Lecture 5: Pseudorandomness

Ryan Cunningham
University of Illinois
ECE 422/CS 461 – Fall 2017

Security News

- “SharknAT&To” vulnerabilities in Arris modems used by AT&T U-verse
- Equifax had “cybersecurity incident,” losing 143 million consumer’s data
- Security researchers discover (patch) vulnerability in IOTA cryptocurrency

ATTACKING HASHES

How do you find a collision?

Pigeonhole principle: collisions must exist

Input space $\{0,1\}^*$ larger than output $\{0,1\}^{256}$

Birthday attack: build a table with 2^{128} entries

With ~50% probability, have a collision

Cycle finding: "Tortoise and hare" algorithm

$h(x)$, $h(h(x))$, $h(h(h(x)$, .., $h^i(x)$

These are **generic** - actual attacks rely on **structure** of the particular hash function

Concrete Parameterization

How large of a digest size should we choose?

1. Estimate an attacker's budget

e.g., the entire NSA

- **2. Factor in hardware improvements**

3. Consider the best known attacks

Reduction from protocol to well-studied problem

4. Add a safety margin

If all goes well, adding 1 bit increases search space by 2x

Other hash functions:

MD5

Once ubiquitous

Broken in 2004

Easy to find collisions today

SHA1

Currently widely used

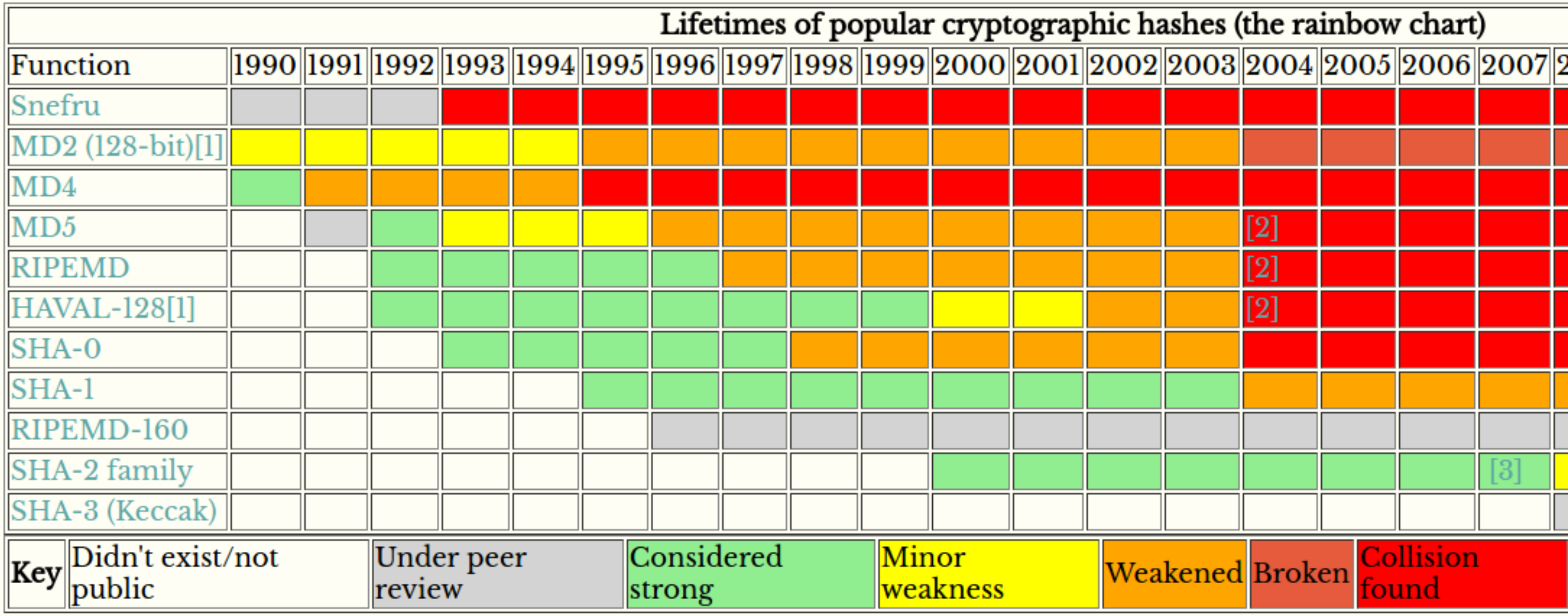
Collisions recently found!

Don’t use in new applications

SHA3

Different construction: “Sponge”

Not susceptible to ***length-extension*** http://valerieaurora.org/hash.html



[1] Note that 128-bit hashes are at best 2^{64} complexity to break; using a 128-bit hash is irresponsible based on sheer digest length.

[2] What happened in 2004? Xiaoyun Wang and Dengguo Feng and Xuejia Lai and Hongbo Yu happened.

[3] In 2007, the NIST launched the SHA-3 competition because "Although there is no specific reason to believe that a practical attack on any of the SHA-2 family of hash functions is imminent, a successful collision attack on an algorithm in the SHA-2 family could have catastrophic effects for digital signatures." One year later the first strength reduction was published.

The Hash Function Lounge has an excellent list of references for most of the dates. Wikipedia now has references to the rest.

Schneier on Security

SHA-1 Collision Found

The first collision in the SHA-1 hash function has been found.

