Automated Evaluation Of Approximate Matching Algorithms On Real Data

By

Frank Breitinger and Vassil Roussev

Presented At

The Digital Forensic Research Conference

DFRWS 2014 EU Amsterdam, NL (May 7th - 9th)

Automated evaluation of approximate matching algorithms on real data

Frank Breitinger & **Vassil Roussev**

CASED, Germany

frank.breitinger@cased.de

University of New Orleans

vassil@roussev.net

outline

APPROXIMATE MATCHING ALGORITHMS /AMA/

PRIOR EVALUATIONS OF AMA

- controlled data experiments
- real data experiments

GROUND TRUTH

- *longest common substring* (LCS) criterion
- *approximate* LCS

APPLICATION TO EXISTING AMA

definition

"Approximate matching is a generic term describing any technique designed to identify similarities between two digital artifacts.

In this context, an artifact (or an object) is defined as an arbitrary byte sequence, such as a file, which has some meaningful interpretation."

DRAFT NIST Special Publication 800-168
"Approximate Matching: Definition and Terminology"

bytewise AM

Treats artifacts as strings of bytes

- and attempts to establish commonality
- w/o parsing, or interpretation.

Rationale

- common representation often implies common semantics

Existing work

- `ssdeep, sdhash, mrsh, …`

Challenges

- AMAs not well understood & not easily compared

overall research goals

ESTABLISH RELIABLE AMA EVALUATION METHODOLOGY

- benchmarks
- reference data sets
- automated implementation

PERFORM EVALUATION OF EXISTING WORK

- run experiments
- characterize strong/weak points
- provide guidance to analysts

today's topic

HOW TO BRING AUTOMATION TO

REAL DATA EVALUATION STUDIES OF

BYTEWISE AMA?

flashback: controlled data studies

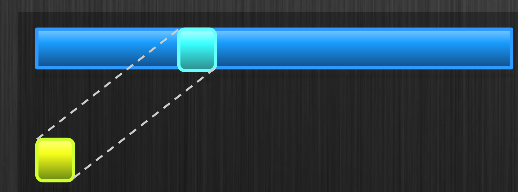
MAIN PROBLEM ➔ ***WHAT IS THE GROUND TRUTH?***

CONTROLLED DATA IDEA (2011)

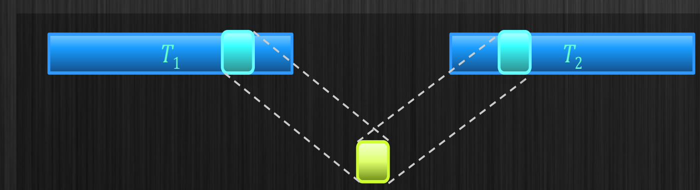
use pseudo-random data to *construct* the data sets
➔ ground truth is trivially known

EXAMPLE SCENARIOS

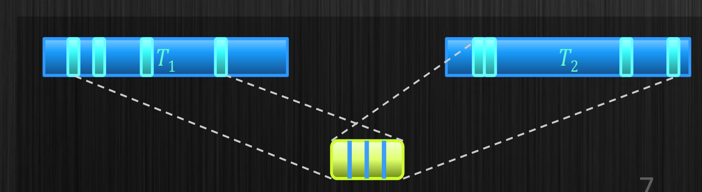
embedded object



single common block



multiple common blocks



summary: controlled data

Pro

- automation
- precise ground truth knowledge
- statistical studies
- arbitrary evaluation scenarios

Con

- not real data → use scenario different from practice

flashback: real data study

USER STUDY (2011)

