Schippers IT, Gilze

Printer Steganography

Reverse Engineering the Machine Identification Code

Author:	Peter Buck
Student number:	410147
Educational Program:	Forensics
Educational Institution:	Saxion University of Applied Sciences
Date:	13 June, 2018
Supervisor:	R. van Enschot
Examiner:	R. Brinkman

Preface

Before starting this research, I was unaware of the presence of yellow dots on printed paper. I find it remarkable that something so common on an everyday object could be this unknown to the public. Learning and becoming a master on a subject few people even know about feels fantastic. Even the majority of law enforcers have only heard rumors that it exists. The research itself was challenging because of the monotonous coloring and reviewing of the patterns. However, the moments when I found something useful were rewarding enough to continue.

This dissertation was written from February to June 2018 at the request of Schippers IT. Last year, they received two independent requests to find the origin of a document. They were aware of the dots but did not know how to use this knowledge in a forensic investigation.

I would like to thank everyone at Schippers IT for helping and supporting me through the process. I also wish to thank everyone whom I spoke to on the subject. Some asked insightful questions, while others just listened. Both were helpful in clarifying my understanding of the subject. Special thanks go out to Seth Schoen from the Electric Frontier Foundation (EFF). He sent me an additional dataset per post while I was struggling to get samples myself.

I hope you learn something from my research.

Peter Buck

Gilze, June 13, 2018

Abstract of the dissertation

Reverse Engineering the Machine Identification Code

By

Peter Buck

Forensic research

Saxion University of Applied Sciences, 2018

R. van Enschot, supervisor

This study analyzed the Machine Identification Code (MIC) used by various Color Laser Printers that this document describes. Previous studies showed that the Xerox pattern contains the time, date and serial number. The goals of this research were: 1) to understand various patterns and 2) to explore the ability to use the MIC during forensic investigations. This method can support investigations on fraud, stalking, theft and other offences. The research concludes that the Machine Identification code can be used to identify specific printers.

Two methods were used to analyze the patterns. The first one used a previously gathered and scanned dataset provided by the EFF and an additional self-scanned dataset. All of the different patterns that were available were reviewed to find out if time is part of each pattern. The patterns of both the “Reverse-L” and the “corner” were analyzed in depth. Other patterns did not have enough samples to discover anything else. The “Reverse-L” pattern, mostly used by HP, includes the manufacturer and specific printer model. At the conclusion of this research, it is unknown whether the pattern is hard coded or changeable. The second method consisted of changing a number of variables on an HP Color LaserJet CP1515n. None of the yellow dots changed position.

Recommendations related to further research and the gathering of a larger dataset. A new dataset should include more printer information, but also metadata including the country and the store that sold the printer. Another option is to create a script which automatically extracts the pattern from a document. Lastly, the in-depth research on the HP CP1515n can be repeated with other printers. These printers should preferably have older unsigned firmware.

Contents

1. Introduction

Last year, a former NSA employee leaked a highly sensitive document to a newspaper. She had printed the document in her office and gave the physical copy to *The Intercept*. They published the entire document on their website for people to read. Partly because of this action, the NSA was able to trace the document back to the leaker. They were able to do so due to a code that the printer that was used printed on the paper. [1] This code gives the printer serial number and the time and date that the documents was printed. To the general public, this is known as the Machine Identification Code (MIC).

The MIC has been around for a long time, presumably since the 80s. [2] Some governments were afraid that counterfeit money would be printed with laser printers. Thus, they wanted an anti-counterfeit measure. The printer manufacturers obliged by adding a mechanism to their printers that adds a secret code to every document. This method remained unnoticed and undecipherable for twenty years until the Electronic Frontier Foundation deciphered the code for all Xerox color laser printers. It contains time and date of printing and the serial number of the used printer. [3] The Canon pattern includes the country and dealer that supplied the specific printer. [4]

According to a Freedom of information appeal (FOIA), analysis of the MIC only happens during investigations of counterfeit money. The Japanese Business Machine Maker Association sells software to decode the dots. They sell it to specific governments. [5]

Linking documents to printers is possible with other techniques. These methods consist of statistically analyzing noise on the document. This noise can lead to the identification of an individual printer. [6] [7] A drawback of this method is that the printer that was used must be available to compare its output to the printed documents. Another drawback is that the noise changes if the toner cartridge is replaced.

In recent years, the few efforts that have been made to analyze the MIC gave few results. Other printer manufacturers use more difficult encoding methods than Xerox. Knowing the encoding method can help with the identification of the person who printed the document. The identification of a person can be beneficial in investigations concerning fraud, or other illegal cases. This research provides an in-depth analysis of the MIC used by printers other than Xerox. It attempts to provide an answer to the central question: How does the machine identification code work for color laser printers other than those manufactured by Xerox.

2. Theoretical Background

This chapter describes the information needed to understand how and why specific steps are taken. It also adds some general information.

2.1 Laser Printers

Laser printers use static electricity, light, photosensitive material and toner to print documents. A LED or laser projects the text or image onto the printer's positively charged cylindrical drum which is covered in a photosensitive material. The light causes the material to lose its charge. The drum rolls through the

toner which sticks to the non-charged material. The drum is then heated up and is rolled over the paper, leaving behind toner. A color laser printer does this multiple times with different colors. [8]

Dots Per Inch (DPI) define the quality of the printout. In computer terms, this means the number of pixels printed on an inch. [9] The maximum DPI of a printer defines the sharpness of the MIC.

Color laser printers print a code onto every piece of paper multiple times. The most known way is by using pale yellow dots. These dots are invisible to the naked eye. This is because of two factors; The dots are very small and pale yellow does not stand out on white paper. These dots are on the entire document in a repeating pattern. Not every color laser printer prints a MIC. The EFF published a list of the laser printers that do. [10]

A freedom of information appeal to the US government in 2012 by Theo Karantsalis revealed that the following printer manufacturers use a MIC: Brother, Casio, Hewlett-Packard, Konica, Minolta, Ricoh, Sharp, and Xerox. [11]

2.2 Information type

Laser printers possess a significant amount of data. This data concerns information about the printer and information about the user. Most of the data cannot be used forensically. Table 1 shows the information that might be usable for the identification of a suspect.

Table 1: Possible printer information

Server name	Hostname
Time zone	Date
Product Name	Time
Serial Number	Mac Address
Standard language	Ip address
Gateway address	DNS server address
Domain name	Server name
Print or copied	Administrator name
Manufacturer	Model

As of March 2018, only the Xerox MIC is decipherable by the general public. [3] The code used by Xerox provides the time, date and serial number. Figure 1 shows the explanation of the Xerox MIC.

The information given by the Xerox MIC indicates how specific data can be hidden in plain view. The parity ensures that it is not possible to misinterpret the code.

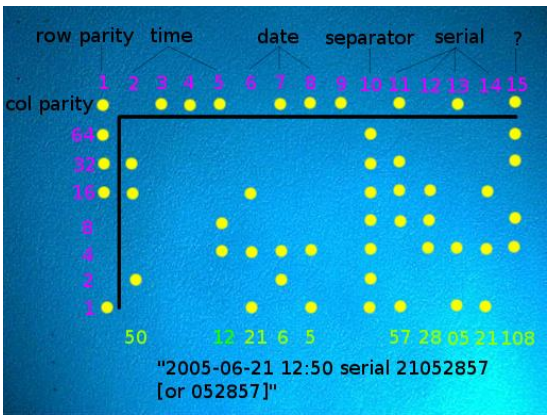

Figure 1: MIC Xerox DocuColor [3]

2.3 Visibility

There are two different methods that can be used to make the MIC visible. These methods are physical and digital. With the physical method, the paper is illuminated with a blue light. This light makes the yellow dots appear black. These dots can then be observed using a magnifying glass. The physical method is useful for an immediate analysis of a document because it can be used at a crime scene. However, it takes too much time for this research project.

The second method is digital. Scanning a document with specific settings can make the yellow dots visible. Essential settings are DPI, sharpness, darkness and contrast. Furthermore, the scanner must use limited compression. Some imaging e-tools can change the color channels of an image. Deleting both the green and red channels results in a grayscale image of the blue channel. This grayscale image then shows the dots. The contrast and lightness can be changed to make the dots appear more visible.

2.4 Pattern alignment

The complication of this research begins with pattern alignment. Depending on the size and variety of the pattern, it can be hard to define where the pattern begins and where it ends.

The most straightforward decipherable pattern is the isolated pattern alignment. These patterns are isolated and because of this, the only thing that needs to be defined is top and bottom of the pattern. This type of alignment is shown in Figure 2.

A moderately tricky pattern is the Grid pattern alignment. This pattern repeats itself in a grid. This pattern is moderately hard to decipher because the start and end corners are unknown. Furthermore, it is unknown whether the pattern should be read from left to right, right to left, top to bottom or bottom to top. This type of alignment is shown in Figure 3.

The most challenging pattern type is the diagonal pattern alignment. The pattern repeats itself diagonally. This pattern type is the most difficult to decipher because nothing is known. For every begin point of the pattern, the pattern changes. This type of alignment is shown in Figure 4.

