Protecting Digital Content

3 - 4.5 hours

Ernst L. Leiss
Department of Computer Science
University of Houston
coscel@cs.uh.edu

Support of this research under NSF grants 0313880, 0453498, 0519316 and 1241772 is acknowledged

Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not reflect the views of the National Science Foundation

Contents

0. Introduction

Information controls access to resources

Data (raw transmission) vs. information (content)

The problem of perfect copies (especially for digital signatures)

Qualitative vs. quantitative change in computing (certain quantitative changes amount to qualitative ones:: queries may simply be infeasible, so they are not posed)

Insiders vs. outsiders: Insiders are much more trouble

Three basic areas of data security

Statistical database security – inference control

Authorization systems

Cryptosystems

1. Intellectual Property

Text: Literary works, programs

Images: Photographs, video

Audio: Music

Securing property rights to intellectual property: through copyright and patents (also trade secrets).

Legal versus technical means of protection

- **The law does not protect against infractions, it only punishes infractions. Technical means provide impossibility.**
- **E. g., prohibiting legally access versus encrypting**

2. Cryptographic Techniques

Allow hiding the **information content** of a message – the message itself is accessible to anybody.

Also permit **authentication**, digital signatures, and verification of the **integrity** of a message.

Stream versus block ciphers

Two fundamentally different approaches:

- Symmetric encryption
- Public-key encryption

Both have advantages and disadvantages.

Attacks:

Cipher-text only

Known plain-text

Chosen plain-text

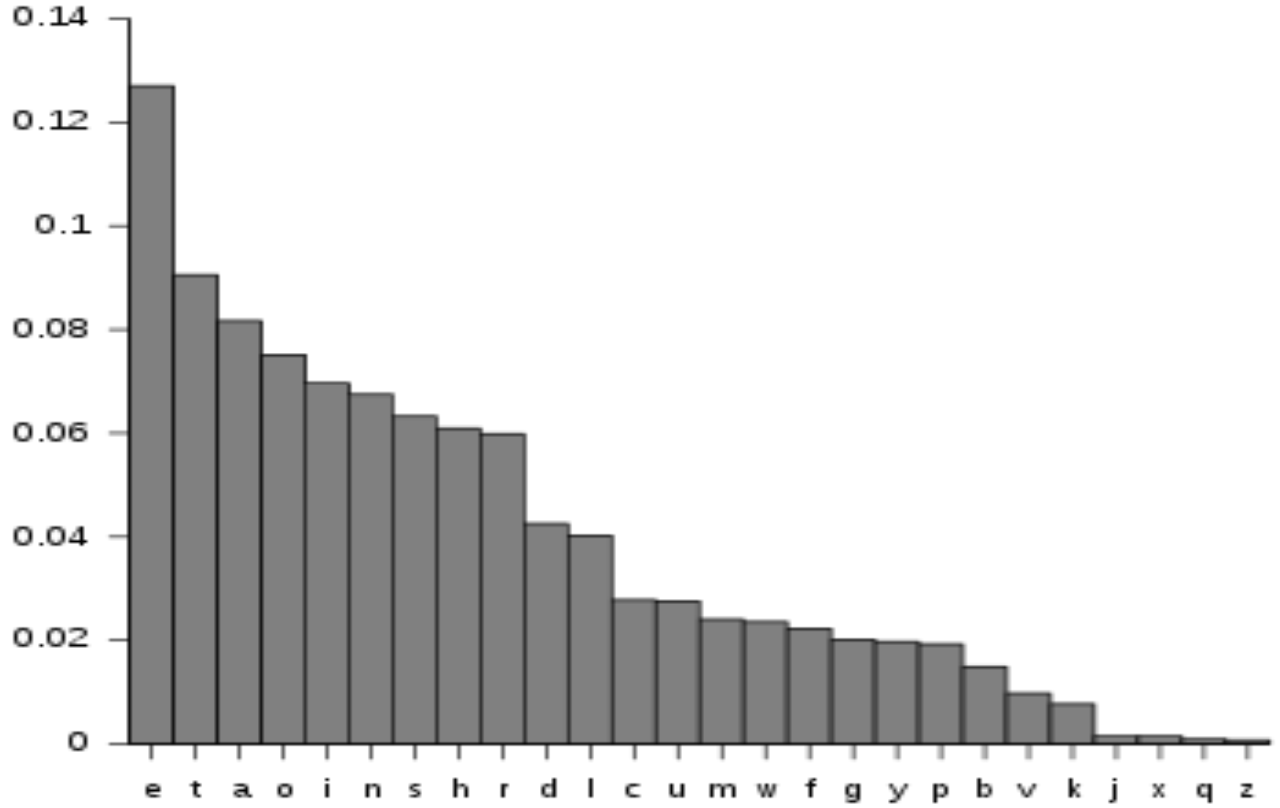
An encryption scheme must protect at least against known plain-text attacks

