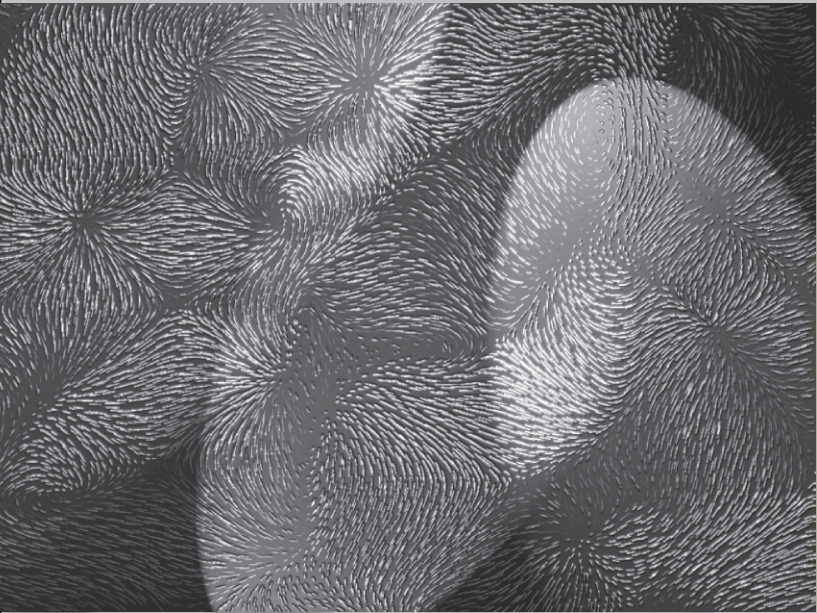


Deep Content Fingerprinting

The Key to Enterprise Content Security

Technical White Paper



Track. Analyze. Protect.

Table of Contents

Introduction	1
What is Deep Content Fingerprinting?	1
Deep Content Fingerprinting Methodology	2
Registering Content Fingerprints	2
RedListing and GreenListing	2
Content Inspection	3
Deep Content Fingerprinting Technology	3
Algorithm	3
Characteristics	6
Performance.	6
Security	7
Use Case Scenarios	7
Summary	8

INTRODUCTION

Code Green Networks™ (CGN) offers content security solutions that help enterprises mitigate the risk of unauthorized disclosure of confidential information. While traditional security products protect against hackers, viruses, spam and other commonly known threats initiated from the outside of a network, few security solutions are available to protect against information leaks initiated from the inside. Without adequate content security, individuals are able to send confidential information, either inadvertently or maliciously, to external parties using common Internet protocols, such as e-mail, webmail, File Transfer Protocol (FTP) and Instant Messaging (IM). Additionally, employees can transfer data onto removable media devices, such as iPods or USB flash drives and physically remove it from the enterprise. Such unauthorized disclosures of confidential information can result in loss of revenue, financial penalties and irreparable damage to a corporation's image, brand and customer loyalty.

The heart of the Code Green Networks Content Security Solution is a unique and sophisticated technique called Deep Content Fingerprinting™ (DCF) which is extremely efficient and accurate at detecting transmission of confidential electronic content on the network and on client PC's. This white paper explains fingerprinting and how it is used to safeguard intellectual property and confidential information.

First, we provide an overview of how content is fingerprinted and how the stored fingerprints are then used to detect and prevent unauthorized disclosure of confidential information. We then examine the unique and tunable parameters of the Deep Content Fingerprinting methodology and contrast it with less powerful, alternative approaches. Finally, we will provide examples of how confidential information is tracked, analyzed and protected by the Code Green Networks Content Security Solution.

WHAT IS DEEP CONTENT FINGERPRINTING?

Authorized use of confidential content is part of normal business activity. Employees manipulate and transfer digital content when they quote from email, cut and paste from internal documents and reports, and collaborate on projects. Digital content is easily manipulated, copied, printed and/or transmitted across multiple channels using multiple protocols. Malicious or negligent employees are creative; instead of simply sending content across the network they download information from their personal computer to an iPod or USB Flash drive.

To protect confidential information, effective identity management and access controls are necessary, but not sufficient. They must be augmented with an inspection and enforcement capability that can actually monitor network traffic, detect unauthorized attempts to transfer confidential content and intercept them. To do this, the content security solution must capture and store a representative signature of the content to be protected. It then compares this signature, at wireline speeds, to content being transmitted on the network. If it detects a match, it can then invoke the appropriate pre-defined security policy such as logging, quarantining and/or blocking. This methodology must scale to the enterprise level, where billions of bytes of confidential content are flowing through the network.