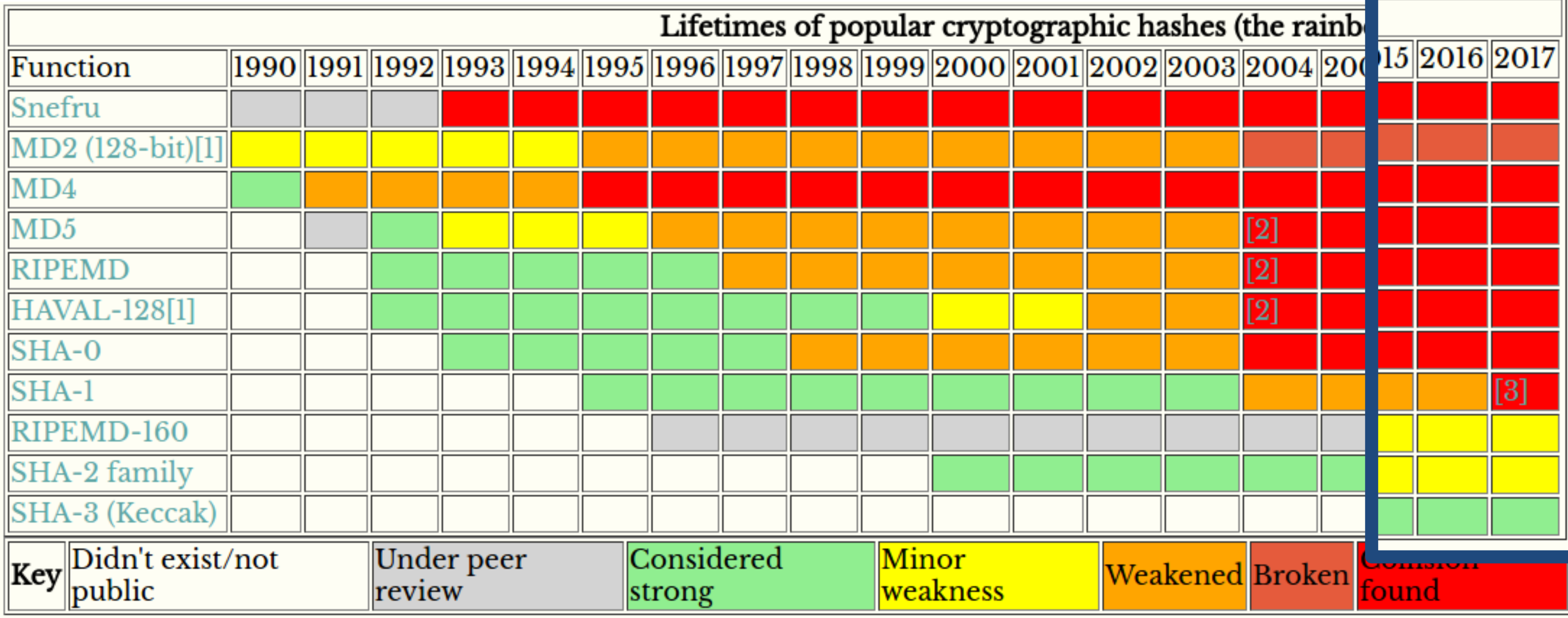
This is not a surprise. We've all expected this for over a decade, watching computing power increase. This is why NIST standardized SHA-3 in 2012.

EDITED TO ADD (2/24): Website for the collision. (Yes, this brute-force example has its own website.)

EDITED TO ADD (3/7): This 2012 cost estimate was pretty accurate.

Tags: cryptanalysis, cryptography, hashes, SHA-1

Posted on February 23, 2017 at 3:29 PM • 31 Comments http://valerieaurora.org/hash.html



[1] Note that 128-bit hashes are at best 2^64 complexity to break; using a 128-bit hash is irresponsible based on sheer digest length.

[2] What happened in 2004? Xiaoyun Wang and Dengguo Feng and Xuejia Lai and Hongbo Yu happened.

[3] In 2007, the NIST launched the SHA-3 competition because "Although there is no specific reason to believe that a practical attack on any of the SHA-2 family of hash functions is imminent, a successful collision attack on an algorithm in the SHA-2 family could have catastrophic effects for digital signatures." One year later the first strength reduction was published.

The Hash Function Lounge has an excellent list of references for most of the dates. Wikipedia now has references to the rest.

MESSAGE AUTHENTICATION CODES

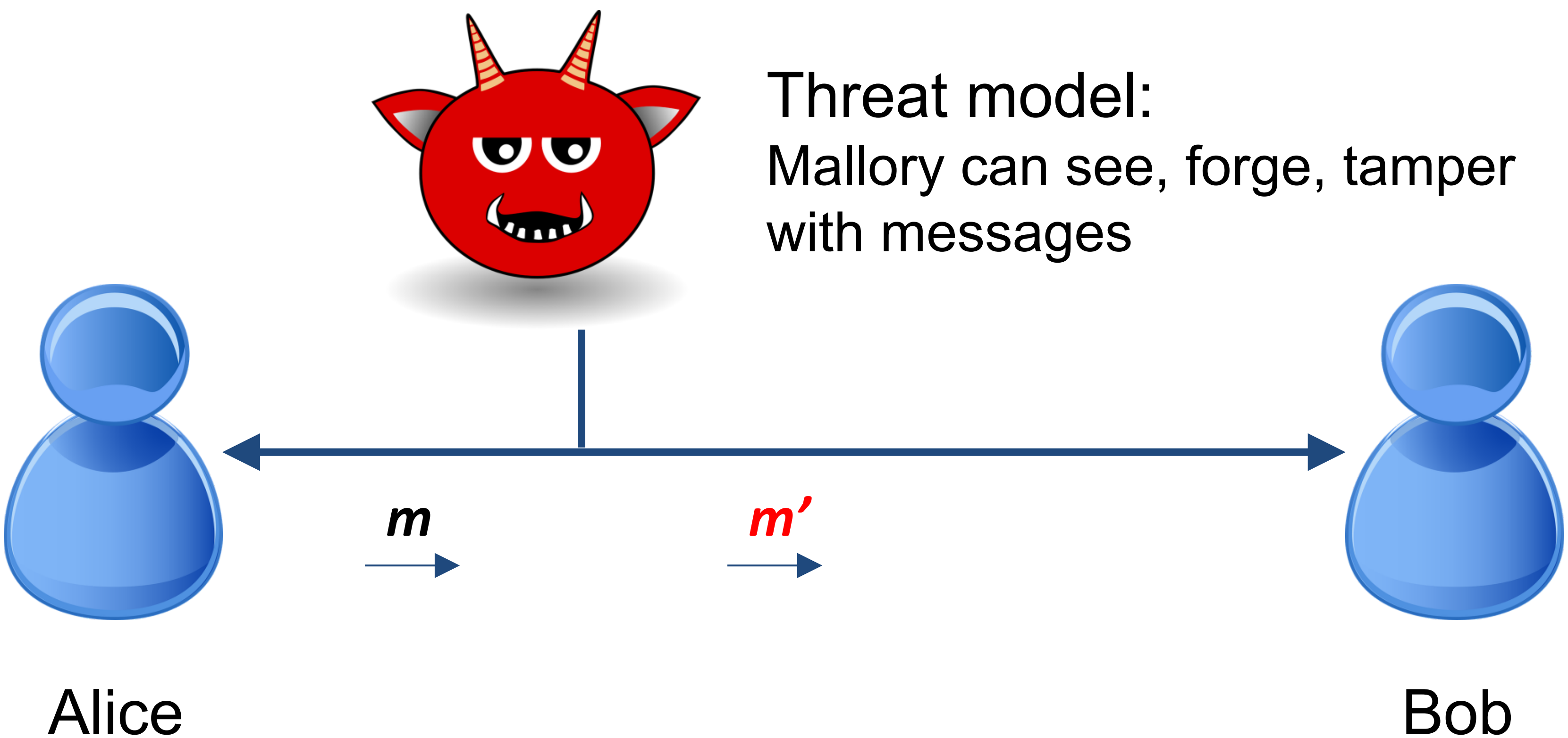
Goal: Message Integrity

Alice wants to send message ***m*** to Bob

Mallory wants to trick Bob into accepting a message Alice didn't send

Threat model:

