TREND UNIVERSITY

Trend University is to Train You

# Open Source Similarity Digests
## DFRWS August 2016

*Jonathan Oliver*

# *Notes on this class*

- Discuss how Similarity Digests work
  - Ssdeep
  - TLSH
  - Sdhash
  - Nilsimsa
- Have practice using them
- Cover important cases

The notes include slides labeled FULL DETAILS – we will not go over these in the class

**Acknowledgement:**

Work done with Scott Forman, Chun Cheng, Yanggui Chen & Vic Hargrave

Also thanks to Jennifer Rihn for notes on the class

# *What are Similarity Digests?*

- Traditional hashes (such as SHA1 and MD5) have the property that a small change to the file being hashed results in a completely different hash

- Similarity Digests and Locality Sensitive Hashes (LSH) have the property that a small change to the file being hashed results in a small change to the hash
  - You can measure the similarity between 2 files by comparing their digests

# *Similarity Digests*

- Similar files / images / documents
  - Spam / attachments
  - Malware families

- Does not solve
  - Packing issues
  - Encryption
  - Compression (zip files, jpg, gif, etc)
  - Encoding

- Use Security / Forensic knowledge to extract the required content
  - Then use Similarity Digests

# *Similarity Digests*

**4 pieces of open source software**

1. Ssdeep is the Industry standard (in Virus Total and NIST)
2. TLSH is Trend Micro’s LSH
    - Less vulnerable to attack
    - Enables fast search
3. Sdhash
    - Literature says the Sdhash is better than Ssdeep
4. Nilsimsa
    - Proposed for spam signatures

# *Licenses*

| | | |
|---|---|---|
| 1. | Ssdeep | GPL |
| 2. | TLSH | Apache |
| 3. | Sdhash | Apache |
| 4. | Nilsimsa | Various |

The Apache license – an important detail.

Variants must include NOTICE.txt

# NOTICE.txt

```
===================================================
==  NOTICE file for use with the Apache License, Version 2.0,          ==
==  in this case for the Trend Locality Sensitive Hash distribution.   ==
========================================================

   Trend Locality Sensitive Hash (TLSH)
   Copyright 2010-2016 Trend Micro

   This product includes software developed at
   Trend Micro (http://www.trendmicro.com/)

   Refer to the following publications for more information:

     Jonathan Oliver, Chun Cheng and Yanggui Chen,
     "TLSH - A Locality Sensitive Hash"
     4th Cybercrime and Trustworthy Computing Workshop, Sydney, November 2013
     https://github.com/trendmicro/tlsh/blob/master/TLSH_CTC_final.pdf

     Jonathan Oliver, Scott Forman and Chun Cheng,
     "Using Randomization to Attack Similarity Digests"
     Applications and Techniques in Information Security. Springer Berlin Heidelberg, 2014. 199-210.
          https://github.com/trendmicro/tlsh/blob/master/Attacking_LSH_and_Sim_Dig.pdf
```

# Log into AWS

If you use Cygwin

(1) Put instance1.pem into some folder sim_digest

(2) In shell / Cygwin

```
$ cd sim_digest
$ ssh -i instance1.pem ec2-user@ec2-a-b-c-d-201.ap-southeast-2.compute.amazonaws.com
```

Where a-b-c-d is replaced with your allocated IP #

Do not use "." use "-" in between the numbers

(3) In AWS

```
$ ./alloc.sh   YOUR_NAME
$ cd Similarity_Digest_YOUR_NAME
```

# *Exercise 1A: Calculating Digests*

**chp1.txt – Chapter 1 of Pride and Prejudice**

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.
..
When she was discontented, she fancied herself nervous.
The business of her life was to get her daughters married; its solace was visiting and news.

```
$ cd Exercise1
$ ./exercise1A.sh
```

# Exercise 1B: Comparing Files

**chp1.txt – Chapter 1 of Pride and Prejudice**

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

..

When she was discontented, she fancied herself nervous.

The business of her life was to get her daughters married; its solace was visiting and news.

**chp1-.txt - Chapter 1 of Pride and Prejudice with last line removed**

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

..

When she was discontented, she fancied herself nervous.

# *Exercise 1B: Comparing Files*

```
$ ./exercise1B.sh

../bin/tlsh -f chp1.txt -c chp1-.txt
17      chp1.txt

../bin/sdhash -g chp1.txt chp1-.txt
chp1.txt|chp1-.txt|100

../bin/ssdeep -d -l chp1.txt chp1-.txt
chp1-.txt matches chp1.txt (100)

../bin/nilsimsa_ut -v1 -f chp1.txt -c chp1-.txt
254     chp1.txt
```

# Exercise 1B: Comparing Files

```
$ ./exercise1B.sh random.txt

../bin/tlsh -f chp1.txt -c random.txt
 324   chp1.txt

../bin/sdhash -t -1 -g chp1.txt random.txt
chp1.txt|random.txt|000

../bin/ssdeep -a -d -l chp1.txt random.txt
random.txt matches chp1.txt (0)

../bin/nilsimsa_ut -v1 -f chp1.txt -c random.txt
 130   chp1.txt
```

# *Score Ranges*

TLSH: distance score

| | |
|---|---|
| 0 | perfect match |
| 1 .. 100 | near perfect (1) to weak match (100) |
| 2000 | very distant files |

Ssdeep: similar score

Sdhash: similarity score

| | |
|---|---|
| 0 | no match |
| 1 .. 99 | match |
| 100 | perfect match |

Nilsimsa: similarity score

| | |
|---|---|
| 0 | perfect disagreement |
| 128 | no similarity |
| 256 | perfect match |

Trend Micro TLSH documentation

# Exercise 1C

Copy chp1.txt to 1c.txt

```
$ cp chp1.txt 1c.txt
```

Do a small change to 1c.txt

```
$ vi 1c.txt
    (if you do not like vi – then use nano)
$ ./exercise1B.sh 1c.txt
```

Exercise: Modify 1c.txt so that

TLSH (chp1.txt, 1c.txt) = 1 or 2

Ssdeep(chp1.txt, 1c.txt) = 99 or 100

Sdhash(chp1.txt, 1c.txt) = 99 or 100

Nilsimsa(chp1.txt, 1c.txt) = 255 or 256

# *Exercise 1D*

Simple transforms

What will some standard transformations do?

- sort
- fmt
- rot13  a->n b->o c->p … z->m
- lowercase A->a B->b …

```
$ ./exercise1D.sh
```

# *Exercise 1E*

Simple encodings

Encode a file and a close variant.

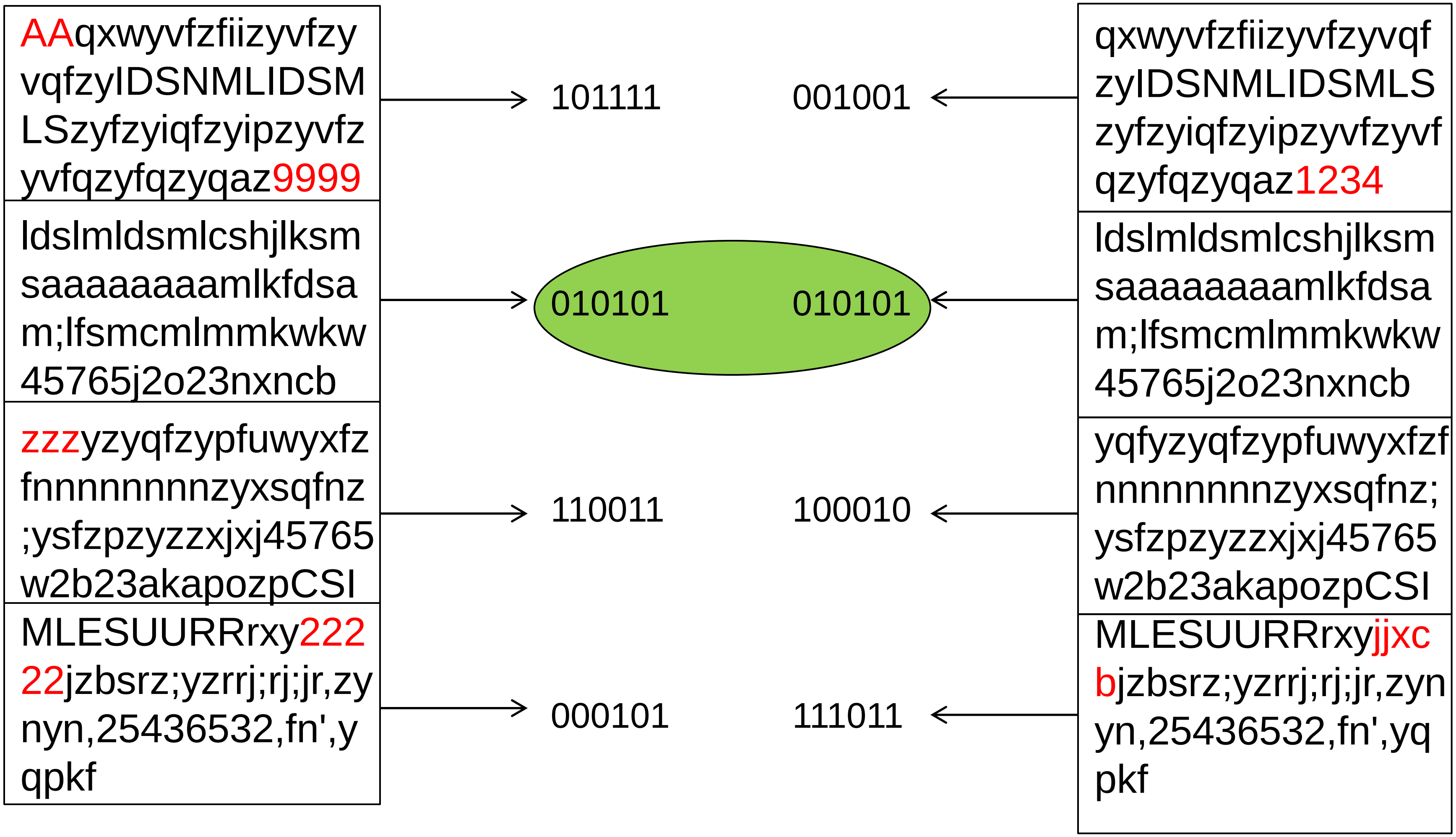
| | | | |
|---|---|---|---|
| File1 | => | base64 | file1.base64 |
| File1+ | => | base64 | file1+.base64 |

```
$ ./exercise1E.sh
```