Fundamental requirement for any encryption scheme: **The distribution of n-grams must be flat**, for n = 1, 2, 3, …

Implication: Any single-bit error in a block of the cipher-text will corrupt about half of all the bits in the resulting decrypted text.

Stream ciphers are even more affected: All bits after the corrupted bit are subject to this problem.

Relative frequencies of letters in English text.



Symmetric encryption: Must keep both encryption and decryption key secret (knowing one allows one to determine the other with great ease).

Public-key encryption: One key is public (similar to a telephone number), the other is private. Crucial is the requirement that knowing one key does not permit determining the other.

One-way functions: Encryption and decryption are inverses of each other; computing one must be easy, computing the other hard.

Symmetric encryption: Transposition and substitution operations.

DES and successors.

The problem of key length.

Main public-key scheme: RSA

Choose two large prime p and q; compute $n = p \cdot q$.

Choose e relatively prime to $(p-1)\cdot(q-1)$; compute d multiplicative inverse of e with respect to $(p-1)\cdot(q-1)$.

Publish n and e; keep private d (as well as p and q).

Encryption of message M to be sent: $C = M^e \bmod n$.

Decryption of cipher-text C received: $M = C^d \bmod n$.

(M and C are viewed as a number between 0 and n-1.)

Fundamentally based on the difference in the computational complexities for **factoring** integers (needed to determine p and q from the known n – allows finding d given e) and testing for **primality** (needed to set up the scheme which requires having two primes, p and q)**.**

Testing whether a given s-digit integer is a **prime number**: Polynomial time in s [$O(s^3)$ under some mild assumptions].

Factoring a given s-digit integer: No algorithm known that works in polynomial time in s.

Current state-of-the-art:

- Determining d when knowing p and q (each of length O(s)): polynomial in s with small exponent.
- Determining d when knowing n: super-polynomial in s.

But: No non-trivial **lower bound** known for factoring.

Assume n participants, message length m

	Symmetric	Public-key
Advantages	O(m) time en/decryption	O(n) keys No prior contact needed
Disadvantages	$O(n^2)$ keys Prior contact required	en/decryption >> O(m) time

Key management typically a problem, but more so for symmetric schemes.

In practice, hybrid schemes: Use public-key to distribute keys, which are then used to encrypt symmetrically the actual messages. Obtains advantages of both approaches.

3. Applications of Cryptography

Digital signatures

Can be based on symmetric encryption

Most often based on public-key schemes:

A sending a signed message M to B: $E(D(M,L_A),K_B)$

with K_B B's public encryption key and L_A A's private decryption key.

Absence of any physical signature: Only A could have sent this message to B, since only A is capable of producing it.

Authentication

Introduce time dependence into a protocol that establishes the identity of the sender.

Data integrity

Insert redundancy into data to be secured and then encrypt.

Since the redundancy is not apparent in the cipher-text, it cannot be forged without decrypting.

Simple example: Duplicate the message before encryption.

