geq \ell_{\mathcal{M}}$

Bounds for Perfect Symmetric Encryption: Details

Plot of Enc function:
$G = \{(m, k, c = \mathrm{Enc}(k, m))\}_{m \in \{0,1\}^{\ell_{\mathcal{M}}}, k \in \{0,1\}^{\ell_{\mathcal{K}}}}$

Correctness implies
$(m, k, c), (m', k, c) \in G \quad \Rightarrow \quad m' = m$
Then, fixing k, we obtain $\ell_{\mathcal{C}} \geq \ell_{\mathcal{M}}$

In addition, Perfect Privacy implies
$(m, k, c) \in G \quad \Rightarrow \quad \forall m' \; \exists k' \quad (m', k', c) \in G$
Then, fixing c, we obtain $\ell_{\mathcal{K}} \geq \ell_{\mathcal{M}}$

go back...

Shannon's Entropy Main Inequality (Submodularity)

The map $t \mapsto t \log_2 t$ is convex in $[0, +\infty)$

By discrete Jensen's inequality for convex functions, for positive numbers a_i, t_i, $i \in \mathcal{I}$

$$\sum_{i \in \mathcal{I}} a_i t_i = 1 \quad \Rightarrow \quad \sum_{i \in \mathcal{I}} a_i t_i \log_2 t_i \geq -\log_2 \sum_{i \in \mathcal{I}} a_i$$

Then use

$$t_{x,y,z} = \frac{\Pr[X = x, Y = y, Z = z] \Pr[Z = z]}{\Pr[X = x, Z = z] \Pr[Y = y, Z = z]}$$

$$a_{x,y,z} = \frac{\Pr[X = x, Z = z] \Pr[Y = y, Z = z]}{\Pr[Z = z]}$$

for all (x, y, z) such that $\Pr[X = x, Y = y, Z = z] \neq 0$.

Shannon's Entropy Main Inequality (Submodularity)

The map $t \mapsto t \log_2 t$ is convex in $[0, +\infty)$

By discrete Jensen's inequality for convex functions, for positive numbers a_i, t_i, $i \in \mathcal{I}$

$$\sum_{i \in \mathcal{I}} a_i t_i = 1 \quad \Rightarrow \quad \sum_{i \in \mathcal{I}} a_i t_i \log_2 t_i \geq -\log_2 \sum_{i \in \mathcal{I}} a_i$$

Then use

$$t_{x,y,z} = \frac{\Pr[X = x, Y = y, Z = z] \Pr[Z = z]}{\Pr[X = x, Z = z] \Pr[Y = y, Z = z]}$$

$$a_{x,y,z} = \frac{\Pr[X = x, Z = z] \Pr[Y = y, Z = z]}{\Pr[Z = z]}$$

for all (x, y, z) such that $\Pr[X = x, Y = y, Z = z] \neq 0$.

Shannon's Entropy Main Inequality (Submodularity)

The map $t \mapsto t \log_2 t$ is convex in $[0, +\infty)$

By discrete Jensen's inequality for convex functions, for positive numbers a_i, t_i, $i \in \mathcal{I}$

$$\sum_{i \in \mathcal{I}} a_i t_i = 1 \quad \Rightarrow \quad \sum_{i \in \mathcal{I}} a_i t_i \log_2 t_i \geq -\log_2 \sum_{i \in \mathcal{I}} a_i$$

Then use

$$t_{x,y,z} = \frac{\Pr[X = x, Y = y, Z = z] \Pr[Z = z]}{\Pr[X = x, Z = z] \Pr[Y = y, Z = z]}$$

$$a_{x,y,z} = \frac{\Pr[X = x, Z = z] \Pr[Y = y, Z = z]}{\Pr[Z = z]}$$

for all (x, y, z) such that $\Pr[X = x, Y = y, Z = z] \neq 0$.