- establish ground truth by manual comparison

PRO

- as close to practice as possible

CON

- not scalable
- difficult to make uniform comparisons

goal: real data + automation

Actual problem

algorithmic ground truth discovery

Approach

use *longest common substring* (LCS) as an approximation of commonality

Pro

compatible with what AMAs actually do

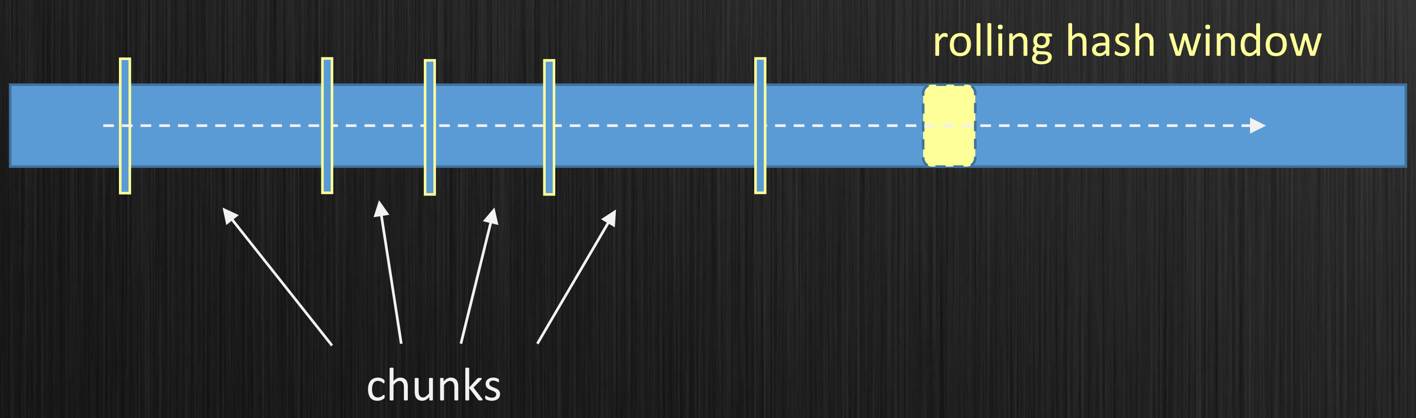
Con

quadratic complexity → not practical for larger files

new idea: use *approximate* LCS

APPROACH

- use rolling hash to break up data into variable-sized chunks (40 bytes avg)
  - this is similar to what *ssdeep* does
- hash chunks and look for the longest match
- linear complexity



ALCS: does it work?

$$d_r = \lceil 100 \times \frac{lcs(f_1,f_2) - alcs(f_1,f_2)}{\min(|f_1|,|f_2|)} \rceil, d_r \in 0, 1, \ldots, 100.$$

X	0	1	2	3	4	5	10	15	20
$Pr\{d_r = X\}$	0.8869	0.0449	0.0155	0.0040	0.0047	0.0116	0.0062	0.0001	0.0000
$Pr\{d_r \leq X\}$	0.8869	0.9318	0.9473	0.9513	0.9561	0.9677	0.9834	0.9992	0.9999

for 95%+ of files, difference is no more then 3%

experimental setup

DATA: *T5 CORPUS*

jpg	gif	doc	xls	ppt	html	pdf	txt
362	67	533	250	368	1093	1073	711

BASELINE SCENARIO

\# *threshold* = 0

➔ *any* positive comparison result is counted

some notation

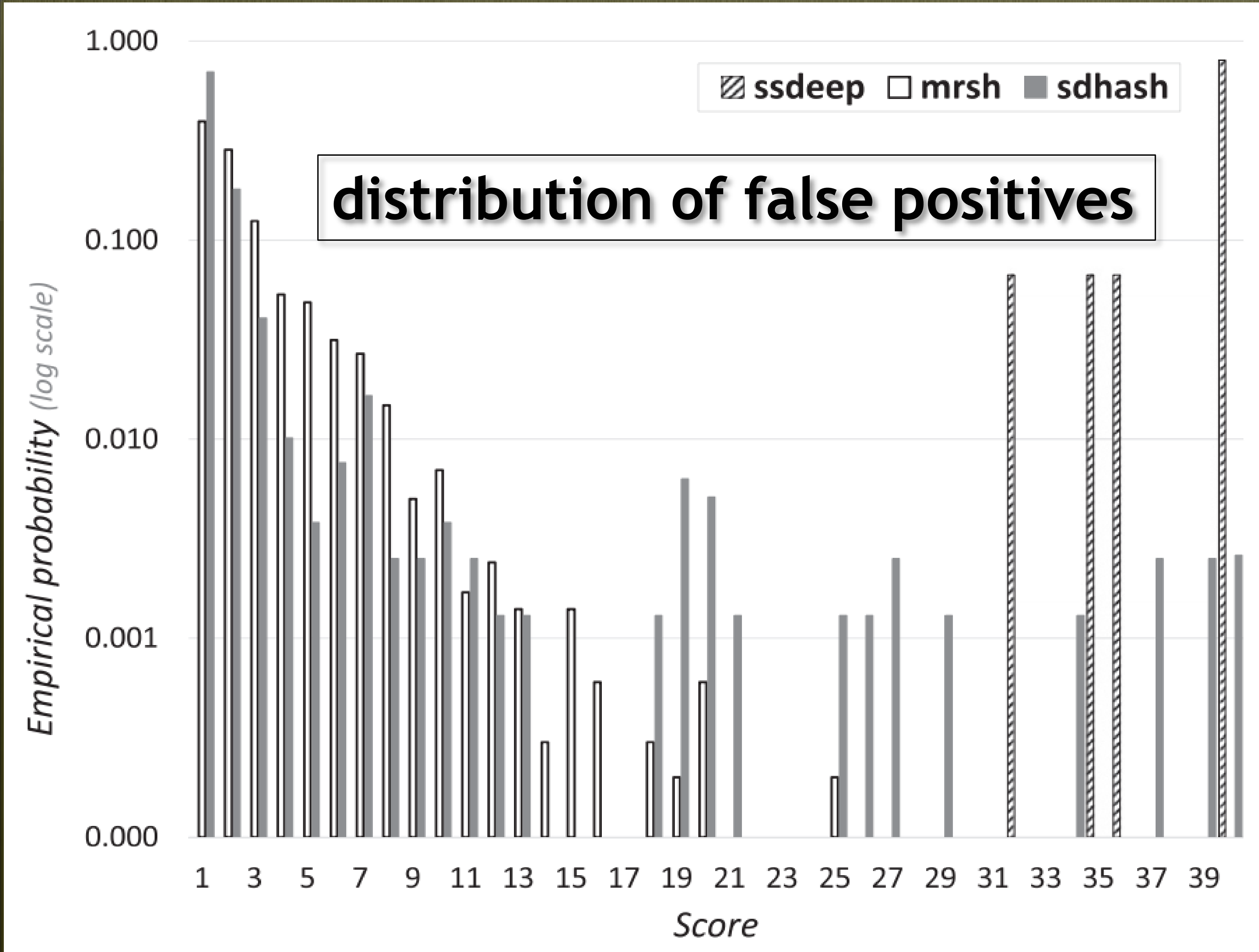
$$L_a = alcs_a(f_1, f_2), \text{ where } 0 \le L_a \le \min(|f_1|, |f_2|).$$