The amount of redundancy is a measure of the probability with which the integrity can be violated:

s bits redundancy $\leftrightarrow$ $1/_2$s probability of successful defeat

4. Text-Based versus Signal-based Information

Text-based: Ordinary text, code – consists of characters for some fixed alphabet

Signal-based: Voice, video – consists of signals or pixels

Size: Kilobytes vs. Gigabytes

Error tolerance: Zero tolerance vs. high tolerance (~5%)

Redundancy: Low versus extremely high

Compression is applicable almost exclusively to signal-based information

5. Security and Integrity of Digital Signal-Based Content

Security: Read access – who may **obtain** information
Integrity: Write access – who may **change** information

Classical cryptography requires **security** for all its keys
Public-key cryptography requires **integrity** for its public keys

Usage of digital content:

Deny access to unauthorized users (security) or
verify that content has not been changed (integrity).

6. Single-Step versus Multi-Step Methods

Digital content is used by the receiver
=> Conversion of transmitted file into useful format

If mechanism does not survive the conversion: Single-step
e. g.: decryption :: recipient has now the unencrypted file and can do anything with it, for example transmit it to others **unencrypted**.

If mechanism does survive the conversion: Multi-step

Traditional methods are all single-step.

7. Traditional Methods

Traditional techniques for safeguarding integrity and security of digital media:

(A) Encryption-based schemes

Encrypt the entire file and transmit only the encrypted file

(B) Signature/witness/hash schemes

Compute a piece of information dependent on the file (signature, witness code, hash) of the file and transmit the original file plus the additional piece of information

Main disadvantages:

- Interference with data compression techniques (A)
- Compute-intensive (A, B)
- Usually require separate, secure part of file (B)
- Lack of robustness: transmission errors can have major effects (A, B)

8. Introduction to Digital Watermarks

The problem of perfect copies

Establishing ownership of intellectual property

Digital watermarks

- Visible, invisible
- Robust, fragile
- Time-invariant, time-variant

Types of attack

- Geometric operations
- Filtering
- Manipulation of images (insertion, deletion of pixels)
- Manipulation of sequencing of frames

9. Requirements for Watermarks

Invisible, robust watermarks

Attempts at removing, destroying, obliterating, or overwriting should result in a severe and very noticeable degradation of the image before the watermark is lost

Must be able to survive

- loss-less and lossy compression techniques,
- other common video processing techniques
  - scaling,
  - cropping,
  - resizing,
  - filtering (including changes in the color scheme, e.g., reducing the color palette [e. g., from 16 bit to 8 bit])
  - transcoding

The watermark must allow the owner to demonstrate ownership conclusively (to a judge)

Sufficient information must be present in the watermark for this purpose

The watermark must not affect the perceived quality of the video (perception is everything!)

Inserting the watermark must not substantially increase the overall complexity (beyond what MPEG already requires) of generating and using the video.

10. The Basic Organization of a Video File

Color **JPEG** **MPEG**

Color

Humans: combinations of the primary colors red, blue, and yellow (the typical rainbow arrangement)

Video hardware: RGB model (Red, Green, Blue)
a pixel is associated with (RGB) representing the color intensities;
(000) represents black in this scheme (absence of everything),
(kkk) white (presence of everything), (k00) pure red, …
k is the quantization granularity for each primary color.
For k is 255: 2^{8+8+8} or 2^{24} different representable colors.
Smaller k == less faithfulness in the color scheme
Larger k == greater faithfulness

Color schemes with more than 24 bits result in improvements in image quality that are virtually imperceptible to the naked human eye.

RGB transformed into (Y, C_b, C_r)

Y luminance,

C_b blue chrominance,

C_r red chrominance.

$$Y = 0.587\,G + 0.299R + 0.114\,B$$

$$C_b = 0.564\,(B - Y)$$

$$C_r = 0.713\,(R - Y)$$

Human eye less perceptive for color than for luminance:

for natural images the chrominance component of a signal can tolerate a more reduced bandwidth than the luminance, without affecting significantly the perceived image quality.

Typically, the bandwidth for chrominance may be one half to one quarter that of luminance without affecting human perception.

Lossy JPEG compression

1. Decomposition of the image into blocks of size 8*8 pixels

2. Discrete Cosine Transform (DCT) applied to each 8*8 matrix

generates a new 8*8 matrix consisting of the coefficients of increasing spatial frequency.