# How do they work?

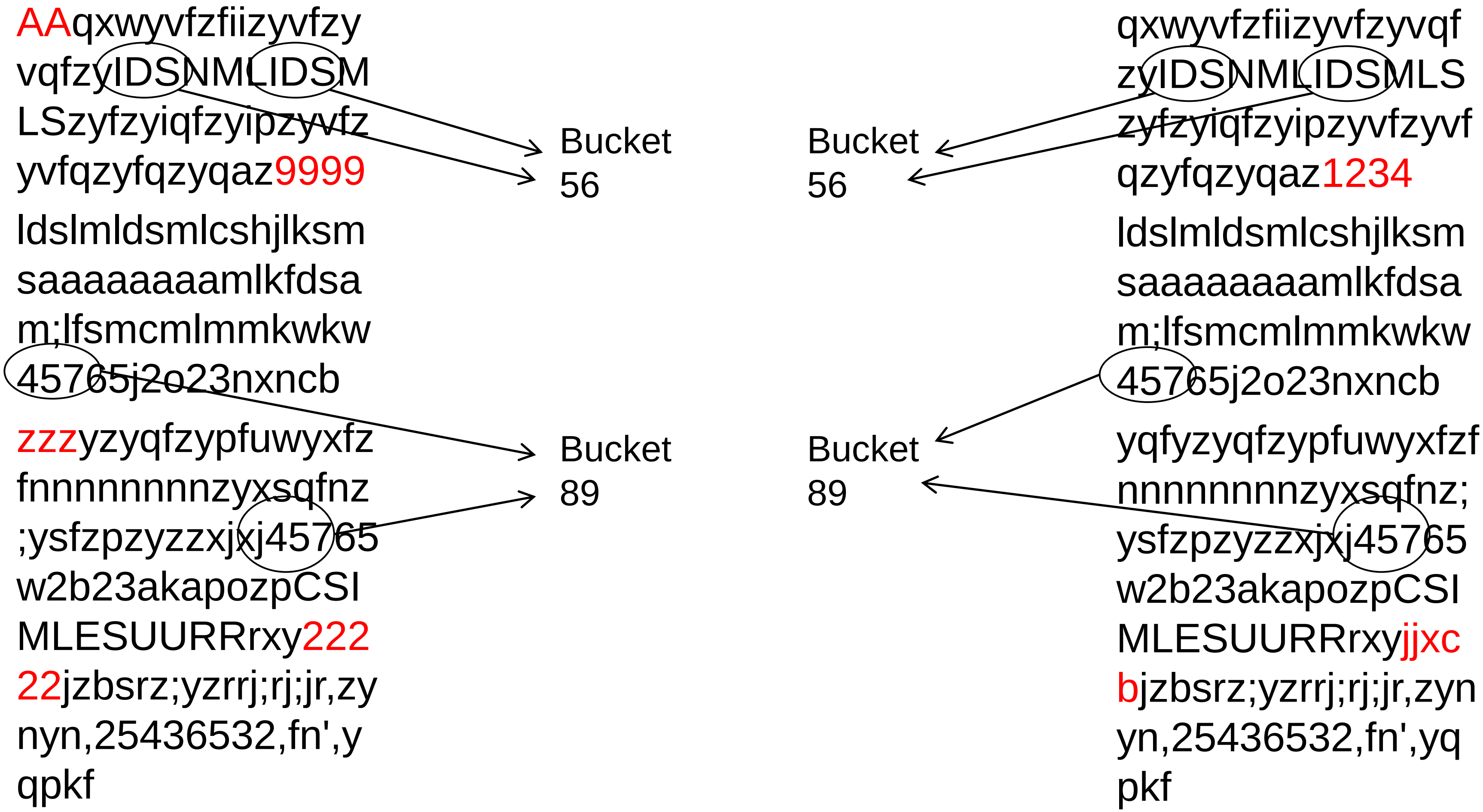
| | Context Triggered Piecewise Hashing | Feature Extraction | Locality Sensitive Hashes |
|---|---|---|---|
| Example | Ssdeep | Sdhash | Nilsimsa, TLSH |
| Creating the digest | Cut up a file into segments<br>Create a checksum for each segment<br><br>The digest is the concatenation of the checksums | Extract relatively long features (64 bytes) which are “interesting”<br><br>The digest is the encoded features | Extract many very small features (3 bytes)<br>Put the features into a histogram<br><br>The digest is the encoded histogram |
| Matching Digests | If enough checksums match – then the files match | If enough encoded features match – then the files match | Score the distance between the histograms |

# *Ssdeep*



| | | | |
|---|---|---|---|
| AAqxwyvfzfiizyvfzy<br>vqfzyIDSNMLIDSM<br>LSzyfzyiqfzyipzyvfz<br>yvfqzyfqzyqaz9999 | 101111 | 001001 | qxwyvfzfiizyvfzyvqf<br>zyIDSNMLIDSMLS<br>zyfzyiqfzyipzyvfzyvf<br>qzyfqzyqaz1234 |
| ldslmldsmlcshjlksm<br>saaaaaaaamlkfdsa<br>m;lfsmcmlmmkwkw<br>45765j2o23nxncb | 010101 | 010101 | ldslmldsmlcshjlksm<br>saaaaaaaamlkfdsa<br>m;lfsmcmlmmkwkw<br>45765j2o23nxncb |
| zzzyzyqfzypfuwyxfz<br>fnnnnnnnzyxsqfnz<br>;ysfzpzyzzxjxj45765<br>w2b23akapozpCSI | 110011 | 100010 | yqfyzyqfzypfuwyxfzf<br>nnnnnnnnzyxsqfnz;<br>ysfzpzyzzxjxj45765<br>w2b23akapozpCSI |
| MLESUURRrxy222<br>22jzbsrz;yzrrj;rj;jr,zy<br>nyn,25436532,fn',y<br>qpkf | 000101 | 111011 | MLESUURRrxyjjxc<br>bjzbsrz;yzrrj;rj;jr,zyn<br>yn,25436532,fn',yq<br>pkf |

# *Locality Sensitive Hashes (Nilsimsa, TLSH)*



AAqxwyvfzfiizyvfzy
vqfzyIDSNMLIDSM
LSzyfzyiqfzyipzyvfz
yvfqzyfqzyqaz9999
ldslmldsmlcshjlksm
saaaaaaaamlkfdsa
m;lfsmcmlmmkwkw
45765j2o23nxncb
zzzyzyqfzypfuwyxfz
fnnnnnnnnzyxsqfnz
;ysfzpzyzzxjxj45765
w2b23akapozpCSI
MLESUURRrxy222
22jzbsrz;yzrrj;rj;jr,zy
nyn,25436532,fn',y
qpkf