Traditionally, digital documents have been compared using hashes of entire files. Using this method is simple and sufficient for reliably detecting exact matches that may be sent outside of a company's secure intranet; however, detecting partial copies or near matches is far more complex and requires a new, unique and robust technology. The digital workflow of today's enterprises requires a content fingerprinting methodology that reliably and accurately detects derivatives of a confidential document in various and multiple arbitrary file locations. This methodology, termed Deep Content Fingerprinting, is a key differentiator that separates the Code Green Networks Content Security Solution from other solutions.

Content Inspection

Once content has been fingerprinted and registered, the Content Inspection engine of the CI Appliance is enabled to compare transmitted data against the RedList and GreenList fingerprints. The Content Inspection process is illustrated in Figure 2. During content inspection, the Content Inspection engine examines the stream of information flowing through the network gateway. It identifies and assembles content for a variety of protocols, including SMTP, HTTP (to include blogs and web-based mail) and FTP. The data is fingerprinted “on the fly” using the same Deep Content Fingerprinting algorithm used in the CI Appliance Content Registration engine.

The fingerprint of the transmitted data is then compared to the content fingerprints stored on the RedLists and GreenLists. If a match is detected, the appropriate user-defined security policy and workflow procedures are initiated.

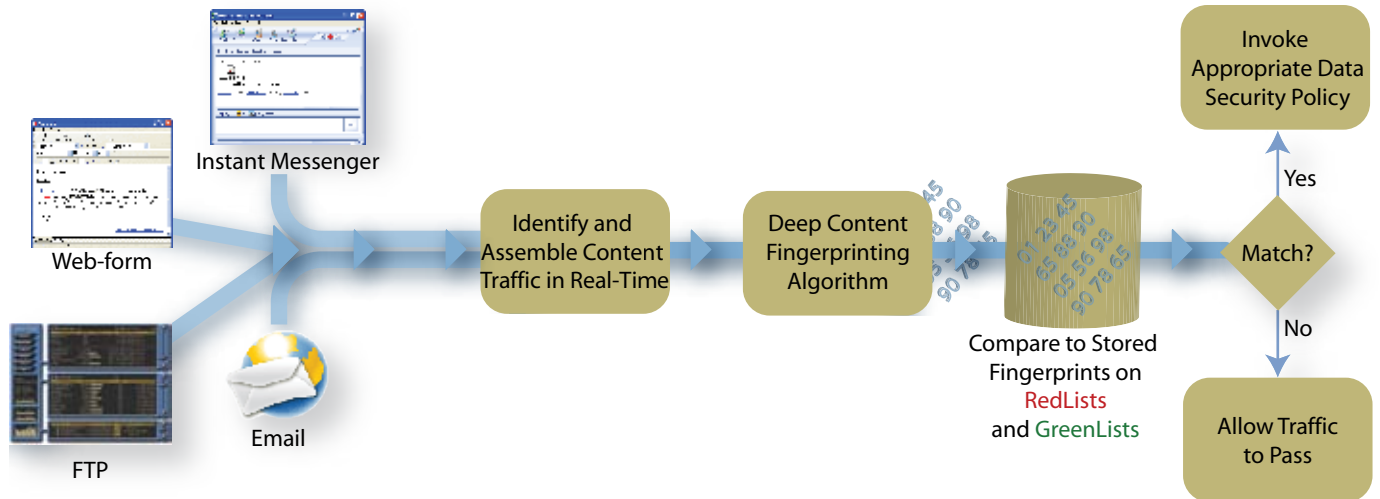


Figure 2: Content Inspection – Network traffic is identified, fingerprinted and compared to RedList and GreenList fingerprints stored on the CI Appliance. If a match is detected, the appropriate security policy and workflow procedure is invoked.

For example, the transmission may be logged and permitted, or the transmission may be blocked and an administrator notified of the attempted transmission.

Besides fingerprinting and inspecting un-encrypted data, the CGN solution can inspect encrypted data. Encrypted data is handled according to security policies that govern users, destinations and protocols. The manner of detection identifies encrypted streams (such as SSL transactions), or encrypted files on normally un-encrypted transmissions (such as sending a password-protected .zip file over email).