Mallory can see, forge, tamper with messages



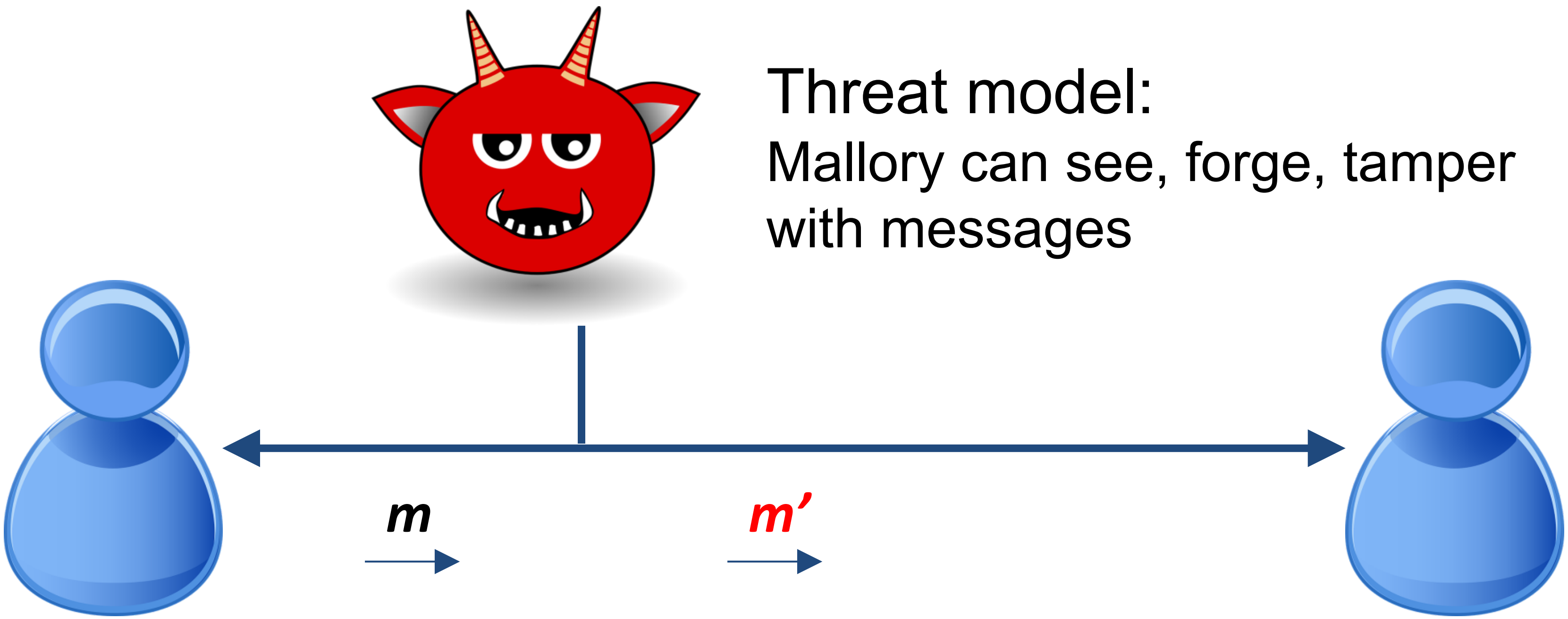
Goal: Message Integrity

Alice wants to send message ***m*** to Bob

Mallory wants to trick Bob into accepting a message Alice didn't send

Threat model:

Mallory can see, forge, tamper with messages



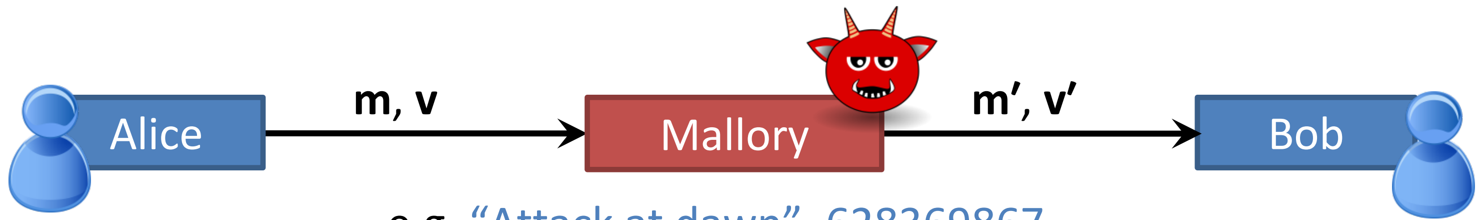
Alice, ***x*** Bob, ***x***

Setup assumption: ***shared secret***

Solution: Message Authentication Code (MAC)

1. Alice computes **v** := ***f***(**m**)

2.



e.g. “Attack at dawn”, 628369867…

3. Bob verifies that **v′** = ***f***(**m′**), accepts message iff this is true

Function ***f*** ?

Easily computable by Alice and Bob;
But <u>NOT</u> computable by Mallory

(Idea: Secret only Alice & Bob know)

We’re sunk if Mallory can learn ***f***(**m’**) for any **x** ≠ **m’**!

Candidate ***f***:

Random function

Input: Any size up to huge maximum

Output: Fixed size (e.g. 256 bits)

Defined by a giant lookup table that’s filled in by flipping coins

0	→	0011111001010001…
1	→	1110011010010100…
2	→	0101010001010000…
⋮		⋮

Completely <u>impractical</u>

Provably <u>secure</u>

Want a function that's practical but "looks random"…

Pseudorandom function (PRF)

Let's build one:

Start with a big *family of functions*

$\boldsymbol{f}_0, \boldsymbol{f}_1, \boldsymbol{f}_2, \ldots$ all known to Mallory

Use $\boldsymbol{f}_{\mathbf{k}}$, where **k** is a secret value (or "key") known only to Alice/Bob

k is (say) 256 bits, chosen randomly

Kerckhoffs's Principle

Don't rely on secret functions

Use a secret key, to choose from a function family

More formal definition of a secure **PRF**:

Game against Mallory

1. We flip a coin secretly to get bit **b**
2. If **b**=0, let ***g*** be a random function
   If **b**=1, let ***g*** = $\boldsymbol{f}_{\mathbf{k}}$, where **k** is a randomly chosen secret
3. Repeat until Mallory says "stop":
   Mallory chooses **x**; we announce ***g***(**x**)
4. Mallory guesses **b**

We say ***f*** is a *secure PRF* if Mallory can't do better than random guessing*

i.e., $\boldsymbol{f}_{\mathrm{k}}$ is indistinguishable in practice from a random function, unless you know k

Important fact: There's an algorithm that always wins for Mallory

A solution for Alice and Bob:

1. Let $\boldsymbol{f}$ by a secure PRF
2. In advance, choose a random **k** known only to Alice and Bob
3. Alice computes $\mathbf{v} := \boldsymbol{f}_{\mathbf{k}}(\mathbf{m})$

1. 

