Hidden Data

Week 5

Steganography

The “art” of hiding data

What is Steganography?

- The word “steganography” comes from Greek words: steganos “covered” and graphie “writing”
- This is quite different from a stegosaurus “covered lizard”
- And different from stenography “narrow writing”

What is Steganography

- Steganography is a science (and art) of hiding a message within another message
  - the secret message is referred to as the payload (or carrier medium)
  - the normal message, that contains the secret message, is the carrier
- Both parties know how the message was hidden and can secretly transfer messages

Cryptography vs. Steganography

- Cryptography
  - does not hide the communication
  - encodes the data to prevent eavesdroppers from understanding the content
  - presence of encrypted data may cause suspicions
- Steganography
  - hides the communication
  - the data may or not be encrypted
  - if they don’t know about it, how can they be suspicious?

History

- Steganography is not a new technology – the idea of secret messages is as old as humanity
- Is has been used since ancient times
  - invisible ink (1st century AD to WW II)
  - tattoos or drawings
  - some characters reflect under special light
  - pin punctures in type
  - microdots

Cryptography vs. Steganography

- Cryptography
  - does not hide the communication
  - encodes the data to prevent eavesdroppers from understanding the content
  - presence of encrypted data may cause suspicions
- Steganography
  - hides the communication
  - the data may or not be encrypted
  - if they don’t know about it, how can they be suspicious?
New techniques have been invented following technological advances.
Steganography can be traced back to 440 BC, from the Histories of Herodotus.

Histiaeus wanted to start a revolt to free his people from the Persians.
To send a message, he shaved off his most trusted slave’s head.
Slave’s head was then tattooed with the message.
The hair was allowed to regrow – hiding it.

At the time, wax tablets were used to write text – similar to today’s white-boards.
Demeratus needed to warn Greece about an upcoming attack by the Persians.
He etched the message into the wooden backing of the wax tablet.
The wax front of the tablet covered the message.
By melting all the wax, the message was revealed.

A null cipher is a technique that embeds a secret message into seemingly innocent (or meaningless) sentences.
For instance, the first letter of each word can be meaningful – when these letters are combined, they form another sentence.

During World War II, a German Spy sent a secret message using a null cipher.
The message was hidden in every second letter of the sentence.

Apparent neutral’s protest is thoroughly discounted and ignored. Isman hard hit.
Blockade issue affects pretext for embargo on by-products, ejecting suets and vegetable oils.

Pershing sails from NY June 1.
April 2006: London's High Court ruled if author Dan Brown had plagiarized *The DaVinci Code*. Suit was brought by another author. Ultimately, the Judge Peter Smith ruled in favor of Brown. But his 71-page written judgment raised eyebrows.

Judge gets Cheeky…

Smith: "I can’t discuss the judgment, but I don’t see why a judgment should not be a matter of fun."

Microdot Technology is a technique of hiding a message inside a single letter or symbol. Basically the message shrunk down in size to about 1 millimeter or less. Just look how much information is stored on a piece of microfilm or microfiche! The message can be placed in a period or the "title" above a j or an i. It was used in World War I, II and even today.

Text Position

Secret messages can also be hidden in what appears to be innocuous data. For instance: data can be hidden in text formatting in subtle ways:

- line spacing
- word or character spacing
- minor changes to shapes of characters

For humans, we might not be able to see the difference… but computers can!

Are These the Same?
Nope…

So, let’s act like spies
What neat trick can we use to hide a message inside something?
Let’s create a new one

Cloak and Dagger

Steganography Today

So much data, so, so much data

Steganography Today

Nowadays, practically everything is stored in digital format
People seldom realize how much data is there!
Digital data is used everywhere
- e-mail
- online images – websites, etc...
- video
- real-time games

Steganography Today

Since everything is binary…
- any binary data can be stored in any binary data
- there are limits depending on the size of the carrier and the payload
- …and some file format issues
- but, for the most part, it is possible
Examples:
- text stored in an image
- image stored in text
- music file stored in an image
- image stored in a music file
- etc….