Bucket
56

Bucket
89

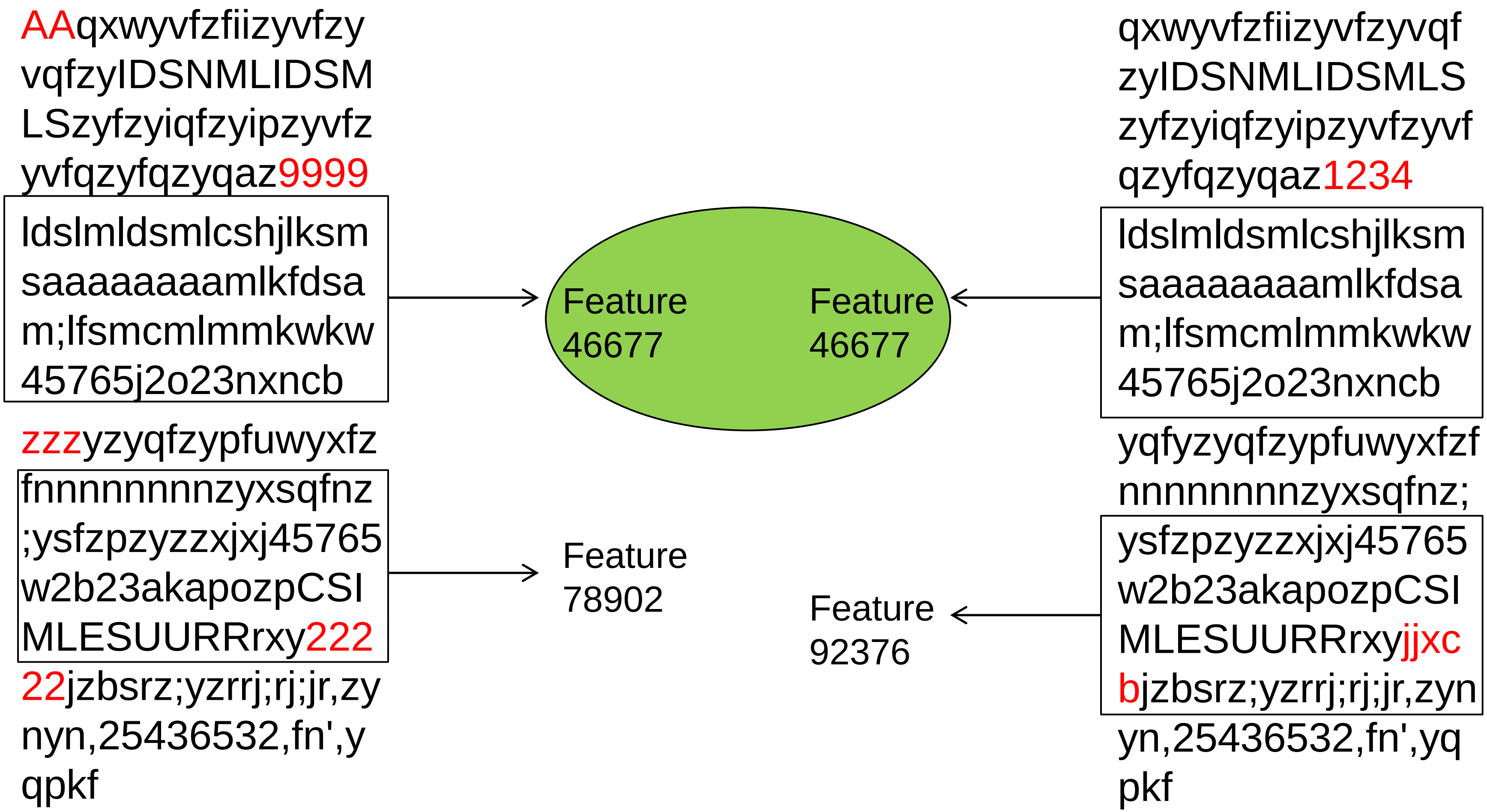
Bucket
56

Bucket
89

qxwyvfzfiizyvfzyvqf
zyIDSNMLIDSMLS
zyfzyiqfzyipzyvfzyvf
qzyfqzyqaz1234
ldslmldsmlcshjlksm
saaaaaaaamlkfdsa
m;lfsmcmlmmkwkw
45765j2o23nxncb
yqfyzyqfzypfuwyxfzf
nnnnnnnnzyxsqfnz;
ysfzpzyzzxjxj45765
w2b23akapozpCSI
MLESUURRrxyjjxc
bjzbsrz;yzrrj;rj;jr,zyn
yn,25436532,fn',yq
pkf

Trend Micro TLSH documentation

# Sdhash (feature extraction)



AAqxwyvfzfiizyvfzy
vqfzyIDSNMLIDSM
LSzyfzyiqfzyipzyvfz
yvfqzyfqzyqaz9999
ldslmldsmlcshjlksm
saaaaaaaamlkfdsa
m;lfsmcmlmmkwkw
45765j2o23nxncb
zzzyzyqfzypfuwyxfz
fnnnnnnnnzyxsqfnz
;ysfzpzyzzzxjxj45765
w2b23akapozpCSI
MLESUURRxy222
22jzbsrz;yzrrj;rj;jr,zy
nyn,25436532,fn',y
qpkf

Feature
46677

Feature
46677

Feature
78902

Feature
92376

qxwyvfzfiizyvfzyvqf
zyIDSNMLIDSMLS
zyfzyiqfzyipzyvfzyvf
qzyfqzyqaz1234
ldslmldsmlcshjlksm
saaaaaaaamlkfdsa
m;lfsmcmlmmkwkw
45765j2o23nxncb
yqfyzyqfzypfuwyxfzf
nnnnnnnnzyxsqfnz;
ysfzpzyzzzxjxj45765
w2b23akapozpCSI
MLESUURRxyjjxc
bjzbsrz;yzrrj;rj;jr,zyn
yn,25436532,fn',yq
pkf

Trend Micro TLSH documentation

# *Processing Directories*

We have set up for you 8 commands

Lists the digests for a directory of files

nil_list DIR

tlsh_list DIR

ssdeep_list DIR

sdhash_list DIR

Does a scoring comparison for every pair of files in a directory

nil_score DIR

tlsh_score DIR

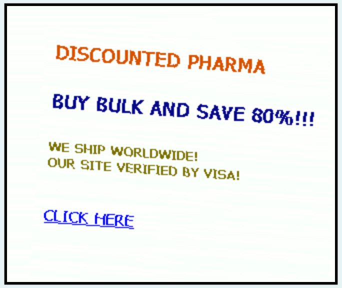
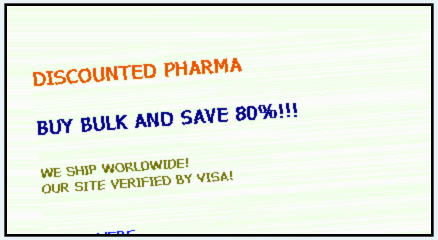
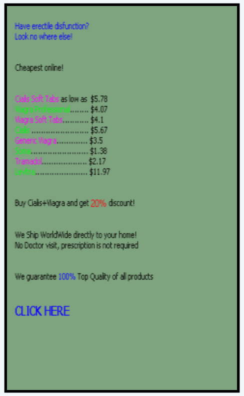
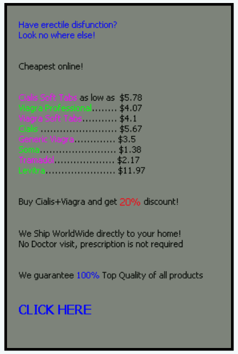
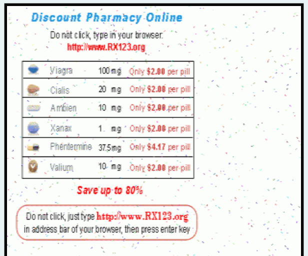
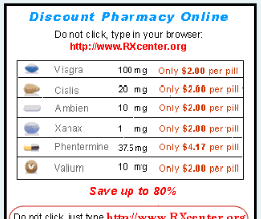
ssdeep_score DIR

sdhash_score DIR

# Processing Directories

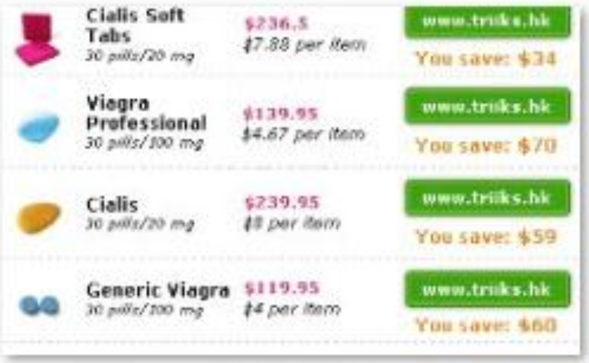
```
alias nil_list="/home/ec2-user/bin/nilsimsa_ut -v1 -r"
alias tlsh_list="/home/ec2-user/bin/tlsh -r"
alias ssdeep_list="/home/ec2-user/bin/ssdeep -r -l"
alias sdhash_list="/home/ec2-user/bin/sdhash -r"

alias nil_score="/home/ec2-user/bin/nilsimsa_ut -xref -v1 -r"
alias tlsh_score="/home/ec2-user/bin/tlsh -xref -r"
alias ssdeep_score="/home/ec2-user/bin/ssdeep -r -l -d -a"
alias sdhash_score="/home/ec2-user/bin/sdhash -r -g -t -999"
```

# *Working with Image Files*

| Folder Name | Manipulation | Image 1 | Image 2 |
|---|---|---|---|
| Angled | Image rotation | <br>DISCOUNTED PHARMA<br>BUY BULK AND SAVE 80%!!!<br>WE SHIP WORLDWIDE!<br>OUR SITE VERIFIED BY VISA!<br>CLICK HERE | <br>DISCOUNTED PHARMA<br>BUY BULK AND SAVE 80%!!!<br>WE SHIP WORLDWIDE!<br>OUR SITE VERIFIED BY VISA! |
| Pharmacy erectile dysfunction | Changing image height and width; Changing background colour | <br>Have erectile disfunction?<br>Look no where else!<br>Cheapest online!<br>Cialis Soft Tabs as low as $5.78<br>Viagra Professional........ $4.07<br>Viagra Soft Tabs........... $4.1<br>Cialis ......................... $5.67<br>Generic Viagra............. $3.5<br>Soma......................... $1.38<br>Tramadol................... $2.17<br>Levitra....................... $11.97<br>Buy Cialis+Viagra and get 20% discount!<br>We Ship WorldWide directly to your home!<br>No Doctor visit, prescription is not required<br>We guarantee 100% Top Quality of all products<br>CLICK HERE | <br>Have erectile disfunction?<br>Look no where else!<br>Cheapest online!<br>Cialis Soft Tabs as low as $5.78<br>Viagra Professional........ $4.07<br>Viagra Soft Tabs........... $4.1<br>Cialis ......................... $5.67<br>Generic Viagra............. $3.5<br>Soma......................... $1.38<br>Tramadol................... $2.17<br>Levitra....................... $11.97<br>Buy Cialis+Viagra and get 20% discount!<br>We Ship WorldWide directly to your home!<br>No Doctor visit, prescription is not required<br>We guarantee 100% Top Quality of all products<br>CLICK HERE |
| Pharmacy_Move | Changing image height and width; Adding dots, and dashes. | <br>Discount Pharmacy Online<br>Do not click, type in your browser:<br>http://www.RX123.org<br>Viagra 100 mg Only $2.00 per pill<br>Cialis 20 mg Only $2.00 per pill<br>Ambien 10 mg Only $2.00 per pill<br>Xanax 1 mg Only $2.00 per pill<br>Phentermine 37.5 mg Only $4.17 per pill<br>Valium 10 mg Only $2.00 per pill<br>Save up to 80%<br>Do not click, just type http://www.RX123.org in address bar of your browser, then press enter key | <br>Discount Pharmacy Online<br>Do not click, type in your browser:<br>http://www.RXcenter.org<br>Viagra 100 mg Only $2.00 per pill<br>Cialis 20 mg Only $2.00 per pill<br>Ambien 10 mg Only $2.00 per pill<br>Xanax 1 mg Only $2.00 per pill<br>Phentermine 37.5 mg Only $4.17 per pill<br>Valium 10 mg Only $2.00 per pill<br>Save up to 80%<br>Do not click, just type http://www.RXcenter.org |