2. Bob verifies that $\mathbf{v'} = \boldsymbol{f}_{\mathbf{k}}(\mathbf{m'})$, accepts message iff this is true

What assumptions are made here?

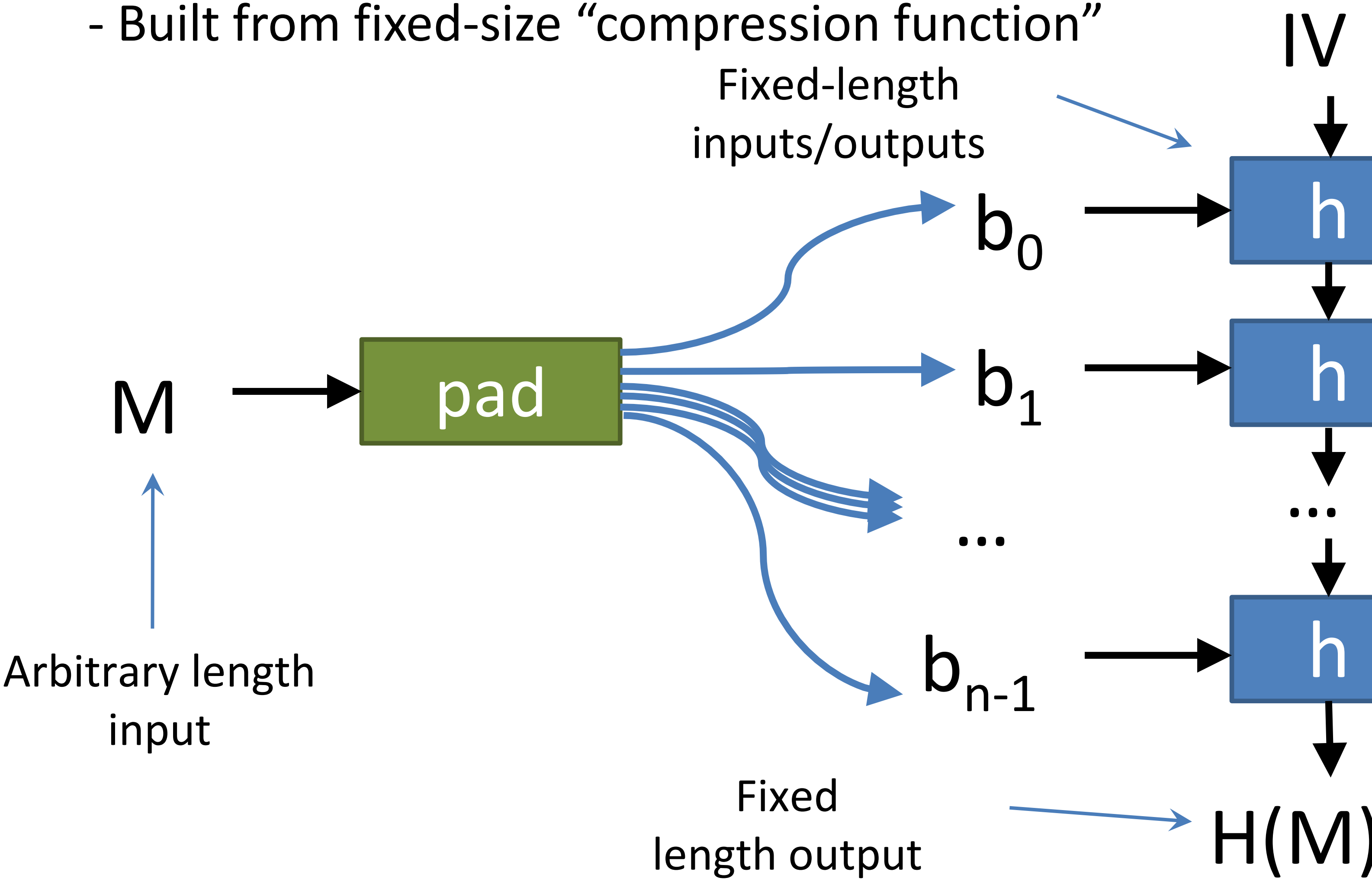
What if Alice and Bob want to send more than one message?

Is this a PRF?

$$\boldsymbol{f}_{\mathbf{k}}(\mathbf{m}) = \mathrm{SHA256}(\ \mathbf{k}\ ||\ \mathbf{m}\ )$$

Merkle–Damgård Construction

- Arbitrary-length input
- Fixed-length output
- Built from fixed-size “compression function”



Recommended Approach: Hash-based MAC (HMAC)

HMAC-SHA256 see RFC 2104

HMAC$_{\mathbf{k}}$($\mathbf{m}$) =

$$\underbrace{\mathrm{SHA256}}_{\text{SHA256 function}}\left(k \underbrace{\oplus}_{\text{XOR}} \underbrace{c_1}_{\texttt{0x3636...}} \underbrace{\|}_{\text{Concatenation}} \mathrm{SHA256}\left(k \oplus \underbrace{c_2}_{\texttt{0x5c5c...}} \| m \right)\right)$$

SHA256 function

takes arbitrary length input,

returns 256-bit output

Message Authentication Code (MAC)

e.g. HMAC-SHA256

vs.

Cryptographic hash function

e.g. SHA256

not a strong PRF

Used to think the distinction didn't matter, now we think it does

e.g., ***length extension attacks***

Better to use a MAC/PRF (not a hash)

```
$ openssl dgst -sha256 -hmac <key>
```

MAC Crypto Game

Game against Mallory

1. Give Mallory MAC(k, m_i) for all m_i in M
In other words, Mallory has an ***oracle***
Mallory can choose next m_i after seeing answer

2. Mallory tries to discover MAC(k, m') for a new m' not in M

We can show the **MAC game** ***reduces*** to the **PRF game**. Mallory wins MAC game → she wins PRF game.

This is a **Security Proof**

What is a **Security Proof**?