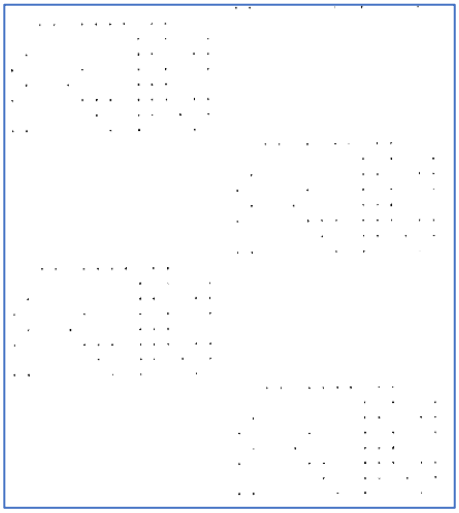

Figure 2: Isolated pattern alignment

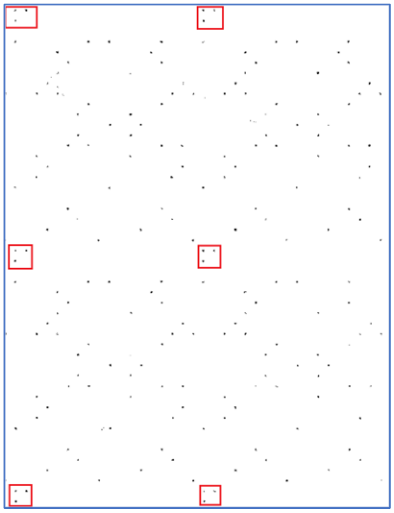

Figure 3: Grid pattern alignment

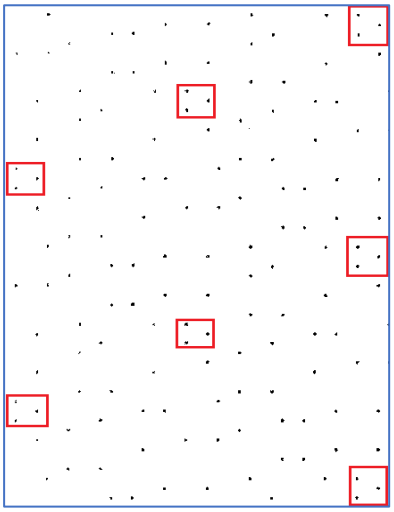

Figure 4: Diagonal pattern alignment

2.5 Machine classification

In November 2013, a research group proposed a new method for the automatic authentication of color laser print-outs using Machine Identification Dots. They used the dataset gathered by the Electronic Frontier Foundation in 2005. [12] This dataset is a collection of print-outs of many different color laser printers and is also used for this research. The proposed method is to categorize the printer types based on the distance between two repeating patterns. The distance is calculated in both the horizontal and vertical directions. Since every printer manufacturer uses their own pattern type, all of their printers belong to a specific pattern separation list. This method makes it possible to quickly analyze the brand of printer used. [13]

2.6 Grid encoding

All patterns are printed on a rectangular grid. Encoding of a grid can use various mathematical formulas. Since a grid is an image, it needs to be translated to numbers or letters. The known Xerox printers use a binary system. Every column contains a maximum of seven dots. A dot can mean: 1, 2, 4, 8, 16, 32, or 64 adding up to a maximum total of 127. This method can be used by other manufacturers as well. The number one is the only indication whether a number is even or odd. Statistically speaking, numbers have a 50% chance of being odd, or even. The easiest way to see whether a binary system is used is to check whether the most filled row or column is statistically 50% filled with dots.

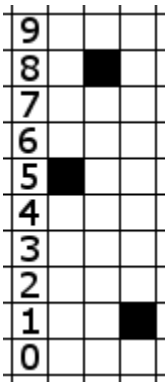

Figure 5: decimal grid encoding

A similar method uses a different number system, for instance hexadecimal, octal or decimal. In this instance, a multi-digit number is written down in multiple columns or rows. Figure 5 shows decimal grid encoding. The number written down can be either 581 or 185. This method can be identified if columns only contain a single dot.

Another approach is not column or row oriented. Instead, it is based on boxes, squares or areas. Every grid can be divided into a finite number of areas, and every area has a specific function.

Something to keep in mind is the notation of characters. For example, an "A" can be 0041 (Unicode), 10 (Morse), 1000001 (binary) or simply A (hex).

3. Research Strategy

This aim of this research is to reverse engineer the Machine Identification Code (MIC) for multiple color laser printer brands. This will be done by analyzing multiple test sheets printed by multiple printers. The entire research approach is based on trial and error until the encoding system is known.

The main experiment has two components. The first components is to repeatedly print a document after changing a single variable (e.g., time, date, serial number, content). This method has been explicitly chosen to make decoding easier. If a single row or column changes along with the variable, it might indicate a connection between the variable and a row or column. This method has been chosen because it generates a clear overview of the variables per printer brand.

The second component is to analyze the variables per column or per row. This analysis will compare the same columns or rows of different test sheets in combination with the changed variable.

The previous method used by EFF used test sheets sent in by people. This provided the community with a dataset of numerous pages printed with different printers. It is complicated to use this dataset for pattern decoding. Each pattern printed by similar printers changes due to the number of variables. Therefore, the available dataset can only be used for general pattern analysis.

Two of the most critical variables are time and date. These are the only variables that change over time without any intervention. If they are present in the pattern, the pattern changes continuously. Therefore, every pattern will be analyzed for time and date changes.

The hypothesis behind this research is that every printer prints a unique pattern. This pattern contains the serial number of the used printer and maybe other variables. Every pattern is decodable if sufficient time and resources are available.

Each experiment results in a statement. This statement is a short description of the result of each pattern analysis. Until the pattern encoding method is known, the statement remains a hypothesis. However, it can support further research in deciphering the MIC.

4. Methods and materials

The first step of the research was to compare documents of the known dataset. The machine classification project categorized all patterns into groups. These groups were reviewed individually. The dots were made visible with Gimp 2.8. First, the red and green channels were turned off. Then a grid was placed on top of the pattern, and each block was colored. This process creates a schematic overview of the pattern, which can be compared to similar patterns.

Multiple samples of every group were taken to find recurring patterns. These were then placed on top of each other, creating an overview of the entire pattern. The patterns were then described based on their visible recurring components. This made it possible to look at a pattern and to immediately identify the type of pattern.

Different documents printed by the same printer were then compared. This indicates whether time was part of the pattern since every document was printed at a different time.

The HP LaserJet pattern was explored with the use of PRET (Printer Exploitation Toolkit). [14] This toolkit made it possible to change variables within an HP Color LaserJet CP1515n. The variables include serial number, formatter number, network or usb, IPv4 address and printer name. For every changed variable, a test document was printed. The document was then reviewed to see if the pattern had changed.

An additional data set was made with additional documents provided by the EFF and new documents. These papers were scanned with a Xerox WorkCentre 7132 equipped with the flatbed scanner. The settings used were:

a. General Settings
    i. Scan color: Color
    ii. 1 or 2-sided: 1-sided
    iii. Original type: Photo and text
    iv. File format: JPEG
b. Print Quality
    i. Photo: Photo improvement on
    ii. Screen Options: Darker +3, Sharpness +1
    iii. Print suppression: Contrast -2, no print suppression
    iv. Shadow suppression: No suppression
c. Layout:
    i. Scan resolution: 600 dpi
    ii. Original format: Automatic registration
    iii. Remove borders: top/bottom 2mm, left/right 2mm
    iv. Decrease increase: 100%
d. Scan layout
    i. Screen compression: -2

The entire known dataset, including some patterns has been made available for online access. [15]

5. Results

The results are categorized into two parts. One being some general observations that have not been mentioned in other articles, the second part is pattern recognition and details. Multiple manufacturers can use the same pattern. Therefore, these results are arranged per pattern type.

5.1 General Observations

As stated by the EFF, not all laser printers use the MIC [10]. The more notable Xerox Phaser is thought not to include these dots. After a thorough check of multiple documents produced by this printer, it turns out that the dots only appear on colored areas as seen in Figure 6 (a) and (b). Blue areas provide the clearest dots.

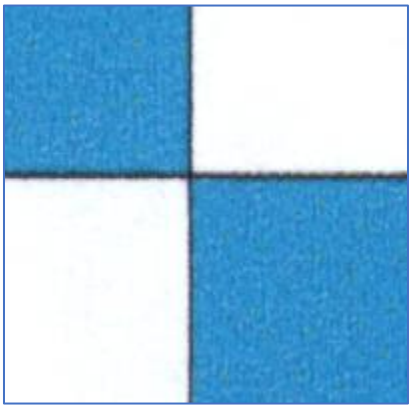

Figure 6a: Colored areas

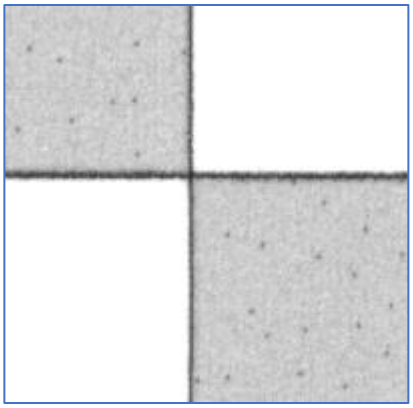

Figure 6b: yellow channel colored areas

Some patterns have not been observed previously because of color intensity. Figure 7 shows a 0.17mm x 0.17mm dot printed by an HP LaserJet M476dw (a) and a Canon imageRUNNER (b). They were both scanned using the same scanner and the same settings. As scanners improve, it might be possible for the dots to decrease in yellow intensity.