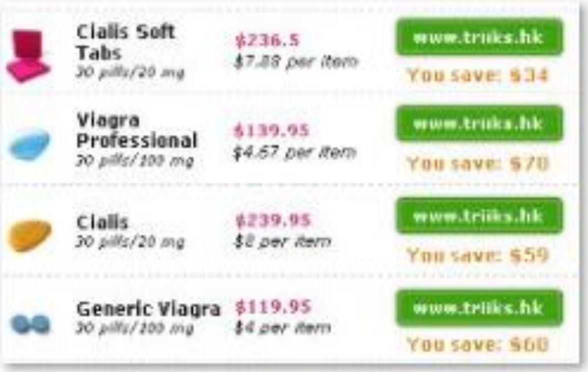
# *Working with Image Files*

Exercise2/Images_sorted_1000/Pharmacy_cialis_softtabs



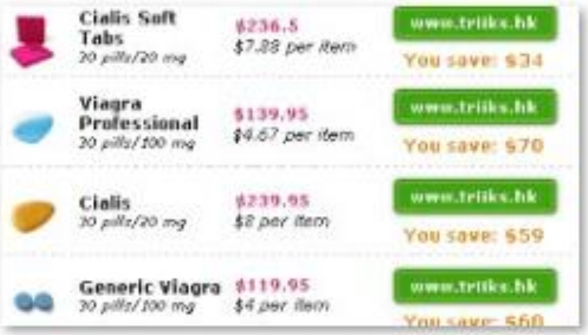
| | | |
|---|---|---|
| **Cialis Soft Tabs**<br>30 pills/20 mg | $236.5<br>$7.88 per item | www.triiks.hk<br>You save: $34 |
| **Viagra Professional**<br>30 pills/100 mg | $139.95<br>$4.67 per item | www.triiks.hk<br>You save: $70 |
| **Cialis**<br>30 pills/20 mg | $239.95<br>$8 per item | www.triiks.hk<br>You save: $59 |
| **Generic Viagra**<br>30 pills/100 mg | $119.95<br>$4 per item | www.triiks.hk<br>You save: $60 |

0e6f3429_0.gif



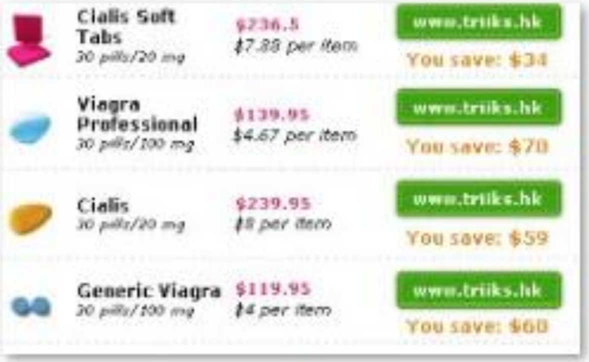
| | | |
|---|---|---|
| **Cialis Soft Tabs**<br>30 pills/20 mg | $236.5<br>$7.88 per item | www.triiks.hk<br>You save: $34 |
| **Viagra Professional**<br>30 pills/100 mg | $139.95<br>$4.67 per item | www.triiks.hk<br>You save: $70 |
| **Cialis**<br>30 pills/20 mg | $239.95<br>$8 per item | www.triiks.hk<br>You save: $59 |
| **Generic Viagra**<br>30 pills/100 mg | $119.95<br>$4 per item | www.triiks.hk<br>You save: $60 |

0f05d804_0.gif



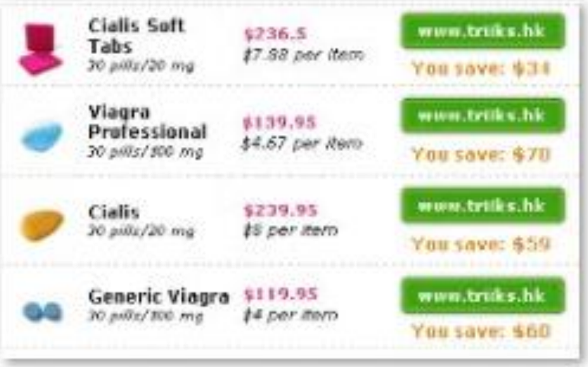
| | | |
|---|---|---|
| **Cialis Soft Tabs**<br>30 pills/20 mg | $236.5<br>$7.88 per item | www.triiks.hk<br>You save: $34 |
| **Viagra Professional**<br>30 pills/100 mg | $139.95<br>$4.67 per item | www.triiks.hk<br>You save: $70 |
| **Cialis**<br>30 pills/20 mg | $239.95<br>$8 per item | www.triiks.hk<br>You save: $59 |
| **Generic Viagra**<br>30 pills/100 mg | $119.95<br>$4 per item | www.triiks.hk<br>You save: $60 |

03303bb4_0.gif



| | | |
|---|---|---|
| **Cialis Soft Tabs**<br>30 pills/20 mg | $236.5<br>$7.88 per item | www.triiks.hk<br>You save: $34 |
| **Viagra Professional**<br>30 pills/100 mg | $139.95<br>$4.67 per item | www.triiks.hk<br>You save: $70 |
| **Cialis**<br>30 pills/20 mg | $239.95<br>$8 per item | www.triiks.hk<br>You save: $59 |
| **Generic Viagra**<br>30 pills/100 mg | $119.95<br>$4 per item | www.triiks.hk<br>You save: $60 |

084a03d7_0.gif



| | | |
|---|---|---|
| **Cialis Soft Tabs**<br>30 pills/20 mg | $236.5<br>$7.88 per item | www.triiks.hk<br>You save: $34 |
| **Viagra Professional**<br>30 pills/100 mg | $139.95<br>$4.67 per item | www.triiks.hk<br>You save: $70 |
| **Cialis**<br>30 pills/20 mg | $239.95<br>$8 per item | www.triiks.hk<br>You save: $59 |
| **Generic Viagra**<br>30 pills/100 mg | $119.95<br>$4 per item | www.triiks.hk<br>You save: $60 |

01047e2a_0.gif

# *Exercise 2A*
# *Working with Image Files*

```
$ cd Exercise2
```

Use the commands

nil_list, tlsh_list, ssdeep_list, sdhash_list
nil_score, tlsh_score, ssdeep_score, sdhash_score

to inspect the digests and similarity scores of

Images_sorted_1000/Pharmacy_cialis_softtabs

```
$ tlsh_list    Images_sorted_1000/Pharmacy_cialis_softtabs
$ tlsh_score   Images_sorted_1000/Pharmacy_cialis_softtabs
```

Can the digests determine that the images are similar?

# *Limitation of Similarity Digests*

Similarity Digests cannot identify files as being similar if they are

-Encrypted

-Compressed

-Packed malware

-Encoded

-…

You have to unpack, un-compress or decrypt first.

What do we need to do with gif, jpeg files?

# *Exercise 2B*
# *Working with Image Files*

Each of the image files has been converted (using CxImage library) to a .bmp file (a image bitmap – identical to the in memory representation of images)

.bmp files can be compared

Use *_list, *_score commands to inspect the digests and similarity scores of

Images_sorted_1000/Pharmacy_cialis_softtabs_bmp

```
$ cd Exercise2
$ ssdeep_list   Images_sorted_1000/Pharmacy_cialis_softtabs_bmp
$ ssdeep_score  Images_sorted_1000/Pharmacy_cialis_softtabs_bmp
```

# *Exercise 2B*
# *Working with Image Files*

Select a folder of images.

Which method(s) works on that folder?

Which method(s) fails on that folder?

If you are having problems with sdhash – you might have to

```
$ export LC_ALL="en_US.UTF-8"
```

A. Angled_bmp/
B. Pharmacy_erectile_dys_bmp/
C. Pharmacy_Move_2col_bmp/

```
$ cd Exercise2
$ tlsh_score     Images_sorted_1000/Angled_bmp
$ ssdeep_score   Images_sorted_1000/Angled_bmp
$ sdhash_score   Images_sorted_1000/Angled_bmp
```

# *Exercise 2C*
# *Working with Image Files*

Use the digests to work out what type of images are in Random_Images/

```
$ cd Exercise2
$ ./exercise2C.sh
```

`ssdeep`

Usage: ssdeep [-m file] [FILES]

-m - Match FILES against known hashes in file

`tlsh`

Usage: tlsh -l listdigests -c file

# *Exercise 2C (cont)*
# *Working with Image Files*

Need a hint?

```
$ cd Exercise2
$ cp ../Answers/answer2C.sh .
$ ./answer2C.sh | less
```

# *Collisions / False Positive Matches*

## Collision

When 2 distinct files have the same digest or hash

## False Positive Match

When the score is a match, but we consider

file1 not similar to file2

Ssdeep(file1, file2) > 0

Sdhash(file1, file2) > 0

TLSH(file1, file2) <= 100

Nilsimsa(file1, file2) >= 220

# *Collisions / False Positive Matches*

Exercise 2D

Do any of the methods suffer from collisions in the collection of image files?

Find one.

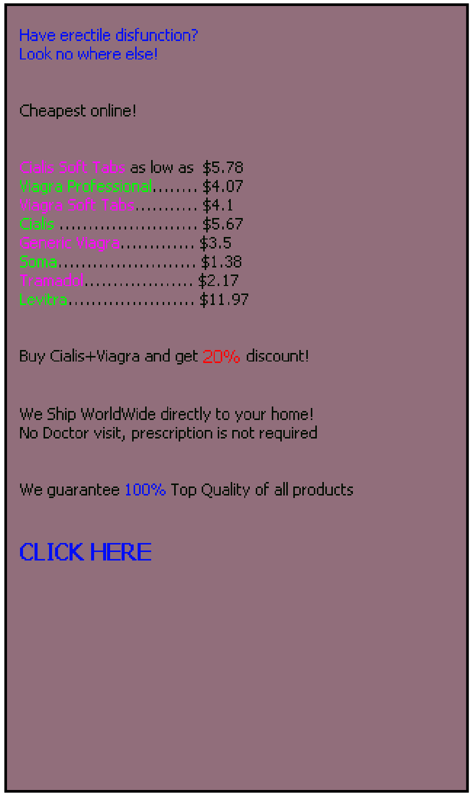
Exercise 2E

Do any of the methods suffer from false positive matches in the collection of image files?