$$L_r = \lceil 100 \times \frac{L_a}{\min(|f_1|, |f_2|)} \rceil, \text{ where } 0 \le L_r \le 100.$$

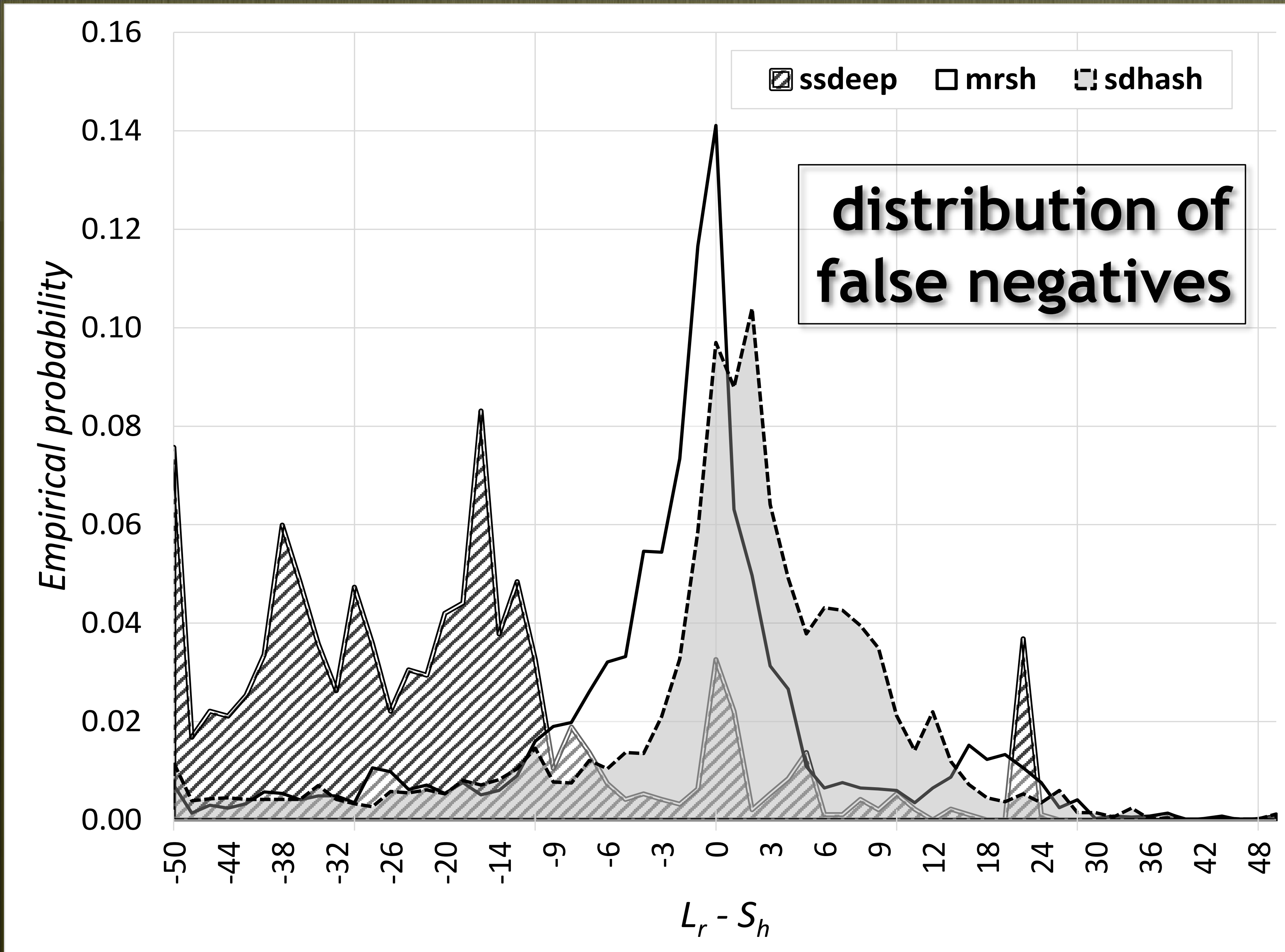
$$TP_{alcs}(f_1, f_2) \equiv L_a \ge 100 \wedge L_r \ge 1.$$