Figure 7a: HP LaserJet dot intensity

Figure 7b: Canon imageRUNNER dot intensity

5.2 Xerox

The Xerox pattern is an isolated pattern. The decryption method is mostly known. However, a comparison between a printed document and a copied document resulted in a new observation. Figure 8 shows two patterns from the same Xerox printer printed at roughly the same time, (a) is a printed document, whereas (b) is a copied document.

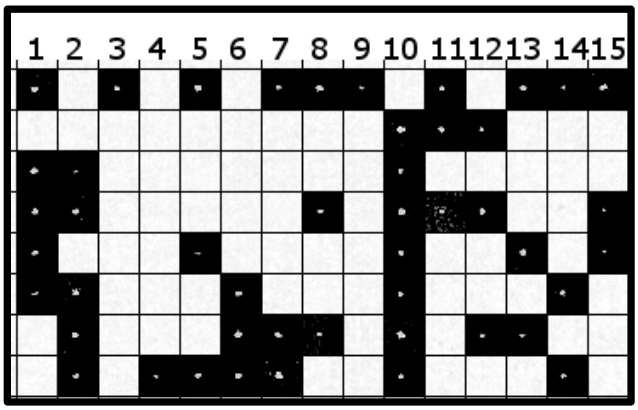

Figure 8a: Printed Xerox Document

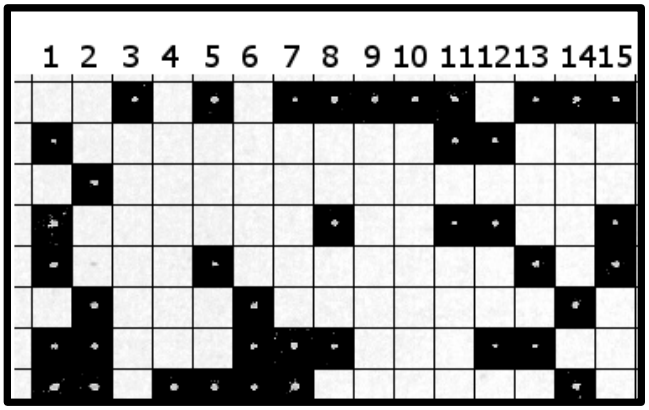

Figure 8b: Copied Xerox Document

Column 10 was previously thought to be a separator between the serial number and the date and time of printing. It is instead a line which defines whether a document has been printed or copied.

Statement: If column 10 of a Xerox pattern is filled, the document is printed. If column 10 is empty, except the parity bit, the document is copied.

Decoding columns 11 – 15 in figure 8 using the binary method of EFF results in 5 numbers. Starting at column 15, these numbers are 24 05 10 82 80.

A hard drive of a Xerox Workcentre 7132 was examined. The drive contains a plaintext file that includes the system information about the printer. It contains information not included on the configuration page. The serial number presented in the pattern in Figure 8 is equal to the SerialNumber 108280 at CheckProduct in Figure 9. Even though it is just the same serial number minus the first three numbers, it could be the variable that the pattern references. The ProductNumber is equal to columns 15 and 14. These columns were previously unknown.

Statement: Columns 14 and 15 of the xerox pattern indicate the product number.

```
 --- System Infomation ---
iotModel:        0
manufacturer:    Xerox
manufacturerURI:http://www.xerox.com/
ProductName:     WorkCentre 7132
PrinterName:     WorkCentre 7132
Location:
ContactPerson:
OperatorInfo:
Description:
ProductID:       D38
StatProduct
              SerialNumber:  108280
              ProductNumber: 2405
              ProductCode:   AYN
              ProductType:   15
              ProductTerritoryInfo:   3
              Country:     826
              ProductPaperSizeGroup:  3
CheckProduct
              SerialNumber:  108280
              ProductNumber: 2405
ExposingProductInfo
              SerialNumber : 3311082808
              ProductName  : WorkCentre 7132
```

Figure 9: System information Xerox Workcentre 7132

5.3 Reverse-L

Multiple manufacturers including HP, Kyocera Lexmark, and Ricoh use this pattern. It is recognizable by a Reverse-L in the top left corner.

i) Main Pattern

Thirty-three different patterns of different HP LaserJets were placed on top of each other to distinguish a recurring pattern in Figure 10. This new image provides a clear overview of the blocks used. In the middle of the image is a white vertical stripe. This stripe indicates that the pattern is divided into two parts, left and right. The sixteenth row also contains a white horizontal line. This stripe divides the pattern into top and bottom.

Furthermore, the space adjacent to dots is always white, resembling a chess-like board. It is possible to recognize this pattern by the reverse L in the top-left corner. This L is an exception to the chess pattern.

Figure 11 shows a sample pattern. Every horizontal row has even parity, except for the topmost. An additional rule is that every row contains two dots, one on the left and one on the right, with the exceptions of row 1, 6, 11 and 16. These rows include 0, 2, 4, 6 or 8 dots.

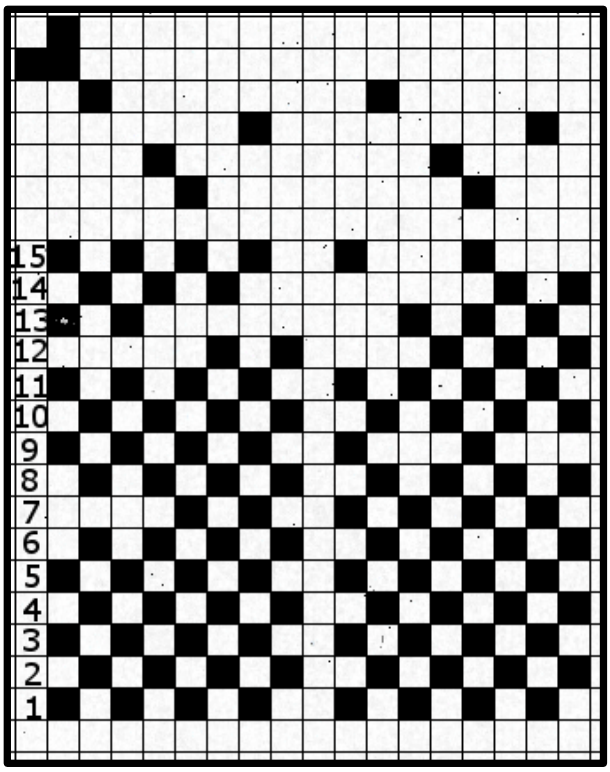

Figure 10: 33 HP LaserJet patterns superimposed

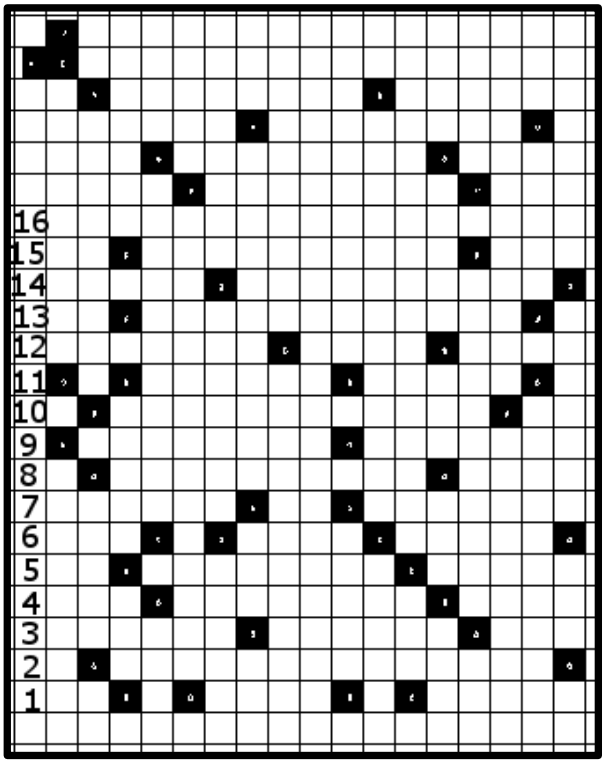

Figure 11: Sample HP LaserJet

ii) Parity rows

The lower part of the Reverse-L pattern consists of two 15x8 grids. As stated in the main pattern paragraph, this pattern has three rows which differ from the rest. Rows 1, 6, 11 and 16 can have 0, 2, 4, 6 or 8 dots.

Since these rows are different, they contain different information compared to other rows. One of the most likely interpretations is that they indicate parity. Dividing the 15x8 grids into blocks of 5*2 supports this theory. All blocks have an odd number of dots. This can be seen in Figure 12.

The parity rows still have an even number of dots, even though parity bits do not need to comply with such rules. This is a result of simple math. Four times an odd number results in an even number. The four rows above always contain four dots. Any even number minus four results in an even number. Therefore, rows 1, 6 and 11 must have an even number of dots.