Find one.

# Exercise 2D
# Nilsimsa Collision

```
0000000000400020000200000000000000009000000000040000000200800000
Images_sorted_1000/Pharmacy_Move_bmp/01aa7260_0.bmp
0000000000400020000200000000000000009000000000040000000200800000
Images_sorted_1000/Pharmacy_erectile_dys_bmp/05f66a41_0.bmp
```



Have erectile disfunction?
Look no where else!

Cheapest online!

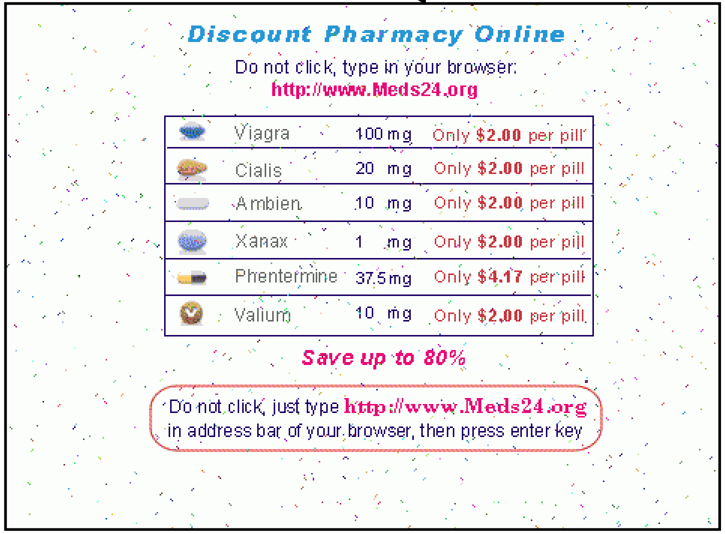
Cialis Soft Tabs as low as $5.78
Viagra Professional........ $4.07
Viagra Soft Tabs........... $4.1
Cialis ........................ $5.67
Generic Viagra............. $3.5
Soma........................ $1.38
Tramadol.................. $2.17
Levitra...................... $11.97

Buy Cialis+Viagra and get 20% discount!

We Ship WorldWide directly to your home!
No Doctor visit, prescription is not required

We guarantee 100% Top Quality of all products

CLICK HERE



Discount Pharmacy Online

Do not click, type in your browser:
http://www.Meds24.org

| | | | |
|---|---|---|---|
| | Viagra | 100 mg | Only **$2.00** per pill |
| | Cialis | 20 mg | Only **$2.00** per pill |
| | Ambien | 10 mg | Only **$2.00** per pill |
| | Xanax | 1 mg | Only **$2.00** per pill |
| | Phentermine | 37.5 mg | Only **$4.17** per pill |
| | Valium | 10 mg | Only **$2.00** per pill |

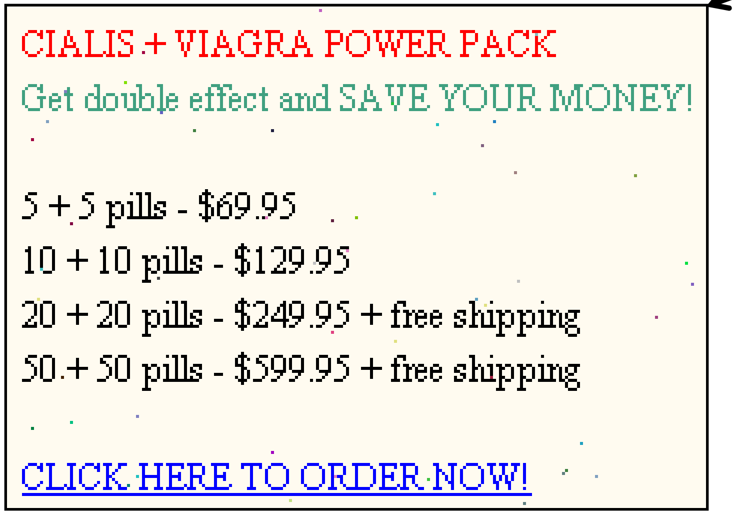
Save up to 80%

Do not click, just type **http://www.Meds24.org**
in address bar of your browser, then press enter key

Trend Micro TLSH documentation

# Exercise 2E
# TLSH False Positives

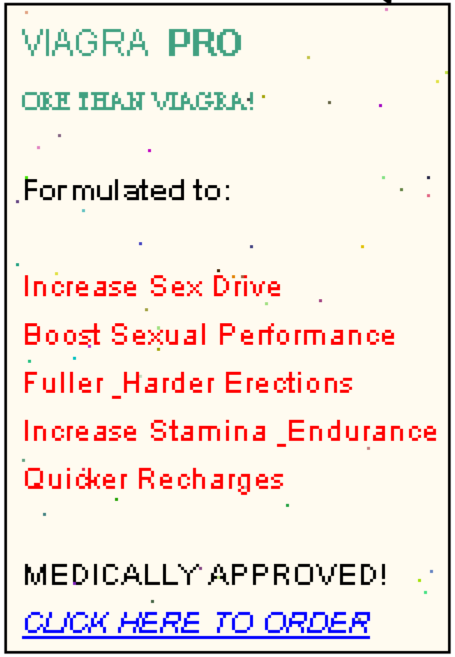
```
../bin/tlsh
        -c Images_sorted_1000/Pharmacy_Viagra_Pro_bmp/01b2bb87_0.bmp
        -f Images_sorted_1000/Pharmacy_power_pack_bmp/01a07331_0.bmp
  78    Images_sorted_1000/Pharmacy_power_pack_bmp/01a07331_0.bmp
```



CIALIS + VIAGRA POWER PACK
Get double effect and SAVE YOUR MONEY!

5 + 5 pills - $69.95
10 + 10 pills - $129.95
20 + 20 pills - $249.95 + free shipping
50 + 50 pills - $599.95 + free shipping

CLICK HERE TO ORDER NOW!



VIAGRA **PRO**
ORE THAN VIAGRA!

Formulated to:

Increase Sex Drive
Boost Sexual Performance
Fuller Harder Erections
Increase Stamina Endurance
Quicker Recharges

MEDICALLY APPROVED!
CLICK HERE TO ORDER

# Comparison on Image files

| Folder Name | Number images | TLSH (100) | Sdhash(1) | Ssdeep(1) | Nilsimsa (240 / 256) |
|---|---:|---:|---:|---:|---:|
| Angled | 20 | 80.0% | 3.7% | 0.0% | 100.0% |
| International Greek | 3 | 33.3% | 33.3% | 0.0% | 100.0% |
| Lotto | 11 | 100.0% | 100.0% | 100.0% | 100.0% |
| Pharmacy cialis softtabs | 5 | 100.0% | 100.0% | 10.0% | 100.0% |
| Pharmacy erectile dys | 147 | 22.1% | 22.6% | 9.6% | 94.7% |
| Pharmacy legal RX | 22 | 0.0% | 0.0% | 0.0% | 64.9% |
| Pharmacy_Move_2col | 22 | 90.5% | 100.0% | 10.8% | 100.0% |
| Pharmacy_Move | 63 | 12.1% | 11.2% | 1.0% | 61.2% |
| Pharmacy_Move_browser | 10 | 64.4% | 62.2% | 4.4% | 64.4% |
| Pharmacy_Move_browser_pikkie | 6 | 100.0% | 100.0% | 6.7% | 100.0% |
| Pharmacy picture | 8 | 57.1% | 3.6% | 7.1% | 57.1% |

# Comparison on Image files

| Folder Name | Number images | TLSH (100) | Sdhash(1) | Ssdeep(1) | Nilsimsa (240 / 256) |
|---|---:|---:|---:|---:|---:|
| Pharmacy pop a pill | 5 | 80.0% | 100.0% | 60.0% | 100.0% |
| Pharmacy power pack | 41 | 47.8% | 47.8% | 20.7% | 55.9% |
| Pharmacy research | 3 | 0.0% | 33.3% | 33.3% | 100.0% |
| Pharmacy Viagra Pro | 11 | 32.7% | 38.2% | 29.1% | 54.5% |
| Pharmacy Viagra Pro2 | 7 | 42.9% | 42.9% | 42.9% | 42.9% |
| Software OEM | 6 | 66.7% | 66.7% | 66.7% | 66.7% |
| Software SOBAKA | 11 | 100.0% | 100.0% | 100.0% | 100.0% |
| StockSpam CYTV | 105 | 1.7% | 1.4% | 0.0% | 27.3% |
| StockSpam EXVG | 389 | 1.2% | 2.8% | 0.6% | 100.0% |
| False Positives | 911 | 0.007% | 0.0% | 0.0% | 4.6% |

# *Working with Executable Files*

```
$ cd Exercise3/
$ ./exe_match.sh
```

Start with default thresholds:

- TLSH <= 100
- Sdhash >= 1
- Ssdeep >= 1

**Exercise 3A**

For each method, find a better threshold for a match

# *Working with Executable Files*

In the paper, I suggest for executable files:

- TLSH <= 52

  Near 52:
    - about half the pairs are related,
    - about half the pairs are unrelated

- Sdhash >= 13

  Near 13:
    - about half the pairs are related,
    - about half the pairs are unrelated

- Ssdeep >= 1 (unclear what is happening between 1 -25)

# *Working with Executable Files*

## Exercise 3B

```
$ cd Exercise3
$ ./exercise3B.sh
```

There are groups of similar looking executables.

Some groups include

- debconf-*
- dpkg-*
- gnome-*

How effective are the methods at identifying these groups?

# *Working with Executable Files*

## Exercise 3C

There are some unexpected matches.

For example, are the executable pairs

uniq wc
du diff

similar?

If so, why?

If not, why not?

