24 – Data Encoding

Goal of Encoding
Simple Ciphers
Other Simple Encoding Techniques
Common Crypto Algorithms
High Entropy Content
Custom Encoding
Custom Decoding Strategies
Goal of Encoding

• Hide configuration information
• Save information covertly before exfiltration
• Store strings and decode just before using to avoid suspicion
• Disguise malware as a legitimate tool
Simple Ciphers

- Caesar shift
  - e.g. rot13: tr [a-mn-zA-MN-Z] [n-za-mN-ZA-M]

- XOR encoding
  - Result of encoding not limited to printable characters.
  - Reversible by double application (just like double rot13)
  - Single-byte XOR encoding common

- Brute-force decoding
  - Try all 256 byte values.
  - Can also brute force by searching for specific xor'd strings (e.g., “This program” for MZ encoded exes)

- Null-preserving xor

- Identify xor decoders in IDA
  - Avoid xor reg, reg
Other Simple Encoding Techniques

- ADD, SUB, ROL, ROR, Multibyte xor, chained/loopback
- Base64:
  - 3 bytes converted to four 64bit (6-bit) characters.
  - Lookup Table (LUT): A...Za...z0123456789+/
  - Often identifiable because of character set and trailing = or == for strings of lengths != 0 mod 3.
  - May find LUT as string in the executable.
  - Indexing may be nonstandard. If you find a 64 character string similar to the normal Base64 LUT, you may have found an alternate encoding.
Common Cryptographic Algorithms

• To solve decryption problem, must find
  i. the encoded string,
  ii. the algorithm, and
  iii. the key.

• May find strings in source code associated with Crypto functions

• May find cryptographic functions in imports table

• IDA FindCrypt2 plug-in looks for magic numbers associated with crypto functions. (Won't catch IDEA or RC4 – which eschew the use of special constants.)

• Krypto ANALyzer (KANAL) PEiD plug-in looks for wide range of constants, Base 64 tables, and numerous crypto imports.
High Entropy Content

- Entropy of a pdf:
  $$- \sum_i p(i) \log(p(i))$$

- Consider entropy of a collection of bytes
  - Range of i: 0..255
  - Uniform density: value is 8
    - $$p(i) = 1/256$$
    - $$\log(p(i)) = -8$$
    - $$-(-8/256) = 8$$
  - All probability clustered around a single point, value is 0
    - At that point, $$p(i)\log(p(i)) = 1*0 = 0$$
    - At other points as $$p(i) \to 0$$, $$p(i)\log(p(i)) \to 0$$
Custom Encoding

- Custom encodings may be simple yet hard to reverse.
- Finding the hard way: trace execution
- Decoding may be
  - Applied to data written as output (to file, or network)
  - Applied to data read as input (from file or network)
- Decoding is often near I/O operations.
- IDA function graph may help identify decoding functions.
Custom Decoding Strategies

- Reprogram the decoding function
- Use the function as it exists in the malware.
  - Self-decoding is economical but
    - Must identify every decryption function and set breakpoints correctly.
    - Must actually decrypt the data of interest.
  - Can direct the decoding by modifying the executable code.
Next Time

- PMA Chapter 14
  (Malware-focused Network Signatures)