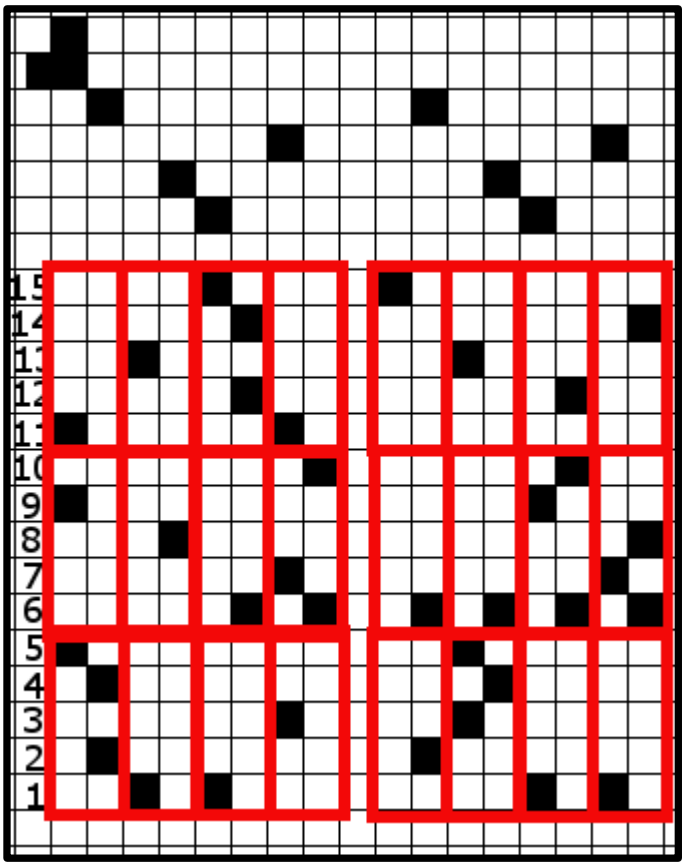

Figure 12: Red rectangles indicate parity blocks

Statement: Rows 1,6 and 11 are parity rows.

If this Statement is agreed upon, another assumption can be made. All the non-parity rows in a single 15x8 matrix contain a single dot. It appears in 1 out of 4 places. Interpreting this as a computable number, results in a quaternary number system. All dots can take the form of 0, 1, 2 or 3.

Statement: The pattern uses a quaternary number system.

Figure 10 shows all patterns superimposed. On the left side, rows 7, 12 and 13 have two spots not filled. On the right side, rows 14 and 15 are empty. A theory for these being empty is that these rows are the most significant bit. The quaternary number system and the most significant bit is used to describe the pattern as seen in Figure 13. This code interpretation is used in the following paragraphs. It is separated in octants with L1 up to L4 for the left side, and R1 to R4 for the right side. Every octant can have 4^4 = 256 different variations. In total there are 256^8 = 18.446.744.073.709.551.616 unique patterns.

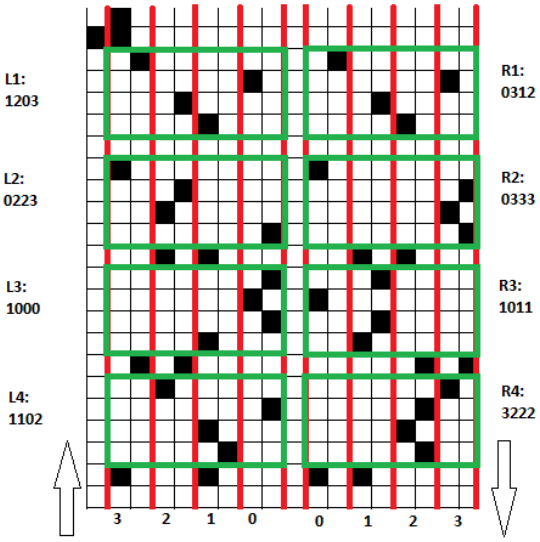

Figure 13: Pattern interpretation

iii) Top part

The main pattern consists of an upper and lower part. This paragraph describes the results observed in the upper part.

Figure 10 shows that the upper part does not change. The upper portion is portrayed in figure 14. A comparison with other manufacturers shows that the upper part changes per manufacturer.

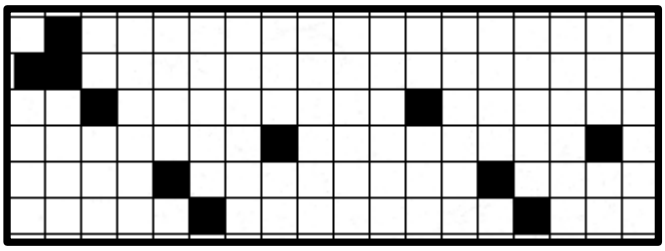

Figure 14a: HP

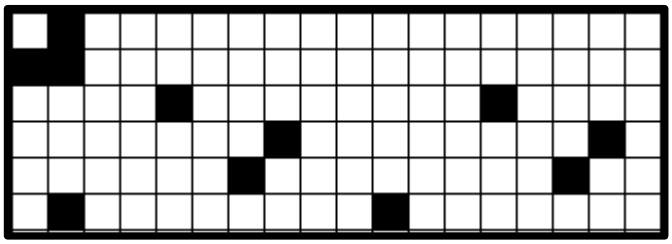

Figure 14b: Kyocera

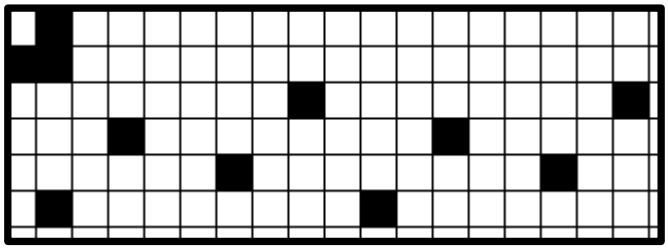

Figure 14c: Lexmark C912

A comparison of two different Lexmark Pattern gives the same upper pattern. Kyocera and Ricoh have multiple unique headers. Table 2 shows the specific numbers belonging to manufacturers.

Statement: Octants L1 and R1 of the reverse-L pattern indicate which printer manufacturer printed the document.

Table 2: List of manufacturers with their corresponding codes in L1 and R1

Manufacturer:	L1:	R1:
Hewlett Packard	1203	0312
Lexmark	3120	3120
Kyocera	3210	3210
Kyocera	3102	1320
Ricoh	2310	3201
Ricoh	0132	1023

iv) Lower part

Similar printer models are placed on top of each other. Figure 15 shows the pattern of five different HP LaserJet 5500 superimposed. Rows 11 up to and including 15 are the same for each pattern. This indicates that these rows are used to indicate the printer model.

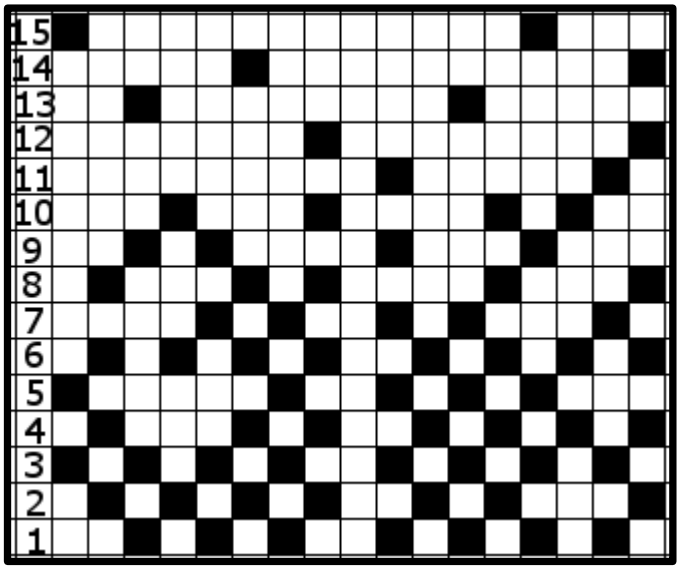

Figure 15: HP LaserJet 5500 multiple patterns

A similar observation is made with the HP LaserJet 4600 in figure 16. Two patterns of the 4600DN are placed on top of each other. A third pattern of the 4600N is also added. Deviating dots are colored red. The model number of the 4600N (C9692A) is slightly different from that of the 4600DN (C9661A). However, a comparison of two LaserJet 2550N printers show a different layout on R2. This could imply that R2 indicates something else, such as manufacturing factory or date. Table 3 contains the numbers in L2 and R2 corresponding to specific HP printers.

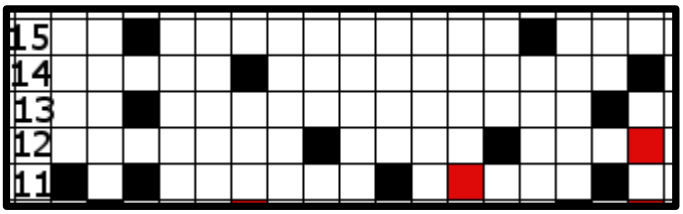

Figure 16: Comparison between 4600DN and 4600N

Statement: Octants L2 and R2 of the Reversed-L pattern represent the model.

Table 3: List of HP LaserJet models with their corresponding codes in L2 and R2.