# *Evaluation: Random Changes*

- 500 lines of Pride and Prejudice
- 200 different version – each more different than the previous

- Random changes
  1. inserting a new word
  2. deleting an existing word
  3. swapping two words
  4. substituting a word for another word
  5. replacing 10 occurrences of a character for another character
  6. deleting 10 occurrences of a character

# Exercise 4A

pp.0 is the first 500 lines of Pride and Prejudice

```
$ cd Exercise4
$ ./pp.sh
```

starting file:        pp_changes/pp.0

Iteratively makes 500 random changes creating files
pp_changes/pp.1001
pp_changes/pp.1500

# *Exercise 4A*

pp.0reason gives an explanation of each change

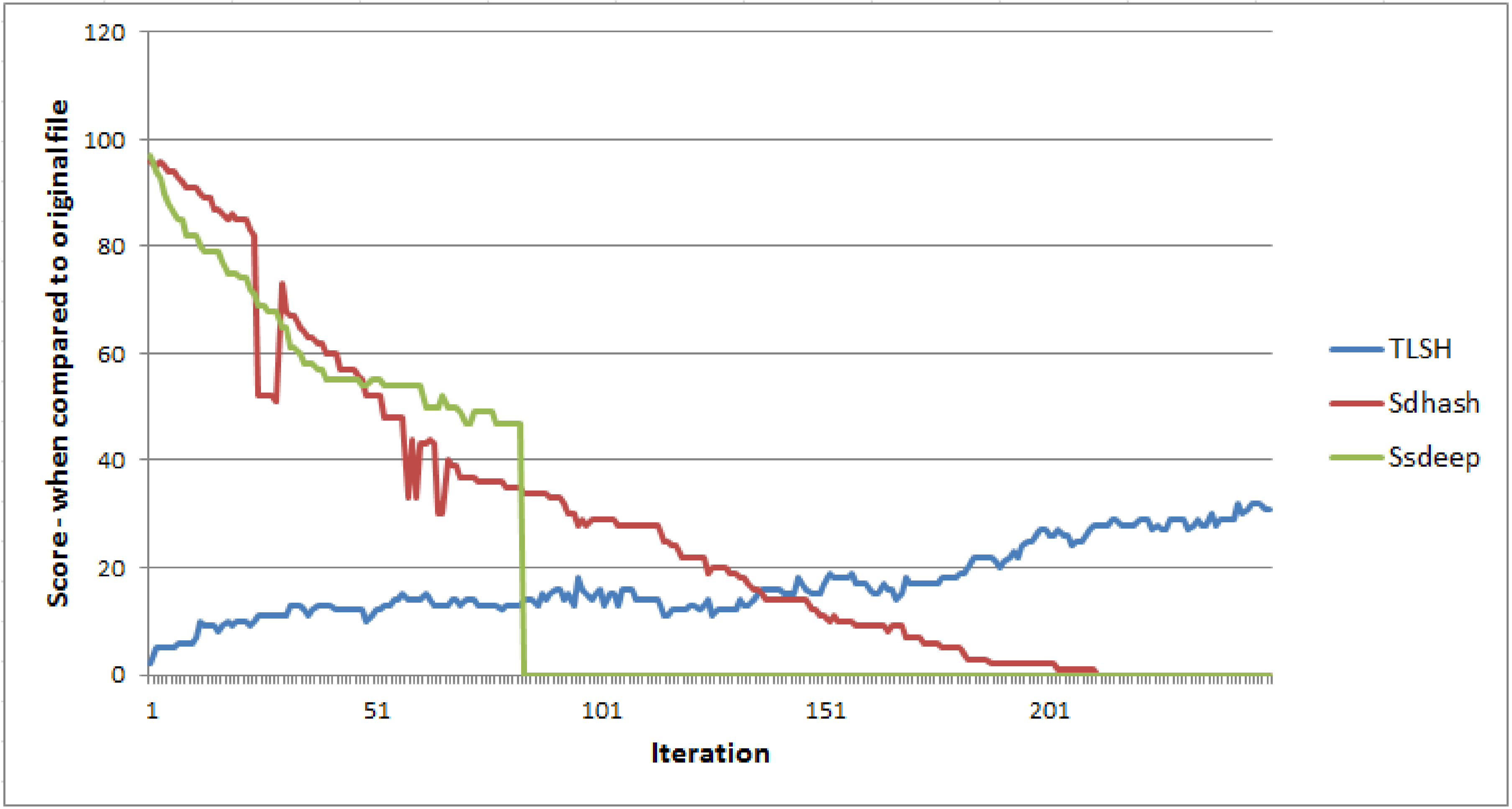
- SWAP-line 251-252 and line 451-452
- DELETE-word 'much' [pos=3882,len=4]
- SUBST-word 'when' [pos=6811,len=4] for 'him' [pos=10012,len=4]

pp.txt gives the score for each iteration compared to the original file

```
tlsh,sdhash,ssdeep
2,096,97
5,095,94
5,096,93
```

At which iteration does TLSH, Sdhash and Ssdeep break?

# Exercise 4A: 500 lines of P&P



Trend Micro TLSH documentation

# *Exercise 4A: 500 lines of P&P*

- First 500 lines after 200th iteration

```
"I hope Mr. Bingley will like it, Lizzy."

Author: Jane Austen

 Date: August 26, 2008 [EBook #1342]
Release Date: June, 1998
[Last updated: October 12, 2012]

Language: English



*** START OF THIS PROJECT GUTENBERG EBOOK PRIDE AND PREJUDICE ***
```

- Ssdeep has failed at iteration 84
- Sdhash has failed at iteration 212

# *Exercise 4B*

Copy pp.0 to pp.4B

Modify pp.4B so that

- `Ssdeep(pp.0, pp.4B) = 0`
- `Sdhash(pp.0, pp.4B) = 0`
- `TLSH  (pp.0, pp.4B) > 100`

```
$ cd Exercise4
$ mkdir 4B
$ cp pp_changes/pp.0 4B
$ cp pp_changes/pp.0 4B/pp.4B
$ vi 4B/pp.4B
$ tlsh_score 4B
$ sdhash_score 4B
$ ssdeep_score 4B
```

# Exercise 4C

Create a version of pp.0 which

- Cannot be detected by the similarity digests
- A human reader would not notice the difference

```
$ cd Exercise4
$ mkdir 4C
$ cp pp_changes/pp.0 4C
$ cp pp_changes/pp.0 4C/pp.4C
$ vi 4C/pp.4C
$ tlsh_score 4C
```

Copy pp.0 to pp.4C

Modify pp.4C so that

- `Ssdeep(pp.0, pp.4C) = 0`
- `Sdhash(pp.0, pp.4C) = 0`
- `TLSH  (pp.0, pp.4C) > 100`
- `The text is easily readable by a person`
- `$ diff -w 4C/pp.0 4C/pp.4C`
  - `Produces no output`

At this point you have “broken” the digests

# *Conclusions*

- Similarity Digests are a great starting place for quickly finding similar content
  - Might want / need to adapt approaches
  - Similarity digests are a general tool. For specific applications (images) consider specific solutions

- Need to consider thresholds for these hashes
  - Each application may needs its own threshold

- When considering you problem / your application
  - An adversary may be deliberately morphing / obfuscating the file / content
  - Consider attacking your own application

- The different Similarity Digests have different strengths
  - Complex applications may require hybrid approaches

# *Papers*

**Introduction to TLSH**

Oliver, J., Cheng, C., Chen, Y.: TLSH - A Locality Sensitive Hash. 4th Cybercrime and Trustworthy Computing Workshop, Sydney, November 2013

https://github.com/trendmicro/tlsh/blob/master/TLSH_CTC_final.pdf

**Vulnerability Paper**

Oliver, J., Forman, S., and Cheng, C.: Using Randomization to Attack Similarity Digests. ATIS 2014, November, 2014, pages 199-210.

https://github.com/trendmicro/tlsh/blob/master/Attacking_LSH_and_Sim_Dig.pdf

**Open sources on Github**

https://github.com/trendmicro/tlsh/

# *Papers*

**SdHash**

"Data fingerprinting with similarity digests"

Vassil Roussev

Sixth IFIP WG 11.9 International Conference on Digital Forensics, Hong Kong, China, January 4-6, 2010

http://roussev.net/pdf/2010-IFIP--sdhash-design.pdf

**Ssdeep**

"Identifying almost identical files using context triggered piecewise hashing"

Jesse Kornblum

Journal Digital Investigation: The International Journal of Digital Forensics & Incident Response archive

Volume 3, September, 2006 Pages 91-97

http://dfrws.org/2006/proceedings/12-Kornblum.pdf

Source code for SSDEEP: http://ssdeep.sourceforge.net/

# *Papers (cont.)*

**Nilsimsa**

Source code for Nilsimsa http://ixazon.dynip.com/~cmeclax/nilsimsa.html

"An open digest-based technique for spam detection"
E. Damiani1, S. De Capitani di Vimercati1, S. Paraboschi2, P. Samarati
Proceedings of the 2004 international workshop on security in parallel and distributed systems. 2004.
http://spdp.di.unimi.it/papers/pdcs04.pdf

**Comparison Paper**

"An evaluation of forensic similarity hashes"
Vassil Roussev
Journal Digital Investigation: The International Journal of Digital Forensics & Incident Response archive
Volume 8, August, 2011
Pages S34-S41

# *End Session*

Thank you

# *FULL DETAILS*
# *Ssdeep*

1. Use a rolling hash to split the document into segments,
2. Produce a 6 bit value for each segment by hashing the segment,
3. Concatenate the base64 encoded 6 bit values from step (2) to form the output signature.