DEEP CONTENT FINGERPRINTING TECHNOLOGY

The Content Registration and Content Inspection engines both use the same unique Deep Content Fingerprinting technology. Besides being able to accurately detect confidential information and any derivatives of confidential information, the Code Green Networks Deep Content Fingerprinting technology is also efficient, robust and tunable.

Algorithm

Deep Content Fingerprinting involves two primary steps: extracting meaningful content and producing a set of fingerprint hashes from the meaningful content.

First, the data to be fingerprinted is pre-processed to extract all meaningful content. This includes certain metadata items (such as document properties) and the body text of documents. Today, the CI Appliance supports over 370 file

types. The output of this step is a Unicode string, representing unformatted content from the original file or input string.

For a simple example, consider the following sentence found as text inside of an arbitrary document:

```
Let's create a fingerprint.
```

Processed to normalize capitalization and remove whitespace and punctuation, it is now:

```
letscreateafingerprint
```

After meaningful content has been extracted, the Deep Content Fingerprinting technique looks at fixed-length segments of content, at one byte intervals throughout the file. These segments of content are known as “k-grams” and have a fixed length throughout the process. This example will use a k-gram length of five (5).

Looking at every five-byte k-gram throughout the file, and sliding along the length of the file at one-byte intervals, yields the following k-grams:

```
letsc  etscr  tscre  screa  creat  reate  eatea  ateaf  teafi
eafin  afing  finge  inger  ngerp  gerpr  erpri  rprin  print
```

This process yields almost as many k-grams as there were bytes in the original content.

Each k-gram is now sent through a one-way hashing algorithm. In Deep Content Fingerprinting, this algorithm takes advantage of previous hashes in order to increase processing speed. These hashes are 64 bits long however, for simplicity, this example will use smaller numbers. The following hashes result from processing the k-grams above:

```
82 17 58 92 01 37 95 12 06
13 77 90 39 54 51 72 23 40
```

To reduce the amount of hash data, the algorithm now selectively chooses hashes in a process termed “winnowing”. It continues by looking at “windows”, containing some number of k-grams, sliding at one k-gram intervals. This method is conceptually similar to the processing done to the original content at one-byte intervals.

This step will use a window size of five (5), producing the following sets of existing hashes:

```
(82 17 58 92 01) (17 58 92 01 37) (58 92 01 37 95)
(92 01 37 95 12) (01 37 95 12 06) (37 95 12 06 13)
(95 12 06 13 77) (12 06 13 77 90) (06 13 77 90 39)
(13 77 90 39 54) (77 90 39 54 51) (90 39 54 51 72)
(39 54 51 72 23) (54 51 72 23 40)
```

Within these windows, the algorithm selects a single hash to represent the group. In this case, it selects the lowest number in each group (represented in bold above). Choosing the lowest number in each group results in the following set of representative hashes, with duplicated hashes removed:

```
01 06 13 39 23
```

Any of these above hashes will match a hash taken from any inspected document that has a matching window anywhere in the document content. We have now created a fingerprint. Deep Content Fingerprinting achieves a high degree of compression by reducing groups of hashes into one representative set of final hashes, termed a fingerprint.

The algorithm has produced a set of hashes that will always represent any of the windows created from the original content. Inspected content must be at least the size of a k-gram (for this example k-gram = 5) in order to trigger a match.

In practice, the algorithm yields a “noise” length number (anything less than a k-gram in length), under which nothing is matched, to weed out common phrases and reduce false positives. It also has a “guaranteed match” length number, in which content of a particular length will always produce a matching hash in the fingerprint database. A match is guaranteed for all phrases of length:

$$\text{Guaranteed Match Length} = [\text{k-gram size}] + [\text{window length}] - 1$$

In this example, this means that any phrase of nine (9) characters or more (5+5-1) would guarantee detection. Any content between the noise number and guaranteed match number will produce a match based upon a normal probability curve.

The size of the fingerprint is quite small compared to the original content, providing the CI Appliance the headroom to process and compare large amounts of data in real-time. Even though the fingerprint size is small, it produces guaranteed matches.

Example: Consider looking for an instance of the word “fingerprint” from the content that we registered above.

The examined text is:

My fingerprint, too.