Model	L2	R2	Model	L2	R2
LaserJet 1500	0221	0333	LaserJet 4600	0212	2331
LaserJet CP2025	0312	2323	LaserJet 4600	0212	2333
LaserJet 2500	0220	2233	LaserJet 4650	0222	2333
LaserJet 2500L	0220	2333	LaserJet 4650	0222	0333
LaserJet 2550	0230	2233	LaserJet 4700	0223	0333
LaserJet 2550	0230	2333	LaserJet 5500	0213	2313
LaserJet 2600	0232	2333	LaserJet 5550	0231	2333
LaserJet 2605dn	0303	0233	LaserJet 9500	0211	2333
LaserJet 3500	0221	2323	LaserJet CP1515n	0311	0213
LaserJet 3700	0221	2333	LaserJet CP1515n	0311	2313
LaserJet 4600	0212	0331	MFP M476dw	1211	0231
LaserJet 4600	0212	0333	MFP M476dn	1211	0312

v) Time and Date

Two different printers were used for this experiment. The first printer was an HP Color Laserjet MFP M476dn and the second was an HP Color Laserjet CP1515n.

Test pages were printed on the M476dn once every minute for 15 minutes. The system was then set to different hours, days and years. For every setting, a test page was printed. All of the printed patterns were the same. The same experiment was repeated with the CP1515n with the same result. This indicates that the printed time and date the page was printed is not part of the pattern. Chinese research gave the same result. [16]

Statement: Printed time and date is not part of the pattern.

vi) Content

The EFF dataset contains eight different printed pages per printer. These different printouts might affect the pattern printed. Comparing the patterns show that they do not differ from each other. This indicates that different content is not part of the pattern. Chinese research gave the same result. [16]

Statement: Content on the printout is not part of the pattern.

vii) Serial number

One of the most significant information components of a printer is its serial number. This unique number can always be found on a panel within the printer. Comparing this number with the patterns is done with two different methods. The first method uses different printers. All octants of the dataset were categorized. If an octant was equal to the same octant from a different printer, the serial number was compared. There appeared to be no similarities in the serial number.

Statement: Serial number is not part of the pattern.

The second method used a single printer. Its serial number was changed multiple times using the PRET tool. The following steps were taken:

1. Open cmd.
2. Type: pret.py (the IP address of the printer) pjl
3. Type: set serialnumber=(new serial number)

These steps were taken ten times with distinctly different serial numbers. For every serial number, a test page was printed. The printed pattern on each paper did not change. This result is not decisive, because the serial number can be hard-coded into the printer. Note: the serial number changes on the configuration page.

Statement: None –

viii) Formatter number

The previous experiment was repeated for the formatter number. Instead of changing the serial number, the following command was issued: set formatternumber=(new formatter number). The pattern did not change. The result is also not decisive, because the formatter number can be hard-coded into the printer. Note: the formatter number did change on the configuration page.

Statement: None –

ix) Even and odd serial number

A serial number is always either odd or even. This fact was used to determine whether a particular pattern repeats itself. For the “Reversed L” pattern, this means that a row can be four numbers. These numbers can be 0, 1, 2 and 3. If there is a row which is always 0 or 2 with even serial numbers and 1 and 3 with odd serial numbers, it indicates that this row encodes the least significant bit.

All patterns originating from a printer with an odd serial number were placed on top of each other. This resulted in figure 17. This figure was compared to the patterns with an even serial number. There were

no differences between both patterns. Two different conclusions can be drawn from this. The first one being that the pattern is not made with an even number system. An even number system is a numeral system that contains an even number of different characters. Every even number system always has the same numbers to determinate whether a number is even (0,2,4,6,8).

Statement: The pattern does not use an even number system.

The second statement is the opposite of the previous one. If it does use an even number system, then the serial number is not stated in the pattern since there are no differences between odd and even serial numbers. The manufacturer might have a database to connect a specific code to a serial number.

Statement: The serial number is not part of the pattern.

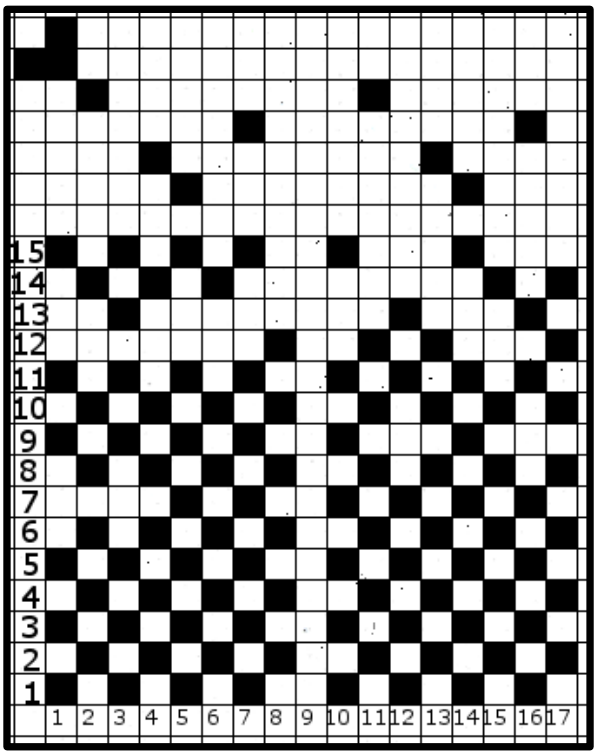

Figure 17: All odd serial number patterns

x) Firmware

Different versions of firmware might show different patterns. This was tested with an HP Color LaserJet CP1515n. The test page was printed with the firmware date code 20100616. Afterward, the printer was updated to a newer firmware version with date code 20130923. The test page was printed another time. Both pages had the same pattern. This indicates that the firmware has no influence on the HP pattern.

Statement: The firmware is not part of the HP pattern.

xi) Network or USB

Most laser printers can receive print commands either over network or via the USB interface. The experiment to test whether this influences the pattern consists of printing two documents. One is printed via the USB interface, the other over the network. The HP patterns were not different.

Statement: The HP pattern does not indicate whether the page was printed over a network or via the USB interface.

xii) Copy or print

This experiment is similar to the previous experiment. One test page is copied, and the other is printed. A comparison between both documents shows that it does not influence the pattern.

Statement: Whether a document is copied or printed is not part of the pattern.

xiii) Country of manufacture

Every serial number starts with two letters. These letters are usually CN (China) or JP (Japan) and state where the printer was built. Comparing patterns with the same country code shows no similarities. This indicates that the country of manufacture is not used in the pattern.

Statement: The country of manufacture is not present in the pattern.

xiv) Local IPv4 address

Printers connected to the internet or a network always have an IPv4 address. If the printers is connected to the internet, it has a unique address. If the printer is only connected to the network, it has a Local IPv4 address. With this experiment two similar pages were printed, each with a different IPv4 address. The patterns were compared and showed no differences. This indicates that the IPv4 address is not part of the pattern.

Statement: The local IPv4 address is not present in the pattern.

xv) Printer or Computer

The pattern is programmed at some point during the printing process. In general, there are two different possibilities. The print command can issue it on the computer through drivers, or the printer adds it locally. This was examined with .spl files usually saved to *"C:\Windows\System32\ spool\PRINTERS"*. These files are sent by the computer to the printer and contain the entire document to be printed. They are deleted once the printer has finished printing. [17]

A print command was issued to an HP LaserJet CP1515n. Before the printer could commence printing, the printer was turned off. The .spl file were virtually printed to pdf and examined for the presence of yellow dots. There were no dots added to the document.

Statement: The pattern is added to the document by the printer itself.

xvi) Decoding

This sub research aimed to reverse engineer part of the pattern to specific printer information. Ten different samples containing the configuration page were examined. The corresponding patterns were analyzed using different decoding methods. All unknown octants were, both individually and together, translated into various character sets. These sets include quaternary (numbers) and HP-Roman (ascii). None of the values defined on the configuration pages was found in the pattern.

Statement: The encoding method of the reverse-L pattern remains unknown.

5.4 Skewed small

One of the patterns that is seen less than the others is the Skewed Small. This pattern has been briefly researched. [18] This pattern has only been seen in Canon printers. It is recognized by a skewed pattern on the entire paper. The pattern always contains two dots next to each other that are not skewed. These two dots (colored red) are shown in Figure 18. The pattern is formed in a 31x31 matrix.

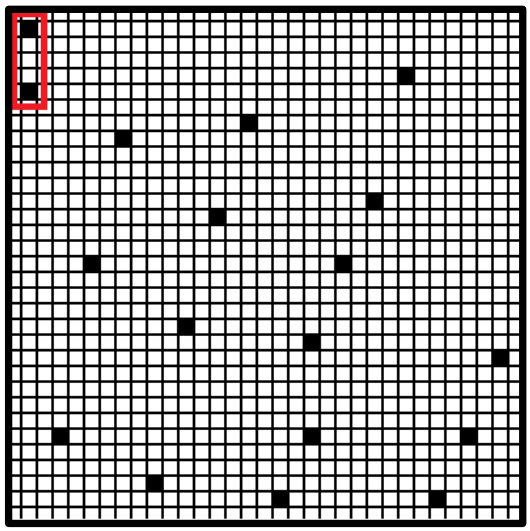

Figure 18: Example Skewed Small

The pattern is diagonally aligned. This means that the pattern is different for every point taken as a central dot. The pattern to the right is one way to interpret the pattern.

i) Main Pattern

Overlaying five different patterns produced Figure 19. This figure does not show a clear base pattern. More samples are needed to find it.