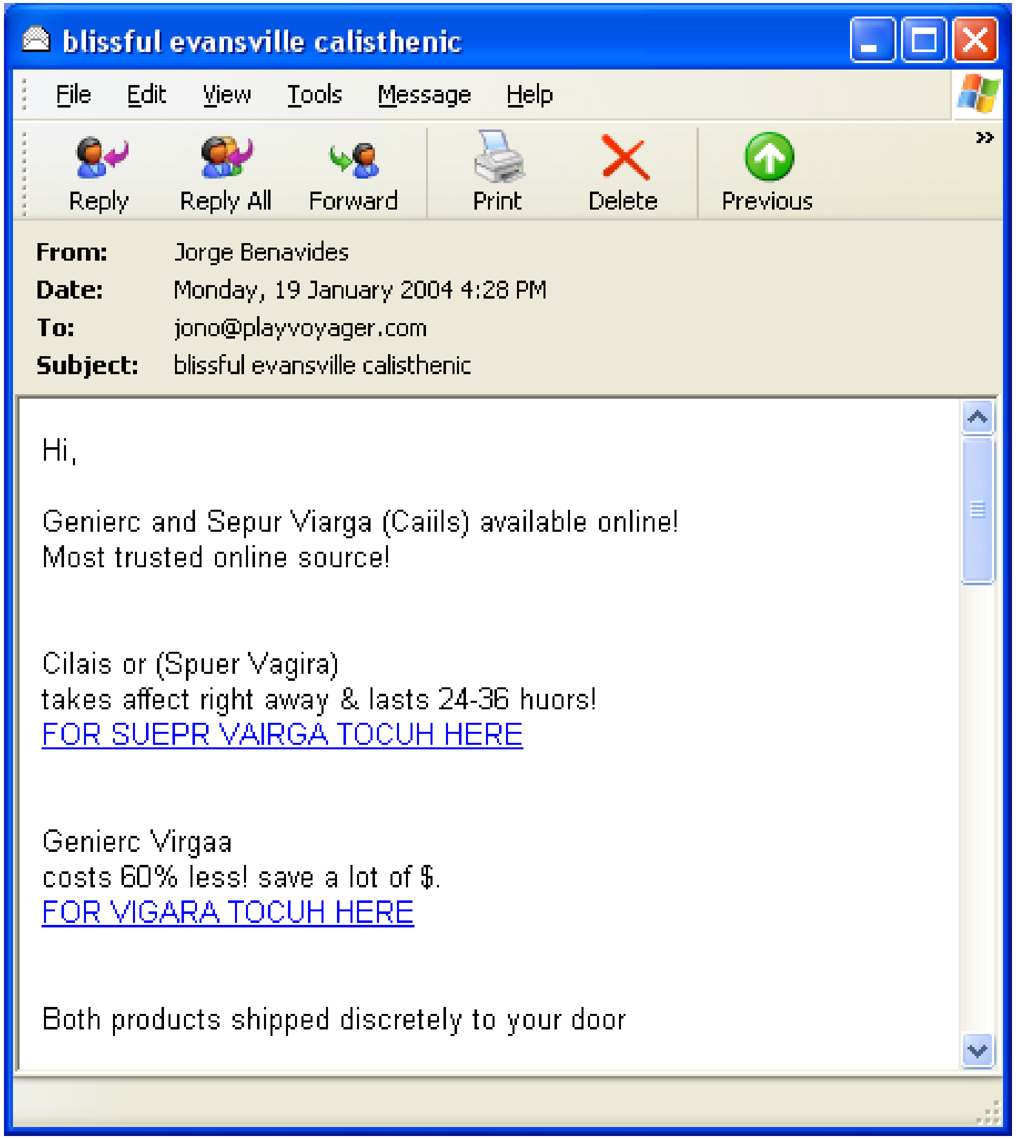
Similarity(digest1, digest2) = 100 - edit distance(digest1 and digest2)

Calculate the probability of the two strings being aligned using Lloyd Allison's Dynamic Programming Algorithm (DPA)

Ref: Dynamic Programming Algorithm for Sequence Alignment

http://www.csse.monash.edu.au/~lloyd/tildeStrings/Notes/DPA/

# FULL DETAILS
# Motivation for Using Edit Distance



**From:** Jorge Benavides
**Date:** Monday, 19 January 2004 4:28 PM
**To:** jono@playvoyager.com
**Subject:** blissful evansville calisthenic

Hi,

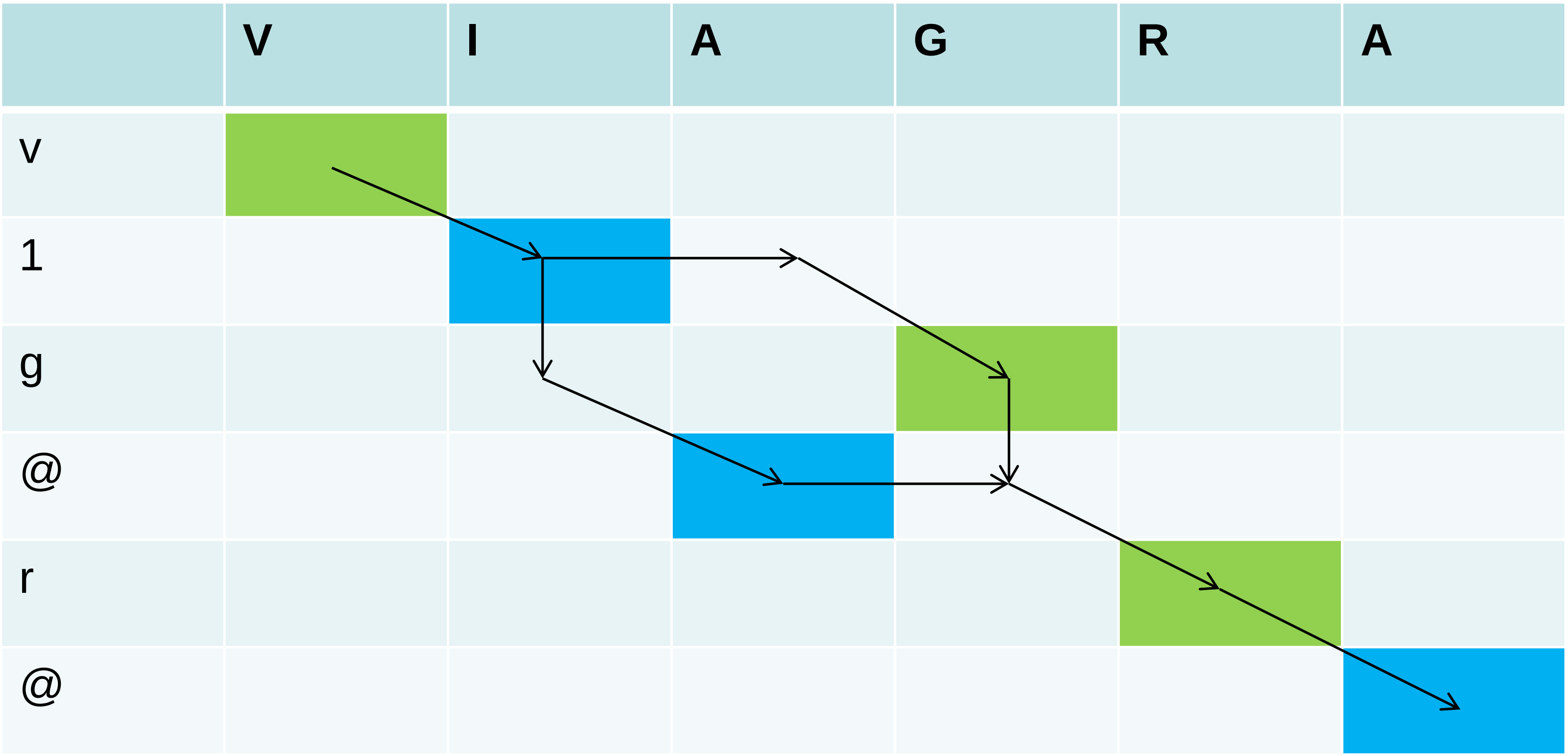
Genierc and Sepur Viarga (Caiils) available online!
Most trusted online source!

Cilais or (Spuer Vagira)
takes affect right away & lasts 24-36 huors!
FOR SUEPR VAIRGA TOCUH HERE