There are a wide variety of techniques
When analyzing a technique, there are several attributes of importance
- *perceptibility* indicates how much the data payload distorts the carrier
- *capacity* is how much data can be hidden
- *robustness* refers to how well the data can survive if the carrier is modified or manipulated

Often, there is a give-and-take…
- increasing one attribute may weaken another
- affects perceptibility, robustness and capacity
- e.g. the higher the capacity → the more compact the data → more perceivable

The person hiding the data must make a decision on the technique depending on the weaknesses / strengths

As the World becomes more based on information, the transmission of hidden data becomes easier
Think of the shear volume of data out there – and how little one message, file, etc… constitutes

Like all technologies, steganography can be used for both good and evil
Good Uses of Steganography

- Watermarks to detect forgeries
  - holding a $20 bill up to the light and seeing a watermark
  - ultraviolet marking on credit cards
- Fighting against government intrusion
  - some states are oppressive (e.g. Iran)
  - resistance groups can use it to talk
- Hiding confidential / value data

Evil Uses of Steganography

- Concealing a plan for terroristic threats
  - al-Qaeda may have used steganographic software to communicate before the 9-11 attacks, this has not yet been confirmed
  - this is a huge threat to the government
- Hiding contraband
  - can allow perpetrators (such as child pornographers) to exchange information
  - stolen data – spying, etc…

al-Qaeda Master Plans

- April 2012
- German officials detained a man in Berlin who appeared on a terrorist watchlist
- On him, they found routine documents and travel items
- However, the suspect had a memory card sewn into his underwear
- The card contained a pornographic videos called “kick ass” and “Sexy Tanja”
- Why sew it into your underwear?
- Cryptologists and steganographers spent weeks on the memory card

al-Qaeda Master Plans

- Hidden, encrypted, within the video there were over 100 hidden files
- Some of the content:
  - "Lessons learned"
  - "Future plans"
  - and more documents detailing strategy
- On “future plan”
  - take control of a cruise ship (low security, tons of victims)
  - dress victims in orange jump suits (like those in Guantanamo Bay)
  - behead them live (and slowly) and upload videos to terrorist websites
Text & Steganography

Very basic, and very hard to detect

---

Text

- Even simple text files are often used to store secret messages
- However, it is not efficient and little data can be put in the carrier
- Classic **null cipher**
  - every first letter of each word (or second, third, etc....) holds the message
  - this restricts the text of the message
  - awkward prose may be a red flag

---

Text: Whitespace

- **Whitespace** can also be used to store a message
- What is whitespace?
  - this is space between words, the blank space after a sentence, etc....
  - ...looks simply "white" on paper
  - it is seen as “empty” by people, and thus a great place to hide data

---

Whitespace: Word Spacing

- The number of spaces between words can contain the message
- e.g. single space $\rightarrow$ 0, two spaces $\rightarrow$ 1
- The text will be **visually** altered, although few may notice
- The file size will also increase

---

Whitespace: End Line

- Spaces can be added to the end of each line (after the text)
- e.g. no space $\rightarrow$ 0, single $\rightarrow$ 1
- Visual appearance of the text...
  - will not be altered
  - but the capacity is far smaller
- The file size will still increase from the original

---

End Line Example

In a society under the forms of which the stronger faction can readily unite and oppress the weaker, anarchy may as truly be said to reign...

James Madison
Federalist Paper 51
In a society under the forms of which the stronger faction can readily unite and oppress the weaker, anarchy may as truly be said to reign...

James Madison – Federalist Paper 51

It is Invisible to the Reader

In a society under the forms of which the stronger faction can readily unite and oppress the weaker, anarchy may as truly be said to reign...

Images & Steganography

Pictures look good for secrets!