- A ***reduction*** from an ***attack on your protocol*** to an attack on a ***widely studied, hard problem (presumed)***
- Excludes large classes of attacks, guides **composition**
    - Proofs are in **models**. So, attack outside the model!
- It does **NOT** ***prove*** that your protocol is ***secure***
- We don’t know if there are any hard problems!
- The field of **Modern Cryptography** is based on proofs
- Most widely used primitives (SHA-256, AES, DSA) have no security proof. We rely on them because they’re widely studied

Randomness
and
Pseudorandomness

Review

Problem:

Integrity of message sent from Alice to Bob

Append bits to message that only Alice and Bob can make

Solution:

Message Authentication Code (MAC)

Practical solution:

Hash-based MAC (HMAC) – ***HMAC-SHA256***$_k$(M)

Where do these random keys **k** come from … ?

Careful: We're often sloppy about what is "random"

True Randomness

Output of a physical process that is inherently random

Scarce, and hard to get

Pseudorandom Function (PRF)

Sampled from a family of functions using a key

Pseudorandom generator (PRG)

Takes small seed that is really random

Generates a stream (arbitrarily long sequence) of numbers that are "as good as random"

Definition: **PRG** is secure if it's indistinguishable from a random stream of bits

Similar game to PRF definition:

1. We flip a coin secretly to get a bit **b**
2. If **b**=0, let ***s*** be a truly random stream
   If **b**=1, let ***s*** be $\boldsymbol{g}_{k}$ for random secret **k**
3. Mallory can see as much of the output of ***s*** as he/she wants
4. Mallory guesses **b**,
   wins if guesses correctly

g is a secure PRG if no winning strategy for Mallory*

Here's a *simple PRG that works:*

For some random k and PRF f,

output: $f_k(0)$ || $f_k(1)$ || $f_k(2)$ || …

Theorem: If f is a secure PRF, and g is built from f by this construction, then g is a secure PRG.

Proof: Assume f is a secure PRF, we need to show that g is a secure PRG.

Proof by contradiction:

1. Assume g is *not* secure; so Mallory can win the PRG game
2. This gives Mallory a winning strategy for the PRF game:
   a. query the PRF with inputs 0, 1, 2, …
   b. apply the PRG-distinguishing algorithm
3. Therefore, Mallory can win PRF game; this is a contradiction
4. Therefore, g is secure

Where do we get true randomness?

Want “indistinguishable from random”
which means: adversary can’t guess it

Gather lots of details about the computer that the adversary will have trouble guessing [Examples?]

Problem: Adversary can predict some of this

Problem: How do you know when you have enough randomness?

Modern OSes typically collect randomness, give you API calls to get it

e.g., Linux:

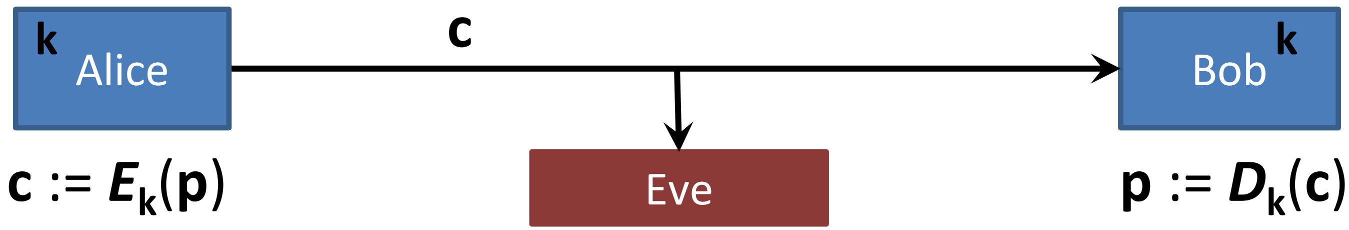
`/dev/random` a device that gives random bits, blocks until available

`/dev/urandom` gives output of a PRG, nonblocking, seeded from `/dev/random` *eventually*

Confidentiality

Confidentiality

Goal: Keep contents of message **p** secret from an *eavesdropper*



Terminology

p	plaintext
c	ciphertext
k	secret key
E	encryption function
D	decryption function

Digression: **Classical Cryptography**

Caesar Cipher

First recorded use: Julius Caesar (100-44 BC)

Replaces each plaintext letter with one a fixed number of places down the alphabet

Encryption: $\mathbf{c}_i := (\mathbf{p}_i + \mathbf{k}) \bmod 26$

Decryption: $\mathbf{p}_i := (\mathbf{c}_i - \mathbf{k}) \bmod 26$

e.g. (**k**=3):

```
          Plain:  ABCDEFGHIJKLMNOPQRSTUVWXYZ
          +Shift: 33333333333333333333333333
=Cipher:          DEFGHIJKLMNOPQRSTUVWXYZABC

          Plain:  fox   go wolverines
          +Key:   333   33 3333333333
=Cipher:          ira   jr zroyhulqhv
```