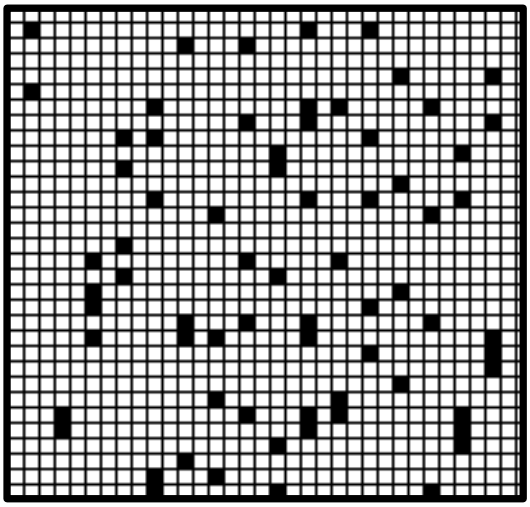

Figure 19: five patterns superimposed

ii) Time

Two different canon printers were used for this experiment. A test document was printed every minute for ten minutes. The pattern did not change. This shows that time is not part of the pattern.

Statement: Time is not part of the Skewed pattern.

iii) Date

Two documents were printed with the same Canon imageRUNNER c5030i. They were printed on different dates. An analysis of the pattern showed that it did not change. This indicates that date is not part of the pattern.

Statement: Date is not part of the Skewed pattern.

iv) Content

The EFF dataset contains eight different printed pages per printer. These different printouts might affect the pattern printed. Comparing the patterns show that they are not different from each other. This indicates that different content is not part of the pattern. It was verified by printing two different pages on a Canon ImageRunner.

Statement: Content on the printout is not part of the pattern.

5.5 Big Triangle

The Big Triangle pattern is a pattern printed by Canon ImageRUNNER C3220-C1. It has been observed in two printers in the EFF dataset (Printers 003 and 037). The pattern is recognizable by a big triangle. This triangle is shown in Figure 20. Its dimensions are 16x15.

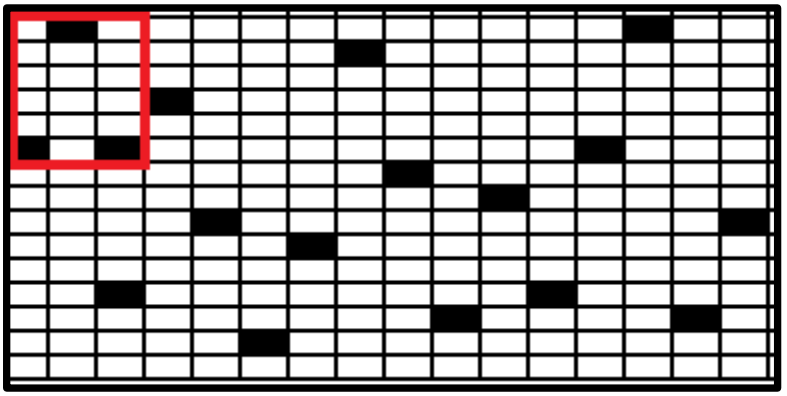

Figure 20: Sample Big Triangle

This pattern is a diagonally repeating pattern. Therefore, the pattern shown on the right is only an interpretation of the pattern. The real corners are unknown.

i) Time

Two different canon printers were used for this experiment. They were taken from the EFF dataset (003 and 037). Each printer had printed eight different printed documents originating from different files. Since the documents are different files, it implies that the print command was issued eight different times. All Xerox samples of the same dataset had different patterns for individual documents. Therefore, the assumption is made that the canon samples are also printed at different times. The patterns of the documents were compared. The pattern did not change. This indicates that time (minutes) is not part of the pattern.

Statement: Time is not part of the Big Triangle pattern.

ii) Content

The EFF dataset contains eight different printed pages per printer. These different printouts might affect the pattern printed. Comparing the patterns show that they do not differ from each other. This indicates that different content is not part of the pattern.

Statement: Content on the printout is not part of the pattern.

5.6 Double type

The double type pattern has been observed in two Konica Minolta ColorForce printers. They were part of the EFF Dataset (Printers 013 and 062). This pattern aligns in a grid with relatively clear sidelines. It is recognizable by two reversed-'L's as seen in the figure to the right.

The pattern consists of two different parts, the upper and the lower part. Just like the Reverse-L pattern, the dots are never horizontally or vertically next to each other. A sample is seen in Figure 21.

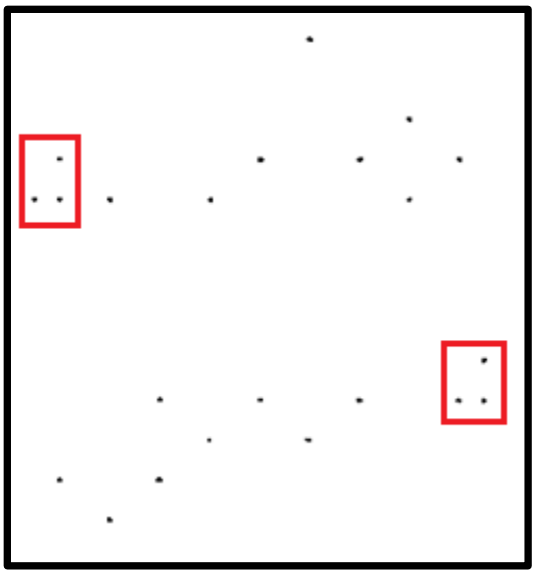

Figure 21: Sample Double type

i) Time

Two different canon printers were used for this experiment. They were taken from the EFF dataset (013 and 062). Each printer had eight different printed documents originating from different files. Since the documents are different files, it implies that the print command was issued eight different times. All Xerox samples of the same dataset had different patterns for individual documents. Therefore, the assumption is made that the Konica Minolta samples are also printed at different times The patterns of the documents were compared. The pattern did not change. This indicates that time (minutes) is not part of the pattern.

Statement: Time is not part of the Big Triangle pattern.

ii) Content

The EFF dataset contains eight different printed pages per printer. These different printouts might affect the pattern printed. Comparing the patterns show that they are not different from each other. This indicates that different content is not part of the pattern.

Statement: Content on the printout is not part of the pattern.

5.7 Corner

The Corner pattern has been observed in Konica Minolta and Epson machines. It consists of a grid of 15x24 dots. The recurring pattern consists of a corner in the left upper corner as shown in Figure 22. Furthermore, the pattern is grid aligned. This makes it hard to find the precise configuration of the pattern.

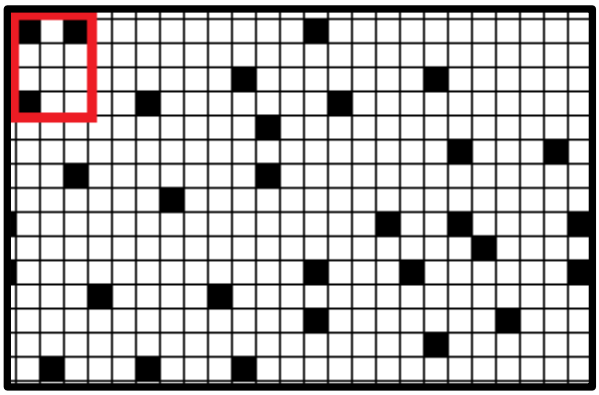

Figure 22: Sample Corner

i) Main Pattern

Fourteen different patterns of different Konica Minolta and Epson printers were placed on top of each other to distinguish the recurring base pattern. Figure 23 provides an overview of the blocks used. It also shows a horizontal line (colored blue) on which there are no dots present. This line might indicate a border of the pattern.