Images are a Popular Choice

- Images are one of the most widely used mediums
- As computers get more and more powerful, the size and complexity of graphics will increase

Images are a Popular Choice

- What we think of "small" images can contain millions of bytes
- Inside that space...
  - there readily available space to hide data
  - larger the image size, the more information you can hide
Images are a Popular Choice

- Digital images are made up of pixels
- Each pixel used 3 (or more) bytes to represent the red-green-blue color
- This means:
  - each pixel can have 16,777,216 unique values
  - changing a red-green-blue value slightly cannot be picked up the human eye
  - … but computers can tell the difference

Completely Under the RADAR

- The most popular technique for images
- Hide the data in the least-significant-bits
  - these are the bits (of each byte) which contain the smallest values (the rightmost bits)
  - usually only the least-significant-bit (bit with a group value of 1) is used
  - but more can be used to increase the capacity

Least-Significant-Bit Encoding

- Advantages
  - simple
  - high capacity – 3 or more bits per pixel
  - low perceptibility – data hides in color “noise”
- Disadvantages
  - not very robust – lossy compression will easily destroy the data
  - … as a result, this technique is used on lossless images such as BMP and PNG

Least-Significant-Bit Encoding

- Least Significant Bit Example
- Byte Can Be Stored in 3 Pixels
Increasing Payload Capacity

- The two least-significant-bits can also be used
- This basically doubles the capacity of carrier
- … but
  - more of the color data is altered
  - it might be visually noticeable (still not likely)
  - easier to detect by steganalysis tools

Using 2 Bits Per Byte → 2 Pixels

Data
11001011

00101011 00101000 00101010 00101001
00101011 00101011 00101011

Other Image Techniques

There is more than pixels!

Demonstration….

Let’s Look at Some Secret Messages

Encoding: Palettes

- Not all images store full color information in each pixel
- Palette images…
  - do not store red-green-blue values separately for each pixel
  - instead, they store a index into a table that contain the RGB color

Images Contain Indexes into the Palette Table
Palette Steganography

- To make it work, the software needs to modify the palette.
- Palette contains seemingly duplicate colors:
  - the two “duplicates” actually differ slightly
  - one color is selected to store the 0 – usually in the least-significant-bit of one of the values
  - the other color stores the 1 in its bit
- When the image is redrawn, the steganography software can select the palette entry that hides the 0 or 1 – as it is needed.

Two Redundant Colors… Almost

<table>
<thead>
<tr>
<th>Image</th>
<th>Palette</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 2 1 1</td>
<td>11011011</td>
</tr>
<tr>
<td>0 6 4 0</td>
<td>10011001</td>
</tr>
<tr>
<td>5 0 2 1</td>
<td>00011111</td>
</tr>
<tr>
<td>2 1 2 4</td>
<td>11011011</td>
</tr>
<tr>
<td>4 3</td>
<td></td>
</tr>
</tbody>
</table>

Palettes: How It Works…

- Drawbacks…
  - basically, there are two ‘duplicate’ palette entries for every real color
  - 256 color palette can only have 128 “real” colors
  - if the image has more than 128 unique colors, some will have to be remapped to their closest matches
- If colors are remapped…
  - it can cause radical color shifts for color images
  - this can alert investigators

As a result, grayscale images are often used

- they only have a max 256 colors!
- shift between two grays is subtle (and hard to detect)

Images that can use palettes:

- PNG – either true color or palette
- BMP – either true color or palette
- GIF – only palette

Metadata and Steganography

- Many file types information about the main data in the file
- This is called *meta-data*
  - its “data about data”
  - various greatly between different file types
  - e.g. created date, author, software used, etc...

Metadata

Hiding data in data about data!
(yes, quite confusing!)
` metadata is often used to store hidden data

- Remember now attachments are stored in e-mail?
- Base64 (and related systems) can store any binary data in simple ASCII text
- So, any text metadata can store any data
- However, this can be detected by steganalysis software

### JPEG Files

- Computer photographs

### JPEGs

- JPEGs are stored using in the JPEG File Interchange Format (JFIF)
- Designed specifically for photographs
- They use lossy compression
  - compression changes the original red-green-blue values
  - so least significant bits cannot be used
The JPEG file header contains information about the image:
This includes
- density of the picture (pixels per inch/cm)
- location data (extension)
- thumbnail graphic (created by the software that saves the JPEG)

- Thumbnail is an independent image
  - ...so, all JPEGs can contain two images
  - it is stored as a true color uncompressed image
  - maximum of 256x256 pixels
- So, the least-significant bit approach can be used on the thumbnail

GIF Files
- There are actually two versions of the file
- They are basically compatible, but differ in a key feature
  - GIF 87 – Original format, no animations
  - GIF 89 – Animation!
- The file format allows a number of features:
  - multiple images (each can have its own palette)
  - timing and placement control
  - looping – simple iteration, nothing fancy
  - transparency
- It can also store "comments"
  - these are hidden – never displayed
  - often used to save the name the generating software
  - can be used to hide a secret message

GIFs were created by CompuServe in 1987 to use with its online software
- The format is popular today to create simple images and basic animation
A Secret Message….. Here?