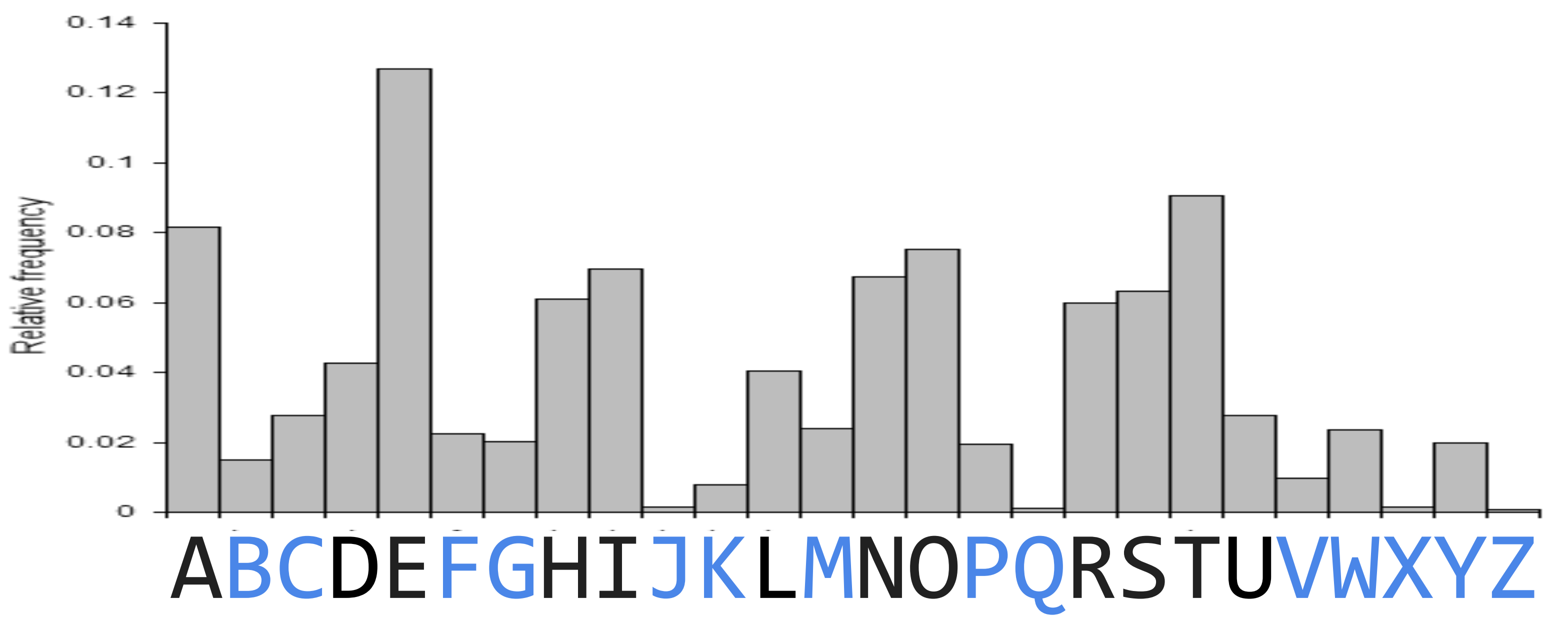
Cryptanalysis of the Caesar Cipher

Only 26 possible keys:

Try every possible **k** by "*brute force*"

Can a computer recognize the right one?

Use *frequency analysis*: English text has distinctive letter frequency distribution



Later advance: Vigènere Cipher

First described by Bellaso in 1553,
later misattributed to Vigenère

Called « le chiffre indéchiffrable »
(“the indecipherable cipher”)

Encrypts successive letters using a sequence of Caesar ciphers determined by the letters of a keyword

For an **n**-letter keyword **k**,

Encryption: $\mathbf{c_i} := (\mathbf{p_i} + \mathbf{k}_{\mathbf{i} \bmod \mathbf{n}}) \bmod 26$

Decryption: $\mathbf{p_i} := (\mathbf{c_i} - \mathbf{k}_{\mathbf{i} \bmod \mathbf{n}}) \bmod 26$

Example: **k**=ABC (i.e. $\mathbf{k}_0=0$, $\mathbf{k}_1=1$, $\mathbf{k}_2=2$)

```
Plain:   bbbbbb    amazon
+Key:    012012    012012
=Cipher:     bcdbcd    anczpp
```

Cryptanalysis of the Vigènere Cipher

Simple, if we know the keyword length, **n**:

1. Break ciphertext into **n** slices
2. Solve each slice as a Caesar cipher

How to find **n**? One way: Kasiski method

Published 1863 by Kasiski (earlier known to Babbage?)

Repeated strings in long plaintext
will sometimes, by coincidence,
be encrypted with same key letters

Plain:	**CRYPTO**ISSHORTFOR**CRYPTO**GRAPHY
+Key:	ABCDABCDABCDABCDABCDABCD
=Cipher:	**CSASTP**KVSIQUTGQU**CSASTP**IUAQJB

Distance between repeated strings in ciphertext is likely a multiple of key length e.g., distance 16 implies **n** is 16, 8, 4, 2, 1

[What if key is as long as the plaintext?]

Kerckhoff’s Principles

1st: The system must be practically, if not mathematically, indecipherable;

2nd: The system must not require secrecy and must not cause inconvenience should it fall into the hands of the enemy;

3rd: The key must be able to be used in communiques and retained without the help of written notes, and be changed or modified at the discretion of the correspondents;

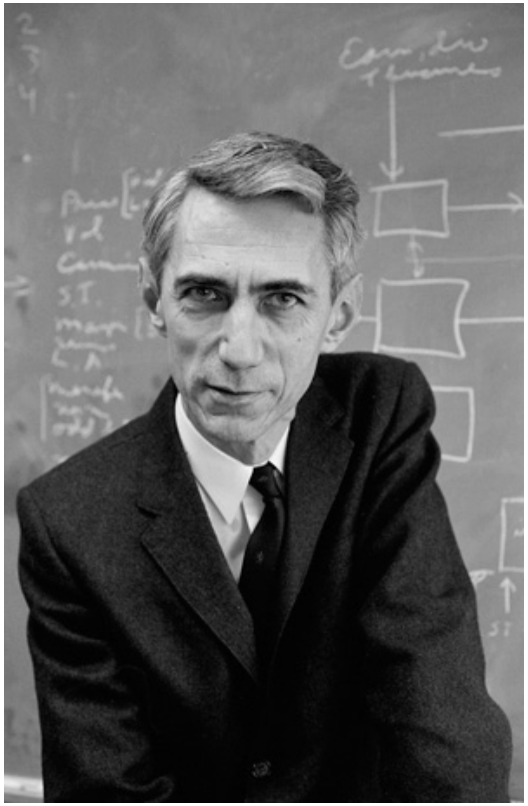
4th: The system must be compatible with telegraphic communication;

5th: The system must be portable, and remain functional without the help of multiple people;

6th: Finally, it’s necessary, given the circumstances in which the system will be applied, that it’s easy to use, is undemanding, not overly stressful, and doesn’t require the knowledge and observation of a long series of rules

"Shannon's Maxim"

The enemy knows the system.



“Schneier's law”

Any fool can invent a cipher that he himself cannot break.

One-time Pad (OTP)

Alice and Bob jointly generate a secret, very long, string of random bits (the one-time pad, **k**)

To encrypt: $\mathbf{c}_i = \mathbf{p}_i$ xor $\mathbf{k}_i$

To decrypt: $\mathbf{p}_i = \mathbf{c}_i$ xor $\mathbf{k}_i$

a	**b**	**a** xor **b**
0	0	0
0	1	1
1	0	1
1	1	0

a xor **b** xor **b** = **a**

a xor **b** xor **a** = **b**

"one-time" means you should never reuse any part of the pad.

If you do:

Let $\mathbf{k}_i$ be pad bit

Adversary learns (**a** xor $\mathbf{k}_i$) and (**b** xor $\mathbf{k}_i$)

Adversary xors those to get (**a** xor **b**), which is useful to him [How?]

Provably secure [Why?]

Usually impractical [Why? Exceptions?]