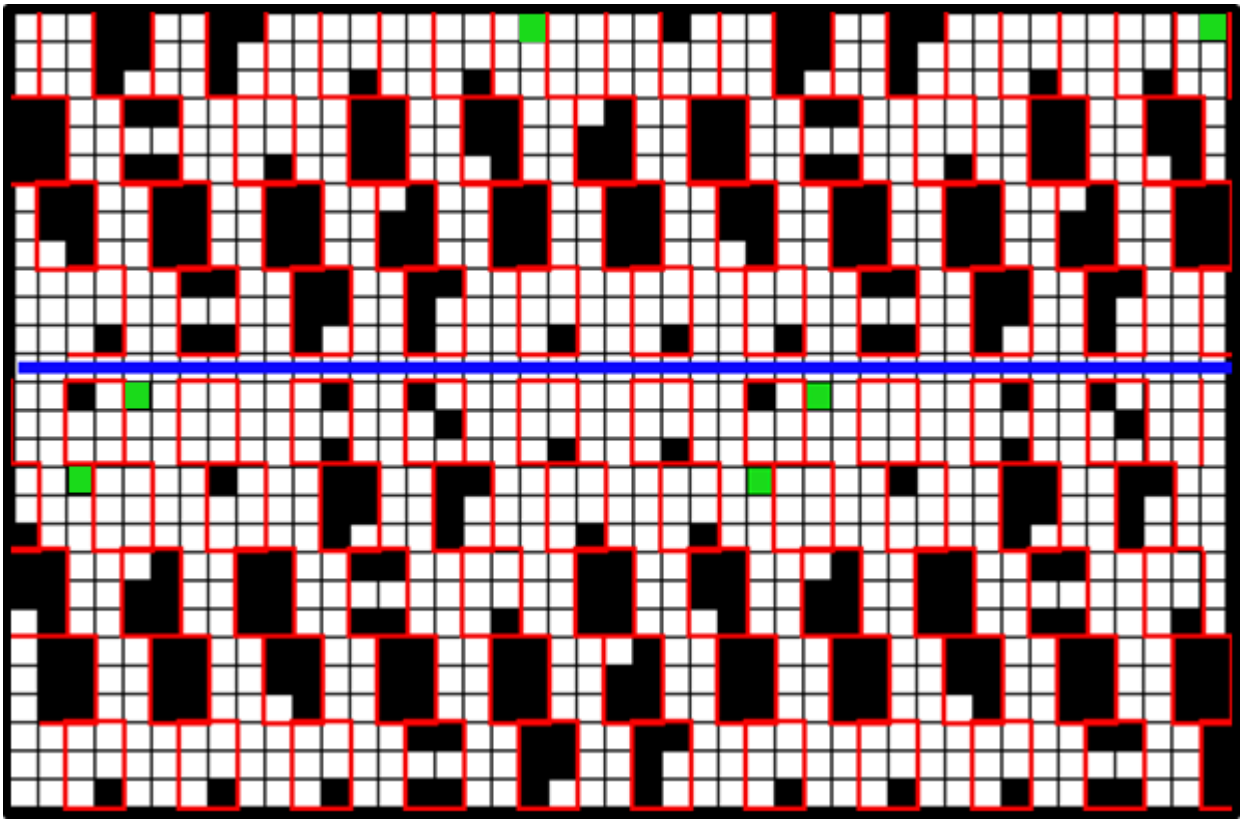

Figure 23: All corner patterns superimposed.

Furthermore, the dots are all situated in blocks of 2x3 (colored red) with blocks of 2x3 whitespace in between. This allocation of blocks also provides an additional proof of the blue line being a border. Two dots do not conform to the red blocks. These dots (colored green) are visible in each pattern.

ii) Similarity

Twenty-one different patterns were compared. Four of these patterns were Epson; the others were Konica Minolta. All patterns appear to have parity dots. A comparison of all patterns is shown in Figure 24. The boxes containing the same numbers are equal for twenty-one different patterns. The chance of two boxes being similar is 1 in 6. For twenty-one different patterns this results in a chance of 1 in 6^21 (21.936.950.640.377.856).

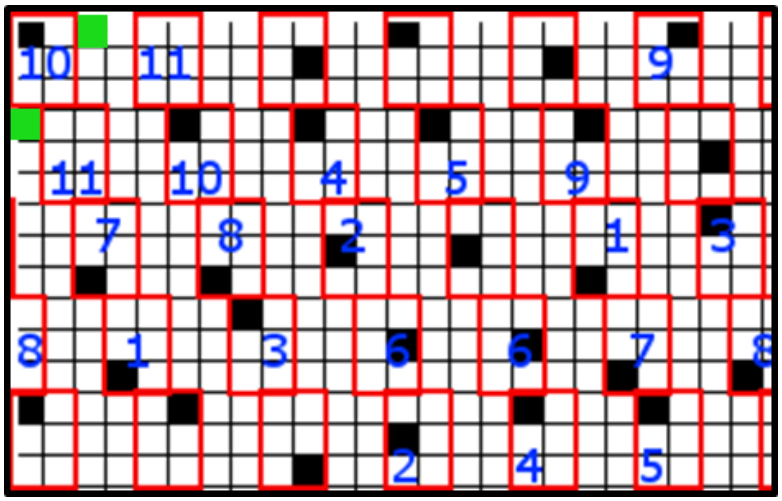

Figure 24: Red rectangles with the same number are always the same.

There are eight unique unnumbered red boxes. Boxes ten and eleven never change. Boxes one to nine have a parity box. This indicates that these dots are present to recognize the specific pattern, but do not contain data. In total, there are 17 dots that contain data (eight unnumbered and nine numbered).

Statement: The Corner pattern consists of 17 useful dots, with parity dots acting as a check.

iii) Time

Six different Konica Minolta printers were used for this experiment. They were taken from the EFF dataset (001, 011, 058,059, 060 and 063). The data set contained eight different printed documents originating from different files for each printer. Since the documents are different files, it implies that the print command was issued eight different times. All Xerox samples of the same dataset had different patterns for individual documents. Therefore, the assumption is made that the Konica Minolta samples are also printed at different times. The patterns of the documents were compared. The pattern did not change. This indicates that time (minutes) is not part of the pattern.

Statement: Time is not part of the corner pattern.

iv) Even and odd serial number

A serial number is always either odd or even. This fact was used to determine whether a specified pattern repeats itself. If a serial number is even and the number system is even, then the least significant bit is always even. For the “corner” pattern, this means that a box of 2x3 has 6 different numbers. These numbers can be 0, 1, 2, 3, 4, 5. If there is a box which is always 0, 2 or 4 with even serial numbers and 1, 3 and 5 with odd serial numbers, it indicates that this box encodes the least significant bit.

Five patterns originating from a printer with an even serial number were placed on top of each other. This resulted in figure 25. This figure was compared to the pattern with five odd serial number placed on top of each other (figure 26). There were no boxes in which all dots were different. Two different statements can be drawn from this. The first one being that the pattern is not made with an even number system since every even number system always has the same numbers to determinate whether a number is even (0,2,4,6,8).

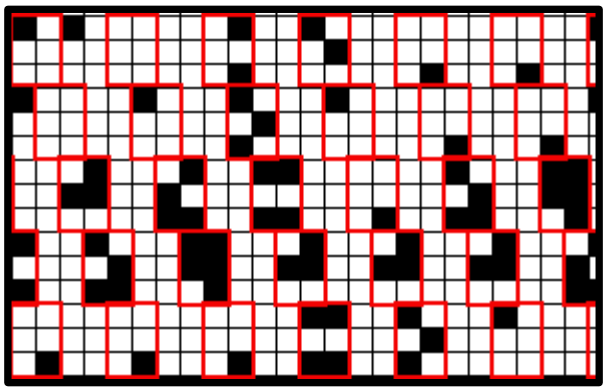

Figure 25: Even serial numbers

Statement: The pattern does not use an even number system.

The second statement is the opposite of the previous one. If it does use an even number system, then the serial number is not stated in the pattern since there are no differences between odd and even serial numbers. The manufacturer might have a database to connect a specific code to a printer.

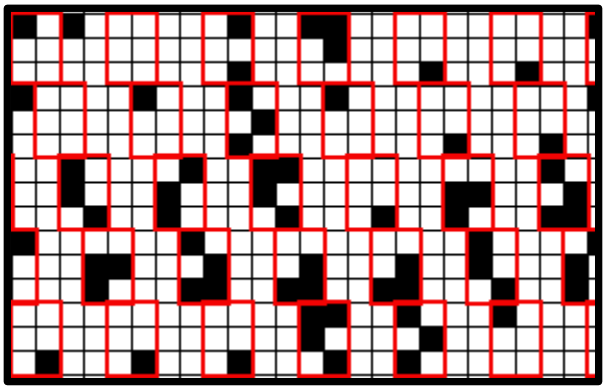

Figure 26: Odd serial numbers

Statement: The serial number is not part of the pattern.

v) Content

The EFF dataset contains eight different printed pages per printer. These different printouts might affect the pattern printed. Comparing the patterns show that they do not differ from each other. This indicates that different content does not influence the pattern.

Statement: Content on the printout does not influence the pattern.

vi) Decoding

This sub research aimed to reverse engineer part of the pattern to specific printer information. Each red box has six different variations, since it has 2x3 different dot positions. This might indicate that it uses a senary number system (0, 1, 2, 3, 4, 5). Decoding the pattern with this number system and comparing them with values on the configuration pages did not conclude in anything.

Statement: The encoding method of the Corner pattern remains unknown.

6. Discussion

The most significant discussion point is the dataset. This dataset was collected by the EFF and scanned by an independent research group. Therefore, there are three different steps where mistakes could have been made. People sending the samples may have misread the serial number or deliberately/accidentally

filled in the wrong information. Furthermore, a technician might have changed the serial number of the printer. This can have visually changed the serial number yet not affect the yellow dots. The EFF may have misplaced certain documents. This could cause information mismatches in the pattern and attached information. Lastly, there is the research group that scanned the material. They too may have mismatched the information.

In addition to the possibility of mismatching information, the dataset did not contain enough information. Only a handful of samples contained a configuration page with all the known variables. Due to the lack of information, many variables could not be investigated.

Other possible errors may have happened when schematically coloring the patterns. Some of the patterns were not as clear as others. This could have caused the schematic overview of a pattern to be wrong. As for the reverse-L pattern, the chance of this happening is low, since it contains a parity row. Even if a mistake had been made, it would be easy to find and fix.

The most important result is time. It does not conform to the hypothesis that it is included in the pattern. I assumed it would be included since the Xerox pattern does use time. After this research, the general conclusion is that only the Xerox pattern uses time. This is an assumed statement for some patterns because there were only one or two sample printers available. These printers may not even have had an internal clock. The HP Color LaserJet CP1515n, used for most of the Reverse-L pattern research, did not contain an internal clock. Additionally, there is a chance that the samples were printed at the same moment. This is however unlikely since the test documents were all different files. Furthermore, all Xerox samples from the same printer had different patterns. This should be the case for other printers as well.