We be spies!

PNG Files

The bling-bling of pngs!

PNG Files

- PNG (pronounced “Ping”) is a relatively new file format
- It was designed to replace other bitmap file formats and work with the Internet

PNG Chunks

- PNGs contain multiple number of “chunks”
- Each chunk…
  - can contain up to 4,294,967,296 bytes
  - contains a 4 byte ASCII identifier
  - are backwards and forwards compatible. If a PNG reader does not understand a “chunk” identifier, it skips the section
  - it also contains a CRC error check

PNG Chunks

- Critical Chunks are necessary to identify the image, set bounds and other items that must be included
- Ancillary Chunks
  - help the image, but are not required
  - they can be safely ignored – though the image might not look correct
  - if a “chunk type” is not recognized, it is ignored
Critical Chunks

<table>
<thead>
<tr>
<th>Type</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>IHDR</td>
<td>The &quot;true&quot; header is contained in this chunk. This is required and is the first chunk</td>
</tr>
<tr>
<td>PLTE</td>
<td>If the image uses a palette, this chunk contains the table</td>
</tr>
<tr>
<td>IDAT</td>
<td>The image data</td>
</tr>
<tr>
<td>IEND</td>
<td>This chunk marks the end of the file</td>
</tr>
</tbody>
</table>

Ancillary Chunks – Just a few

<table>
<thead>
<tr>
<th>Type</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>bKGB</td>
<td>Default background color</td>
</tr>
<tr>
<td>cHRM</td>
<td>Chromaticity settings (for color correction)</td>
</tr>
<tr>
<td>gAMA</td>
<td>Gamma information</td>
</tr>
<tr>
<td>sBit</td>
<td>Color accuracy</td>
</tr>
<tr>
<td>tIME</td>
<td>Time stamp for the image</td>
</tr>
<tr>
<td>tRNS</td>
<td>Transparency information</td>
</tr>
</tbody>
</table>

Text Ancillary Chunks

- PNGs also allow chunks that store text data
  - these are used to store comments, information about the image, or anything the user needs
  - hidden may be stored here
- Using Base-64 *(or a similar approach)*, these text fields can contain *any* data

Ancillary Text Chunks

<table>
<thead>
<tr>
<th>Type</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>iTXT</td>
<td>Unicode text. This can be compressed</td>
</tr>
<tr>
<td>tEXT</td>
<td>Stores text with a key = value time format. This allows image attributes to be stored similar to how INI values. The format uses a 00 byte rather than an equals</td>
</tr>
<tr>
<td>tXT</td>
<td>Compressed version of tEXT</td>
</tr>
</tbody>
</table>

Fake Ancillary Chunks

- Since unrecognized "chunks" are ignored
  - fake “chunks” can be snuck into a file and used to store hidden data
  - although, these are easy to find
- So, in addition to the least-significant bit method, there many ways to hide data in PNGs

Other Techniques

This “art” is only limited by human imagination
Audio

- Data can be hidden in audio files using *perceptual coding*
- Inject signal into areas that will not be detected by humans
- Human ears are poor – certain “white noise” and frequencies are beyond our abilities

Audio

- Some common techniques:
  - Least-Significant-Bit
  - Phase Coding
  - Echo Data Hiding
- Don’t worry, we will not go any further – this stuff is complex!

Audio

- Data may be destroyed by lossy compression algorithms
  - MP3s, WMAs and MP4s use lossy compression
  - Wave files are lossless
- However, data might be inserted during compression – great for catching pirates!