baseline results

	ssdeep	mrsh-v2	sdhash
TP	951	3679	5474
FP	15	23,453	790
TN	9,472,047	9,448,609	9,471,272
FN	457,183	454,455	452,660
Precision	0.98447	0.13560	0.87388
Recall	0.00010	0.00039	0.00058
TNR	1.00000	0.99752	0.99992
Accuracy	0.95396	0.95187	0.95434
F_1	0.00020	0.00078	0.00115
F_2	0.00013	0.00049	0.00072
$F_{0.5}$	0.00050	0.00192	0.00288
MCC	0.04412	0.02232	0.09913



distribution of false positives
ssdeep
mrsh
sdhash
1.000
0.100
0.010
0.001
0.000
Empirical probability (log scale)
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Score



ssdeep
mrsh
sdhash
distribution of false negatives
Empirical probability
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
-50
-44
-38
-32
-26
-20
-14
-9
-6
-3
0
3
6
9
12
18
24
30
36
42
48
L_r - S_h

similarity: containment vs. resemblance

Containment query

- small vs. larger object
- interpretation
  - » does the larger object contain (trace of) the smaller one?

Resemblance query

- two peer object (~ same size)
- interpretation
  - » what is the level of commonality b/w these two objects?

For this study

- $|f_1| >= 2|f_2|$ ➔ containment; otherwise, resemblance

SAMPLES

	TP	TP_{ratio}	TN	TN_{ratio}	Total	$Total_{ratio}$
gt-con	354,914	0.775	7,382,141	0.779	7,737,055	0.779
gt-res	103,220	0.225	2,089,921	0.221	2,193,141	0.221
gt	458,134	1.000	9,472,062	1.000	9,930,196	1.000

RESULTS BY SIMILARITY SCENARIO

	Precision	Recall	F_1	F_2	$F_{0.5}$	MCC
ssdeep-con	0.93671	0.00001	0.00002	0.00001	0.00005	0.01361
ssdeep-res	0.98873	0.00042	0.00084	0.00052	0.00209	0.08944
ssdeep	0.98447	0.00010	0.00020	0.00013	0.00050	0.04412
mrsh-con	0.12647	0.00030	0.00060	0.00038	0.00149	0.01834
mrsh-res	0.15217	0.00070	0.00140	0.00088	0.00345	0.03297
mrsh	0.13560	0.00039	0.00078	0.00049	0.00192	0.02232
sdhash-con	0.87478	0.00047	0.00094	0.00059	0.00235	0.08976
sdhash-res	0.87233	0.00096	0.00191	0.00120	0.00476	0.12612
sdhash	0.87388	0.00058	0.00115	0.00072	0.00288	0.09913

conclusions

INTRODUCED NEW APPROACH FOR AUTOMATED TESTING OF AMA ON *REAL DATA*

PROPOSED AN ANALYTICAL FRAMEWORK FOR QUANTIFYING AMA PERFORMANCE

ANALYZED EXISTING AMA

- recall rates are low
  - i.e., absence of proof ***definitely*** not proof of absence
- precision rates for ssdeep & sdhash are high
  - i.e., positive results are a strong hint
- overall, *sdhash* does best

caveats & future work

THIS IS JUST A FIRST STEP; NEED TO

- make a bigger study with more artifacts
- study correlation of commonality and user-observable similarity
  - i.e., does the user see the commonality?
- control for sparse data
  - e.g., long strings of zeroes
- make this part of a tool (FRASH)

thank you!

Q & A