Using the Deep Content Fingerprinting technology, the text is normalized and creates the k-grams:

myfin yfing finge inger ngerp gerpr erpri rprin
print rintt intto nttoo

The hashes are then created (noting that exact k-gram matches will produce the same hash, represented in bold):

81 27 **90 39 54 51 72 23**
40 65 88 94

Now, these again are separated into windows of five k-grams each (smallest number represented in bold):

(81 **27** 90 39 54) (**27** 90 39 54 51) (90 **39** 54 51 72)
(39 54 51 72 **23**) (54 51 72 **23** 40) (51 72 **23** 40 65)
(72 **23** 40 65 88) (**23** 40 65 88 94)

The resultant fingerprint contains hashes:

27 39 23

The 39 and 23 hashes are duplicates of hashes found in the original content and this new text would trigger a match.

In practice, since the hashes themselves are 64 bits long, and widely distributed throughout the space, the chances of finding two dissimilar pieces of content that produce the same hash is very small. This means that any match found by the DCF algorithm definitely corresponds to matching content.

For further information on the fingerprinting algorithm, see the paper by Schleimer, Wilkerson and Aiken (2003)¹.

¹ Schleimer, Saul; Wilkerson, Daniel S., and Alex Aiken. *Winnowing: Local Algorithms for Document Fingerprinting*. SIGMOD 2003, June 9-12, 2003, San Diego, CA.

Characteristics

The Deep Content Fingerprinting algorithm provides several key advantages, each of which ties directly to business needs:

Accurate and Precise Detection – reduce false alarms and protect content

CGN Deep Content Fingerprinting provides an extremely high degree of accuracy in detecting confidential content. For instance, a particular paragraph of text removed from a corporate report and pasted into an arbitrary Microsoft Word document, will generate a positive detection in 100% of cases, regardless of placement in the document, or other document content.

Non-significant content Insensitivity – increase performance and protect content

CGN Deep Content Fingerprinting is insensitive to trivial changes to a content object such as the amount of white space added or removed from a content object. This improves the performance and accuracy of the CI Appliance.

Noise Insensitivity - reduce false alarms that can affect business continuity

CGN Deep Content Fingerprinting is also insensitive to noise in a content object, such as small phrases that would produce many false positives. The algorithm produces indexes that completely ignore matches of insignificant size (less than the k-gram size described above), while still maintaining the data's significance to a larger match. This reduces the number of false alarms and ensures business processes continue without interruption.

Positional Independence - to detect confidential content regardless of its location

CGN Deep Content Fingerprinting is insensitive to the changes in the positional occurrence of a piece of confidential content within a larger content object. For example, an employee may cut a small section of a RedListed document and then paste the information into a non-confidential document before transmitting across the corporate network. During inspection, the CI Appliance will detect the derived confidential content that was previously registered and fingerprinted.

Language Independence - to span all natural and computer languages for global data protection

Unicode representations of data are used consistently within the CGN Deep Content Fingerprinting algorithm, ensuring support for any standard character set and natural language. There are no language limitations and data from worldwide locations can be simultaneously scanned for matches.

Fingerprinting also works remarkably well to detect similar portions of computer programming or scripting languages, independent of knowledge of the particular language structure. This has obvious uses for protecting source code repositories, but can also be used to detect embedded scripting in other documents, such as spreadsheets and CAD drawings.

Performance

Monitoring outbound enterprise network traffic requires real-time analysis of immense volumes of confidential data. Deep Content Fingerprinting is a high-throughput technique that enables real-time network traffic inspection and enforcement. The CI Appliance is a high-availability system providing automatic failover and redundant storage capabilities. Deep Content Fingerprinting is not only accurate and precise at detecting confidential data; it is also robust, scalable and high-performing.

Space Efficient

CGN Deep Content Fingerprinting performs space-efficient fingerprint encoding of raw confidential data, producing an index which is considerably smaller in size than the original content set. 1 TB of confidential source information produces approximately 5 GB of fingerprints, while maintaining 100% detection capability on matching content.

High Performance

The index can be queried fast enough to enable real-time content monitoring, inspection and enforcement speeds of hundreds of Mbps on a single CI Appliance.

Security

Fingerprints are an accurate and unique numerical representation of the original confidential content and do not include the original data. The creation of the fingerprint representation is a “one-way” transformation – the representation cannot be “reverse engineered” to reconstruct the original source information. The CI Appliance does not store the original confidential data, thus reducing fingerprint storage requirements and removing the security risk of having all confidential data stored in one repository.

USE CASE SCENARIOS

In this section, we present two common security scenarios and describe how Deep Content Fingerprinting would enable the threat to be detected and mitigated.