Video

- Video files are huge in comparison to other file formats
- As a result, there are great targets for hiding data
- Movies typically show 24 frames a second
  - enough to fool the human eye
  - e.g. 1.5 hour movie has 129,600 still pictures

Video Watermarking

- Video is really a combination of images and sound
  - so all techniques used for images and sound also apply
  - there is a few more still

Video Watermarking

- It is commonly used by movie companies to catch pirates
- What companies do
  - on a single frame, they hide a special number/symbol that is unique for *that* copy of the movie
  - when a pirated movie appears on the Internet (or black market), they find that symbol
Other Techniques

- Besides hiding data into the least-significant-bit, data can be hidden in file-specific locations
- Common techniques:
  - Hide the message in unused areas of a file
  - Add the message to the end of the file

Steganalysis

- Steganalysis is concerned only with identifying the existence of a payload
- It does not deal with extracting or reading the contents

Steganalysis

- In the process of locating the existence of a payload it might be possible to identify the software that added it
- Steganalysis looks for anomalies in the file
  - these can concern the file’s format or contents
  - it might also be able to identify the signature of the software that encoded the payload
  - then the same software can be used to read it

Recovering The Data

- Often the generating software is needed – along with passwords, etc....
- Where can you find this information?
- Investigate...
  - hard drives
  - written notes
  - e-mail
  - personal information (pet names, etc...)
  - etc...

Some Possible Anomalies

- Data anomalies
  - visual – Does the data “look” abnormal?
  - statistical – pattern changes in pixels or bits
  - histogram – occurrence of colors, bits, etc... does not match a “normal” view

- Structural anomalies
  - file size – is the file larger than it should be?
  - date/time – internal timestamp doesn’t match files
  - checksum – hash value is abnormal
  - comparison – differences from a copy
Anomalies: File Comparison

- It is incredibly useful if you have a copy of the original file – free from any hidden data
- Comparing the two on a bit-level can reveal differences that will be of interest to investigators
- There are a number of tools such as WinHex, TextPad, etc.

Anomalies: Research

- Comparing numerous files can reveal information on how data is hidden
- Using the same software:
  - encode with the same message
  - examining these and the original can reveal where the data is hidden and how

Anomalies: Bit Planes

- Anomalies in the red-green-blue values can be examined
- For each color, the value of each level will turn on or off bits depending on their value
- For instance, the most significant bit will be 1 if the value is 128 or greater

Bit Planes

- So, often, it is useful to look at an image by only looking at one bit at a time
- These are called *bit-planes* since an image can be viewed, conceptually, as overlaid grids of single bits

Bit Planes

- By looking at a single plane
  - *images will create patterns*
  - this is especially true for increasing/decreasing levels – in particular, gradients
  - steganography can cause obvious breaks in this pattern

Bit Planes

*Single red, green or blue value viewed in 8 different bit-planes*
Bit Planes

- The least-significant-bit is the most chaotic, but some patterns are still visible
- Even least-significant-bit encoding may be detected

Watermarking

- Watermarking is a technique that puts a unremovable message on a target
- Target is often marked...
  - in such a way that its identity or source is known
  - this mark is designed to be unremovable and attempts to altering it will destroy the item
  - it might be obvious or more subtle

Watermarking

- Used for both security and, in the case of software, to indicate the identity of the software used to create it
- Examples:
  - message displayed on shareware software (buying the full version removes it)
  - verification codes on currency
  - barcodes
Steganography vs. Watermarking

- **Steganography:**
  - designed to avoid detection
  - the largest message is desired – the more secret data
- **Watermarking:**
  - designed to avoid distortion or removal
  - usually small hidden or visible message

UV Watermarking

Hidden Codes in Laser Printers

- Some laser printers secretly embed hidden messages on printouts
- Laser barely touches the paper – making little yellow dots
- Dots contain the manufacturer and possible more information

Example: Laser Printer Hidden Code

w2.eff.org/Privacy/printers/docucolor
What Do You Think of This?

- Do you agree with these hidden messages on laser printers?
- Are there Constitutional issues are involved?
- What are the benefits and abuses?