Genierc Virgaa
costs 60% less! save a lot of $.
FOR VIGARA TOCUH HERE

Both products shipped discretely to your door

# *FULL DETAILS*
# *Calculating Edit Distance*

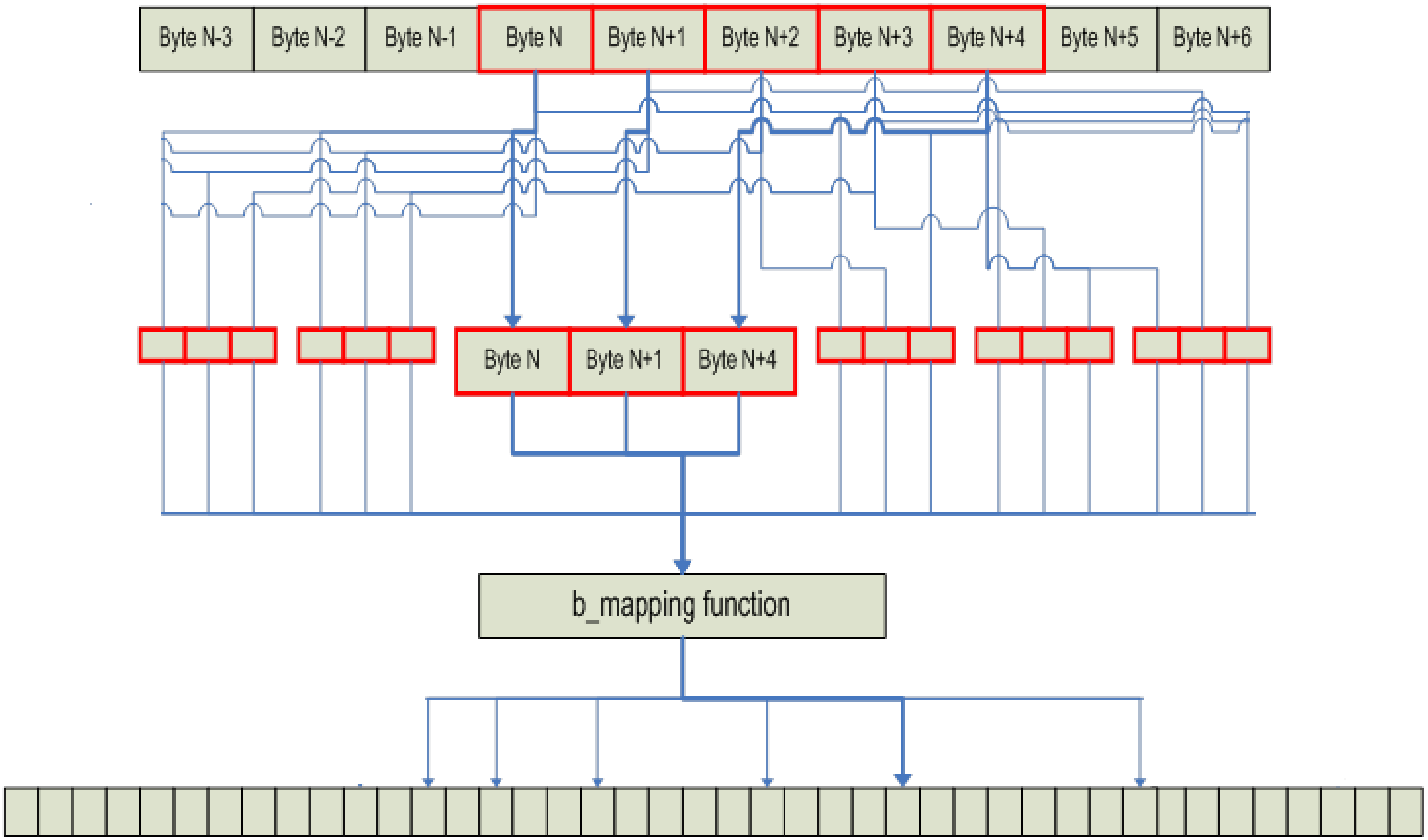


| | **V** | **I** | **A** | **G** | **R** | **A** |
|---|---|---|---|---|---|---|
| v | 🟩 | | | | | |
| 1 | | 🟦 | | | | |
| g | | | | 🟩 | | |
| @ | | | 🟦 | | | |
| r | | | | | 🟩 | |
| @ | | | | | | 🟦 |

Ref: "Using Lexigraphical Distancing to Block Spam“ Jonathan Oliver, MIT Spam Conference 2005.

# Deeper Explanation of How TLSH works

# *FULL DETAILS*
# *Algorithm to determine TLSH*



Each sliding window contains 5 bytes

Those 5 bytes form trigrams

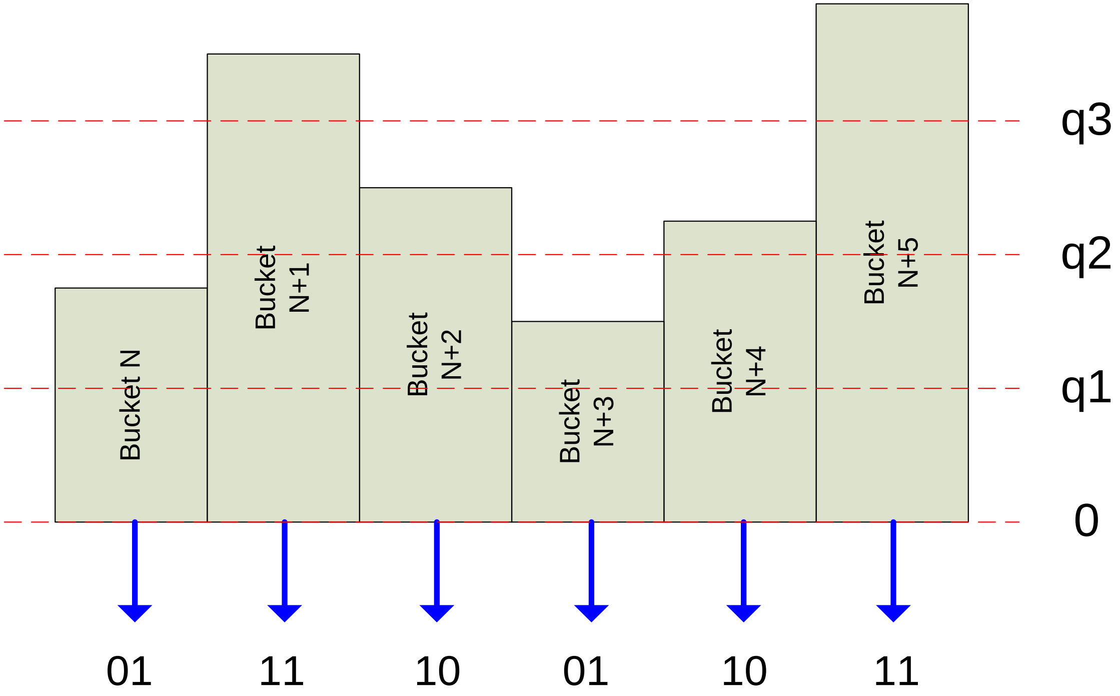
b_mapping function maps each trigram into one of the buckets

For each mapping, increase the bucket count by 1

# *FULL DETAILS*
# *Algorithm to determine TLSH*

TREND UNIVERSITY
Trend University is to Train You

- TLSH uses a 4-way to reflect the differences between different histograms
- The q2 point is at the median bucket count
- The q1 and q3 are the lower and higher quartiles respectively



Trend Micro TLSH documentation

# *FULL DETAILS*
# *The Distance Function*

- To calculate the TLSH distance score we iterate through the buckets, scoring the distance between each bucket value
- For digest1 and digest2 we have bucket values in the range 0 .. 3

```
dist = 0
for each bucket i
    diff = abs(digest1[i] - digest2[i])
    if (diff == 3)            dist += 6
    else if (diff == 2)       dist += 2
    else if (diff == 1)       dist += 1
end for
return(dist)
```

# How EXACTLY does TLSH work?

- `Login to AWS`

```
$ cd Similarity_Digests_YOURNAME
$ ./bin/tlsh_unittest_verbose -f Exercise1/chp1.txt  |
less
```

# *How EXACTLY does TLSH work?*

`Chp1.txt`

It is a truth universally acknowledged, that a single man in possession

```
Append BF BB EF to first window
Append BF BB    to second window

WINDOW       A  B  C  D  E
             I  t     i  s
WIN1   E D C
WIN2   E D B
WIN3  E C B
WIN4  E C A
WIN5  E D A
WIN6  E B A
```

# How EXACTLY does TLSH work?

- low quartile=90
- median=103
- high quartile=118

- bucket[0]=87 => Emit 00
- bucket[1]=138 => Emit 11 1100 = C
- bucket[2]=100 => Emit 01
- bucket[3]=148 => Emit 11 1101 = D
- bucket[4]=109 => Emit 10
- bucket[5]=103 => Emit 01 1010 = 6
- bucket[6]=98 => Emit 01
- bucket[7]=110 => Emit 10 1001 = 9

Header

C991C7 1FA380036685B052B9761E3E17F706C1381764C635981FA12A3332EAAC6F96DC

# Experiments

# Experiments

- Mismatch file set
  - 109 binary malware files (different malware families)
  - 290 randomly constructed HTML fragments
  - 100 pieces of random text from dictionary (no overlap)
  - 79 distinct text files about different topics

- Match file set
  - 3 malware families 20 files each family

- Random created 15 variants of each of the 79 distinct text files
  - 8766 matched file comparisons
  - 55822 different file comparisons

# *Experiments*

| TLSH | | | Nilsimsa | | | Sdhash | | | Ssdeep | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Score | FP rate | Detect | Score | FP rate | Detect | Score | FP rate | Detect | Score | FP rate | Detect |
| < 300 | 79.30% | 98.8% | > 120 | 99.86% | 100.0% | > 0 | 0.04711% | 37.1% | > 0 | 0.09966% | 31.2% |
| < 250 | 69.06% | 98.8% | > 130 | 99.20% | 100.0% | > 5 | 0.02718% | 36.6% | > 5 | 0.09785% | 31.2% |
| < 200 | 50.10% | 98.8% | > 140 | 98.11% | 100.0% | > 10 | 0.02174% | 36.1% | > 10 | 0.09603% | 31.2% |
| < 150 | 24.33% | 98.1% | > 150 | 96.98% | 100.0% | > 20 | 0.01812% | 35.4% | > 20 | 0.09422% | 31.2% |
| < 100 | 6.43% | 94.5% | > 160 | 94.26% | 100.0% | > 30 | 0.01268% | 34.4% | > 30 | 0.05617% | 30.9% |
| < 90 | 4.49% | 92.3% | > 170 | 89.52% | 100.0% | > 40 | 0.00544% | 32.7% | > 40 | 0.01812% | 29.3% |
| < 80 | 2.93% | 89.0% | > 180 | 81.38% | 100.0% | > 50 | 0.00362% | 29.7% | > 50 | 0.00362% | 27.3% |
| < 70 | 1.84% | 83.6% | > 190 | 69.69% | 99.7% | > 60 | 0.00362% | 26.0% | > 60 | 0.00362% | 25.9% |
| < 60 | 1.09% | 76.0% | > 200 | 54.45% | 98.8% | > 70 | 0.00181% | 18.8% | > 70 | 0.00181% | 23.1% |
| < 50 | 0.52% | 65.3% | > 210 | 36.73% | 96.4% | > 80 | 0.00181% | 12.4% | > 80 | 0.00000% | 16.2% |
| < 40 | 0.07% | 49.6% | > 220 | 18.29% | 91.9% | > 90 | 0.00181% | 4.6% | > 90 | 0.00000% | 8.8% |
| < 30 | 0.00181% | 32.2% | > 230 | 5.52% | 72.0% | > 99 | 0.00000% | 1.0% | > 99 | 0.00000% | 3.5% |
| < 20 | 0.00181% | 17.3% | > 240 | 1.26% | 35.2% | | | | | | |
| < 10 | 0.00181% | 6.4% | > 250 | 0.49% | 9.5% | | | | | | |