In addition to time, the date variable is a less researched variable. This variable can only be researched manually since the printouts from the dataset were most likely printed on the same day. The experiment to test whether the date was a factor in the pattern can be performed if such a printer is available. I did not have access to most of these printers.

Comparing two different documents in a forensic investigation needs certainty that the pattern can be linked to the used printer. If a printer remains untouched, only its time and date change. This research proved that time is not included in most patterns. A document can be linked to a printer by comparing the pattern on the evidence document to documents from possible printers. The pattern does not change if the researched variables change, unless it is the Xerox pattern. If it is a Xerox pattern, it can be linked using the known decoding.

An independent Chinese article finds the same base pattern for the Reverse-L [16]. This is an additional reason to believe that the proposed pattern is the correct pattern. The same method has also been applied to the “corner”. It does not necessarily have to be correct, but it does show a pattern.

The in-depth research on an HP Color LaserJet CP1515n consisted of changing variables in its system, expecting the pattern to change. It did not change. This can be explained by two independent reasons.

One reason is that the correct variable stated in the pattern has not been found. The other reason is that the code is hard-coded.

Useful articles about this subject are very scarce. The three known articles [3] [13] [16] are resourceful but do not result in a thorough analysis. There are some forum posts and news articles with similar content. They are all based on the research done by EFF, with little to no additional information.

7. Conclusion

This research aimed to understand how to link specific documents to the printer used to print them using the Machine Identification Code. An analysis was performed on the dataset supplied by the EFF and Joost van Beusekom et al. [13] Another analysis was performed on a newly formed dataset. Both confirmed that time and date were not part of the patterns of most patterns. Time and date are the only variables that change over time without any manual intervention. Thus, linking a document to a printer is possible by comparing patterns. If the patterns are the same, the printer was used to print the document.

Global analysis found that the MIC does not necessarily have to be present on the entire page. It can be locally present on colored areas. Furthermore, the base pattern of both the Reverse-L and the corner is found. This can aid in recognizing and decoding the pattern.

Individual analysis of the Reverse-L pattern indicates that most variables are not present in the MIC. These variables include time, date, firmware, serial number, IP-address, and content. The Reverse-L does include the manufacturer and specific printer model.

8. Recommendations

There are many opportunities for follow-up research. The first step is to create a more extensive dataset with more printers, preferably newer ones. Even if decoding the pattern is not successful, specific information can still be used. Example given: The Reverse-L encoding is unknown. However, the dots stating which printer series it is are known. Many patterns of unknown origin can be connected to a specific series. If the dataset is enlarged with newer printers, there is a higher chance of finding the correct printer. An important detail to keep in mind is to gather as much information as possible from each printer. As for the Reverse-L pattern, the serial number is most likely not included. Other data can still be relevant.

I spent much time finding the patterns and marking them schematically. This process can be speeded up and made more time efficient by creating an automatic program or script. Joost van Beusekom et al. found a way to identify patterns automatically. [13] The pattern can be converted to a matrix of ones and zeros using a script.

Another option for follow-up research consists of repeating the experiments made on the reverse-L pattern with different printers. If possible, use a printer with old firmware. HP firmware has contained a digital signature since March 2012. The signature prevents anyone from altering the firmware (and

possibly other variables) without a key. [19] Most other manufacturers use similar methods to secure their printers. [20]

The Xerox printer contained an internal file with system information. Other printers might also have these documents on their hard drive. Extracting this information from other laser printers can give insight into specific data included in the pattern.

Forensic investigations on suspicious documents can use this research to retrieve the used printer. If the MIC is not visible by scanning, adjust the settings of the scanner. Altering the contrast, darkness, and sharpness can make the pattern visible. If it remains invisible, look at colored areas (preferably blue).

9. References

[1] R. Graham, "How Intercept outed reality winner," 5 June 2017. [Online]. Available: http://blog.erratasec.com/2017/06/how-intercept-outed-reality-winner.html#.Wp0FpWrOWUl.

[2] J. Tuohey, "Government uses color laser printer technology to track documents," 22 November 2004. [Online]. Available: https://www.pcworld.com/article/118664/article.html.

[3] S. D. Schoen, R. Lee and P. Murphy, "Docucolor Tracking Dot Decoding Guide," 2005. [Online]. Available: https://w2.eff.org/Privacy/printers/docucolor/.

[4] W. de Vries, "Dutch track counterfeits via printer serial number," ITWORLD, 25 October 2004. [Online]. Available: https://www.itworld.com/article/2812035/data-center/dutch-track-counterfeits-via-printer-serial-numbers.html. [Accessed 17 May 2018].

[5] Various, "FOIA Printer Dots," 19 December 2008. [Online]. Available: https://www.eff.org/files/filenode/foiaprinterdots/20081219_bep_printer_dots_01.pdf . [Accessed 14 May 2018].

[6] J.-H. Chois, D.-H. Im, H.-Y. Lee, J.-T. Oh, J.-H. Ryu and H.-K. Lee, "COLOR LASER PRINTER IDENTIFICATION BY ANALYZING STATISTICAL FEATURES," in *IEEE International Conference on Image Processing (ICIP)*, Cairo, 2009.

[7] P.-J. Chang, N. Khanna, A. K. Mikkelineni, M. V. Ortiz Segovia, S. Suh, J. P. Allebach, G. T. Chiu and E. J. Delp, "Printer and Scanner Forensics," 2009. [Online]. Available: https://engineering.purdue.edu/~prints/public/papers/sp_article_09_chiang.pdf. [Accessed 17 May 2018].

[8] M. Rouse, "Laser Printer," Mei 2010. [Online]. Available: http://whatis.techtarget.com/definition/laser-printer.

[9] Xerox, "Print Resolution," 9 December 2012. [Online]. Available: http://www.office.xerox.com/latest/XOGFS-17.

[10] S. Schoen, "List of printers Which do or do not display tracking dots," Electronic Frontier Foundation, 2017. [Online]. Available: https://www.eff.org/pages/list-printers-which-do-or-do-not-display-tracking-dots. [Accessed 7 May 2018].

[11] K. L. Prewitt, "Microdots.pdf," 13 February 2012. [Online]. Available: https://www.scribd.com/doc/81897582/microdots-pdf.

[12] S. Schoen, "Machine Identification Code (MIC) Dataset," EFF, 2005. [Online]. Available: https://madm.dfki.de/downloads-ds-mic. [Accessed 9 April 2018].

[13] J. van Beusekom, F. Shafait and T. Breuel, "Researchgate," November 2013. [Online]. Available: https://www.researchgate.net/publication/257471976_Automatic_authentication_of_color_laser_print-outs_using_machine_identification_codes. [Accessed 5 March 2018].

[14] J. Mueller, "Printer Exploitation Toolkit," Ruhr University Bochum, 2017. [Online]. Available: https://github.com/RUB-NDS/PRET. [Accessed 12 April 2018].

[15] P. Buck, "Machine Identification Code Dataset," 4 June 2018. [Online]. Available: https://perudo.stackstorage.com/s/qdl8KHkfJ54p9fM. [Accessed 4 June 2018].

[16] 吴玉宝孔祥维, "Yellow spot array information extraction method in colored laser printing files," 04 Juli 2012. [Online]. Available: https://patents.google.com/patent/CN101853384B/en. [Accessed 19 April 2018].

[17] S. Bunting, Encase Computer Forensics The Official EnCE: Encase certified Examiner, Indianapolis: John Wiley & Sons. Copyright, 2012.

[18] N. Gessler, "Stegonography - Color Laser Copier Dot Codes," [Online]. Available: https://people.duke.edu/~ng46/collections/steg-color-copier-dot-code.htm. [Accessed 12 April 2018].

[19] HP Development Company, 22 March 2012. [Online]. Available: https://support.hp.com/nl-nl/document/c03102449. [Accessed 23 April 2018].

[20] "Firmware Updates," 3 Juli 2017. [Online]. Available: https://hacking-printers.net/wiki/index.php/Firmware_updates. [Accessed 23 April 2018].

10. Attachment

Table 4: Summary of the known or researched variables of all pattern types. Yes means it is included in the pattern. This table can be enlarged by the community/further research.

Variable researched/known	*Xerox*	*Reverse-L*	*Skewed small*	*Big Triangle*	*Double type*	*Corner*
Time	Yes	No	No	No	No	No
Date	Yes	No	No	-	-	-
Serial number	Yes	No	-	-	-	No
Even/Odd serial number	Yes	No	-	-	-	No
Printed/copied	Yes	No	-	-	-	-
Parity	Yes	Yes	-	-	-	Yes
Manufacturer	No	Yes	-	-	-	-
Model	No	Yes	-	-	-	-
Content	No	No	No	No	No	No
Formatter Number	No	No	-	-	-	-
Firmware	No	No	-	-	-	-
Network/USB	No	No	-	-	-	-
Country of manufacturing	No	No	-	-	-	-
Local IPv4	No	No	-	-	-	-
Printer/computer	Printer	Printer	-	-	-	-
Encoding	Binary	Unknown	-	-	-	Unknown