Obvious idea: Use a **pseudorandom generator** instead of a truly random pad

(Recall: Secure **PRG** inputs a seed **k**, outputs a stream that is practically indistinguishable from true randomness unless you know **k**)

Called a **stream cipher**:

1. Start with shared secret key **k**
2. Alice & Bob each use **k** to seed the PRG
3. To encrypt, Alice XORs next bit of her generator's output with next bit of plaintext
4. To decrypt, Bob XORs next bit of his generator's output with next bit of ciphertext

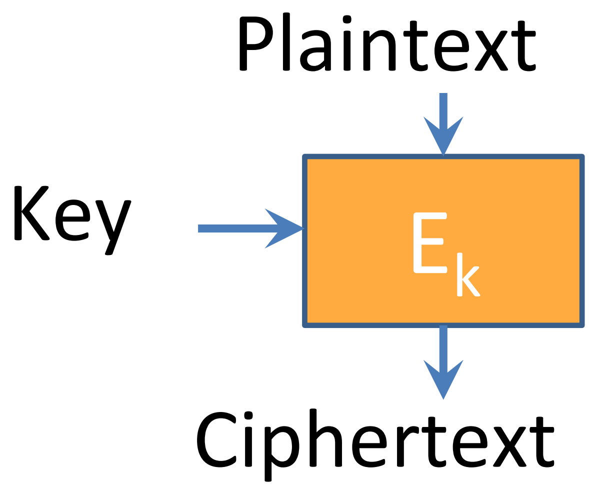
Works nicely, but: don't *<u>ever</u>* reuse the key, or the generator output bits

Another approach: **Block Ciphers**

Functions that encrypts fixed-size blocks with a reusable key.

Inverse function decrypts when used with same key.

The most commonly used approach to encrypting for confidentiality.



A block cipher is <u>not</u> a pseudorandom function [Why?]

What we want instead:

pseudorandom permutation (PRP)

function from **n**-bit input to **n**-bit output

distinct inputs yield distinct outputs (one-to-one)

Defined similarly to **PRF**:

practically indistinguishable from a *random permutation* without secret **k**

Basic challenge: Design a hairy function that is invertible, but only if you have the key

Minimal properties of a good block cipher:

- Highly nonlinear (“confusion”)
- Mixes input bits together (“diffusion”)
- Depends on the key

Definition: a cipher is "Semantically Secure"

Similar game to PRF/PRG/PRP definition:

1. We flip a coin secretly to get a bit **b**, random secret **k**
2. Mallory chooses arbitrary $\mathbf{m_i}$ in **M**, gets to see $\boldsymbol{Enc_k(m_i)}$
3. Mallory chooses two messages $\mathbf{m'_0}$ and $\mathbf{m'_1}$ not in **M**
4. If **b**=0, let **c** be $\boldsymbol{Enc_k(m'_0)}$
   If **b**=1, let **c** be $\boldsymbol{Enc_k(m'_1)}$
5. Mallory can see **c**
6. Mallory guesses **b**, wins if guesses correctly

We can prove this follows from a PRP definition. **[Fun to try!]**

Also known as: IND-CPA "Chosen plaintext attack"

Today's most common block cipher:

AES (**Advanced Encryption Standard**)

- Designed by NIST competition, long public comment/discussion period
- Widely believed to be secure, but we don't know how to prove it
- Variable **key size** and **block size**
- We'll use 128-bit key, 128-bit block (are also 192-bit and 256-bit versions)
- Ten **rounds**: Split **k** into ten **subkeys**, performs set of operations ten times, each with diff. subkey

Each AES round

128-bits in, 128-bit sub-key, 128-bits out

$S_{0,0}$	$S_{0,1}$	$S_{0,2}$	$S_{0,3}$
$S_{1,0}$	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$
$S_{2,0}$	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$
$S_{3,0}$	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$

Four steps:
picture as operations on a
4x4 grid of 8-bit values

1. Non-linear step
   Run each byte through a non-linear function (lookup table)

2. Shift step: Circular-shift each row: **i**th row shifted by **i** (0-3)

3. Linear-mix step
   Treat each column as a 4-vector; multiply by constant invertible matrix

4. Key-addition step
   XOR each byte with corresponding byte of round subkey

To decrypt, just undo the steps, in reverse order

Remaining problem:

How to encrypt longer messages?

Padding:

Can only encrypt in units of cipher blocksize, but message might not be multiples of blocksize

Solution: Add padding to end of message

Must be able to recognize and remove padding afterward

Common approach: Add **n** bytes that have value **n**

[Caution: What if message ends at a block boundary?]

Cipher modes of operation

We know how to encrypt one block,

but what about multiblock messages?

Different methods, called “cipher modes”

Straightforward (but bad) approach:

ECB mode (encrypted codebook)

Just encrypt each block independently

$\mathbf{C_i} := \boldsymbol{E}_{\mathbf{k}}(\mathbf{P_i})$

[Disadvantages?]

Cipher modes of operation

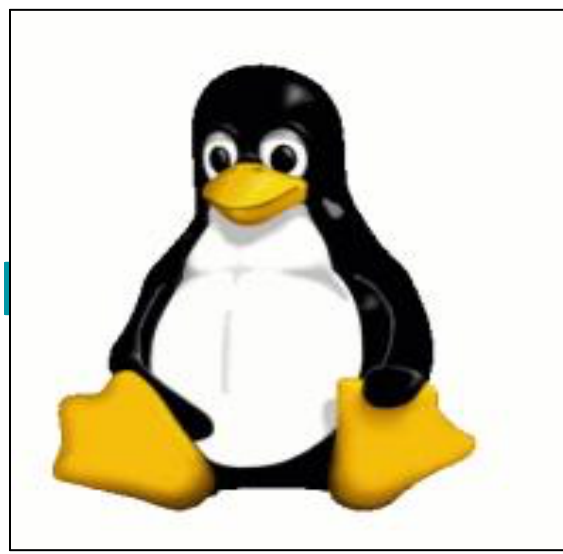
We know how to encrypt one block,

but what about multiblock messages?

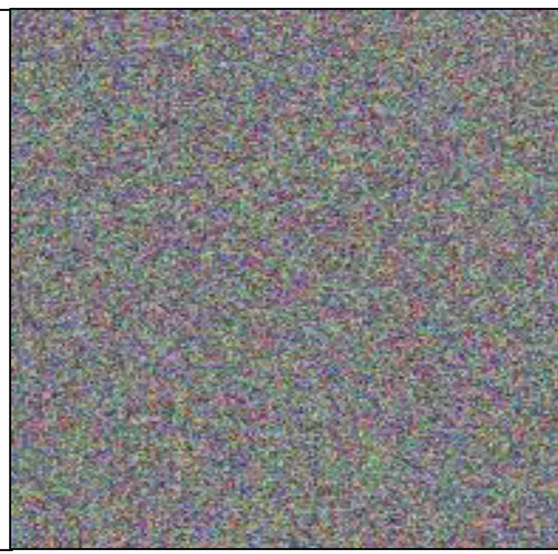
Different methods, called “cipher modes”