WEB-LOG (BLOG)

Threat: The use of blogs has increased over the past several years and is becoming part of popular culture. ABC News reports that blogs are created at the rate of almost one every second. Blogs are easy to implement, use and maintain, thus they are popular with the general public. Many users feel little inhibition about posting detailed information about their personal and work lives to a public blog. With the increased use of blogs comes the threat of trusted employees intentionally or unintentionally posting confidential corporate information on an external blog.

Solution: The Code Green Networks CI Appliance can automatically detect and (if desired) prevent the posting of confidential information to blogs. To prevent this threat, the Content Authority registers critical confidential information for RedList content fingerprinting. For example, one can fingerprint original source code or trade secrets. When the trusted insider attempts to post the source code to a blog outside the corporate network (even if they just try to post a few lines of code) the CI Appliance detects the attempted transmission by using the Deep Content Fingerprinting technology and applies the appropriate security policy. In this case, the data security policy alerts the Content Authority to the attempted transmission, logs the activity and if desired, blocks the transmission.

INTERNAL EMAIL INCIDENT

Threat: Not all information security breaches are intentional. One of the greatest threats is loss of confidential information accidentally through corporate email. Most emails are created for specific employees and contain confidential or private information. During the lifecycle of an email the data contained in the email or attachments can be manipulated, cut and pasted, forwarded or stored in a personal folder. In one example, a continuing email discussion concerning quarterly earnings eventually results in the inclusion of a company outsider and the email still contains original confidential information. The internal employee includes the outsider to the “To:” box of the email and the confidential information has now been disclosed outside of the corporate network. The email contains the original information, but the email programs have also introduced several ‘>’ characters on a per line basis indicating increasing quote levels. The entire original email is also not included, as the mail programs have stripped the original headers (and any attached v-cards and extraneous mime attachments) and only quoted the body text of the email. Since there were no content security solutions in place, the confidential information was permitted to leave the network and quarterly earnings were released earlier than SEC allowance.

Solution: The CGN Deep Content Fingerprinting detection algorithm would reliably flag this email and operate at speeds preventing the disclosure of the confidential information. Non-significant content (the ‘>’ characters, along with any extraneous space) is ignored in the context of the relevant text. Despite the additional email content surrounding the confidential portion of the transmission, the Deep Content Fingerprinting technology will accurately detect the presence of any meaningful portion of the original email.

SUMMARY

Code Green Networks Deep Content Fingerprinting differs from other content detection methods in that it is extremely accurate and efficient at inspecting large volumes of data on the network, while ensuring confidential content transmitted on the network is detected. Since content fingerprints are a unique and accurate representation of the original content, they can later be used to identify confidential content even if it has been cut and pasted into another document, compressed or modified. For example, if an employee cut and pasted a section of C++ source code and attempted to email the code outside of the network, the CI Appliance would detect the derivative work. Some of the key characteristics of CGN Deep Content Fingerprinting are:

- Reliability in detecting whole or partial sections of confidential data
- Insensitivity to insignificant data
- Resistance to noise
- Non-reliance on any particular position of detected content within a larger document
- Language independence

The CGN patent-pending Deep Content Fingerprinting technology accurately identifies registered confidential content. Uniquely tunable parameters reduce the incidence of false positives and false negatives and storage-efficient encoding scheme allows 1 TB of source confidential data to be represented by a 5 GB fingerprint database. By applying workflow procedures, logging network and client activity, and preventing data leaks before they occur in real-time at the source, enterprises have peace of mind knowing their data is secure and data forensic details are always available.

Through the use of proprietary technology and dedicated hardware, the CGN Content Security Solution provides management with the visibility and control necessary to track, analyze and protect a company's confidential data.

For additional information,
please contact Code Green Networks at:

3975 Freedom Circle, Suite 900
Santa Clara, CA 95054
408.213.2300
408.213.2301 (fax)
info@codegreennetworks.com
www.codegreennetworks.com



www.altaware.com
sales@altaware.com
(866) 833-4070

Your Code Green Networks Reseller

The Code Green Networks name, logo and products are trademarks of Code Green Networks, Inc. All other company and product names, brands or service names are trademarks of their respective owners.

Notice: This document is for information purposes only and does not set forth any warranty, expressed or implied, concerning product, product feature or service offered or to be offered by Code Green Networks Inc. All information is subject to change.