3. Quantization applied to the 64 DCT coefficients

yields an 8*8 Quantization table Q(u,v) consisting of integers

4. The 63 AC coefficients in Q are concatenated into a **zigzag scan**; in terms of (u,v), this scan is

DC: **00**

AC:

01 10
20 11 02
03 12 21 30
40 31 22 13 04
05 14 23 32 41 50
60 51 42 33 24 15 06
07 16 25 34 43 52 61 70
71 62 53 44 35 26 17
27 36 45 54 63 72
73 64 55 46 37
47 56 65 74
75 66 57
67 76
77

Note that the values are addresses that provide access to the actual coefficients; thus, 00 is the address of the DC coefficient.

5. Run-length coding

Replaces a sequence of zeroes by the number of zeroes in the sequence.

Major compression in JPEG

from a certain point p on in the sequence of the 63 AC coefficients

of the zigzag scan, replace the rest by zeroes

p parameter:

small, e.g., 5, image quality reduced and compression greatly improved

large, e.g., 30, image quality is virtually unaffected but reduced compression

6. Apply Huffman coding to the resulting sequences.

Significant compression rations (up to 10+) can be achieved while retaining high image quality.

MPEG is based on JPEG

Aims to remove temporal redundancies (from one frame to the next) after JPEG has been applied to remove the spatial redundancies within each frame

Temporal redundancies are detected by motion estimation whereby portions of images in consecutive frames are matched up

I-pictures, P-pictures, and B-pictures

Intra or I-pictures: encoded (using JPEG) without any reference to other frames; most expensive to store/transmit

Predicted or P-pictures depend only on the preceding I- or P-picture

Bi-directionally Predicted or B-pictures depend on both preceding or following I- or P-pictures

P- and B-pictures require less storage/transmission bandwidth.

B-pictures fill in the gaps between I- (and P-) pictures and provide the largest savings.

The number of P- and B-frames between two consecutive I-pictures constitutes a tradeoff between compression and accuracy:

Making this value large results in more savings: P- and B-frames are cheaper, I-frames more expensive.

Making it too large affects the quality of the interpolated frames: P- and B-frames are inferred from surrounding frames.

Compression ratios of 200 with MPEG without sacrificing much quality

11. Implementing Time-Variant Watermarks

Embed the watermark in the middle AC coefficients

(at the beginning: noticeable to human eye,
at the end: can be removed without any consequences)

Time-invariant watermarks

Embed the same image into every picture to be watermarked

Typically, I-pictures are watermarked

Time-variant watermarks

Embed a watermark **video** into the video to be watermarked

A watermark **video** consists of N_0 of pictures

embed this video, frame by frame.

Into the video frame number i to be watermarked, embed the watermark frame number i, for i=1, 2, …, N_0. If there are more than N_0 pictures in the video to be watermarked, the next batch of N_0 pictures get a second copy of the watermark video embedded, etc.

N_0 parameter; if N_0=1: ordinary, time-invariant watermarking scheme.

If N_0 is greater than 1, the approach is time-variant.

Upper bound for N_0: number of I-pictures in original video

(do not insert watermark frames into P- or B-pictures)

Information in the watermark video

include sequencing information

N_0 = number of I-pictures in original video

12. Conclusion for Watermarks

Summary of operations

loss-less compression:	INV, VAR
lossy compression:	INV, VAR
filtering:	INV, VAR
conversion (digital $\leftrightarrow$ analog representation)	INV, VAR
cropping:	INV, VAR
scaling:	INV, VAR
translation:	INV, VAR
rotation:	INV, VAR
superposition of another watermark:	INV, VAR
adding new frames:	INV, VAR
removing original frames:	**VAR**
permuting original scenes:	**VAR**
legal demonstration of ownership:	INV, VAR

Watermarks are a multi-step integrity mechanism

File can pass through many hands; at each step the watermark can be verified, but verification does not imply that an unwatermarked file is produced.

Watermarks cannot achieve security, only integrity.

Watermarks are inexpensive

Little processing is required, beyond what MPEG requires already.

Watermarks are convenient

Users do not need to do anything if they are not interested in the integrity of the file. Only concerned users must carry out additional steps.