Straightforward (but bad) approach:

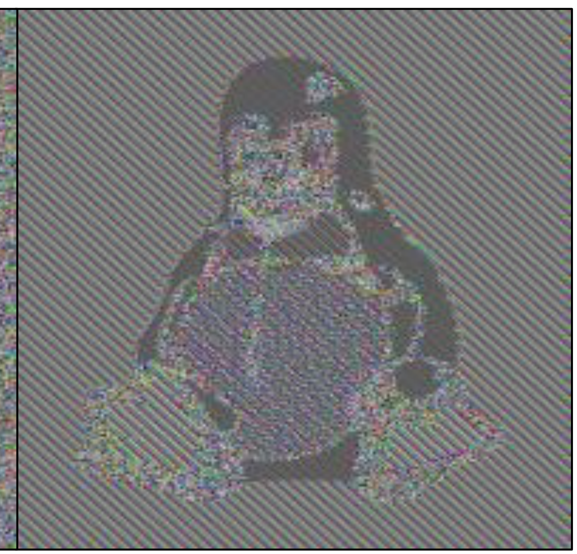
ECB mode (encrypted codebook)



Plaintext



Pseudorandom



ECB mode

Better (and common):

CBC mode (cipher-block chaining)

Lame-CBC (for illustration only)

For each block $\mathbf{P}_i$:

1. Generate random block $\mathbf{R}_i$
2. $\mathbf{C}_i := (\mathbf{R}_i \mid\mid \boldsymbol{E}_k(\mathbf{P}_i \text{ xor } \mathbf{R}_i))$

[Pros and cons?]

Real CBC

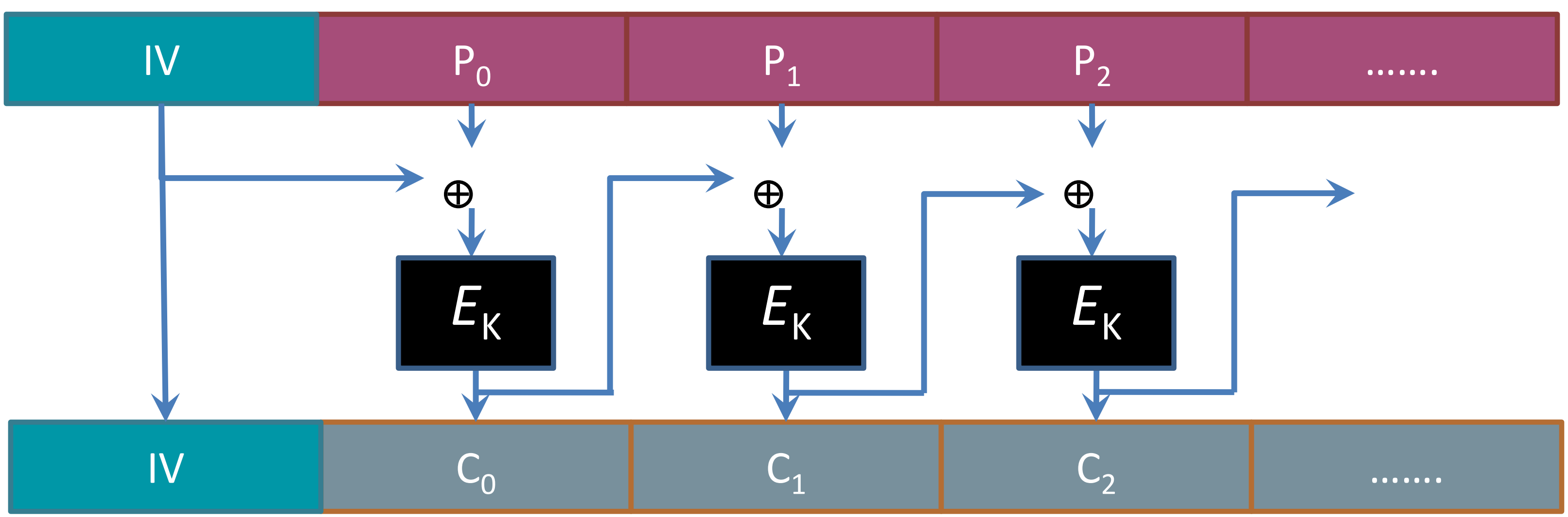
Replace $\mathbf{R_i}$ with $\mathbf{C_{i-1}}$

No need to send separately

Must still add one random $\mathbf{R_{-1}}$ to start, called "**initialization vector**" ("**IV**")

[Is CBC space-efficient?]

Illustration: CBC Encryption



Using OpenSSL to do AES encryption from the command line

```
$ KEY=$(openssl rand -hex 16)
```

Generates a random string

```
$ openssl aes-256-cbc -in mymsg.txt -out mymsg.enc
-p -K ${KEY} -iv $(openssl rand -hex 16)
   key=8582D9E1A36DA4DB065394FB1F401DB3
   iv =DBB272FE6486C4D9B09DBE464E080468
```

Prints the key and IV

```
$ openssl aes-256-cbc -d -in mymsg.enc -out mymsg.txt
-K ${KEY} -iv <iv from above>
```

- By default, uses the standard padding described earlier
- Unfortunately, you have to handle prepending/extracting the IV on your own

Other modes

OFB, CFB, etc. – used less often

Counter mode

Essentially uses block cipher as a pseudorandom generator

XOR **i**th block of message with $\boldsymbol{E}_{\mathbf{k}}$(message_id || **i**)

[Why do we need message_id?]

[Do we need a message_id for CBC mode?]

[Recover after errors? Decrypt in parallel?]

What is **NOT** covered by Semantic Security?

- "Malleability" attacks

Given just some ciphertexts, can the attacker create new ciphertexts that Bob decrypts the wrong value?

- Encryption does NOT IMPLY integrity!

Often you really want both ("authenticated encryption")

- Chosen Ciphertext attacks

The "semantic security" definition does not allow the adversary to see decryptions of (potentially garbage) ciphertexts chosen by the adversary

Assumption we've been making so far:

Alice and Bob ***shared a secret key*** in advance

Amazing fact:

Alice and Bob can have a **public** conversation to derive a shared key!

Security News

Report on lightweight cryptography

NISTIR 8114

Report on Lightweight Cryptography

Kerry A. McKay
Larry Bassham
Meltem Sönmez Turan
Nicky Mouha
Computer Security Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8114

March 2017

2.3 Lightweight Cryptographic Primitives 4
2.3.1 Lightweight Block Ciphers 5
2.3.2 Lightweight Hash Functions 6
2.3.3 Lightweight Message Authentication Codes 6
2.3.4 Lightweight Stream Ciphers 6