Data and Network Security

(Master Degree in Computer Science and Cybersecurity)

Lecture 9

SAPIENZA
UNIVERSITÀ DI ROMA

Outline for today

- Literature Analysis Info
- Recap last lecture
- Digital forensics

Examination

The final grade is calculated as follows:

- 45% Literature Analysis and active participation
- 45% Written research report
- 10% Active participation to Question & Answer section.

***If for any reason students do not turn in most of the required above tests, then those students will be required to take an oral exam on the entire course programme (TDB)**

The weight of this final oral exam = 100%.

Literature analysis

- Choosing the topic
- Choosing the third paper
- Studying the works

Choosing the third paper

*Last 5 years is preferrable

Categories > Engineering & Computer Science > **Computer Security & Cryptography**

	Publication	h5-index	h5-median
1.	IEEE Symposium on Security and Privacy	98	152
2.	IEEE Transactions on Information Forensics and Security	98	152
3.	ACM Symposium on Computer and Communications Security	93	149
4.	USENIX Security Symposium	92	153
5.	Computers & Security	89	131
6.	Network and Distributed System Security Symposium (NDSS)	78	133
7.	IEEE Transactions on Dependable and Secure Computing	69	117
8.	International Conference on Theory and Applications of Cryptographic Techniques (EUROCRYPT)	63	92
9.	International Cryptology Conference (CRYPTO)	59	96
10.	Journal of Information Security and Applications	57	81
11.	IACR Transactions on Cryptographic Hardware and Embedded Systems	52	81
12.	Security and Communication Networks	51	76
13.	International Conference on Financial Cryptography and Data Security	46	95
14.	IEEE European Symposium on Security and Privacy	45	72
15.	International Conference on The Theory and Application of Cryptology and Information Security (ASIACRYPT)	42	61
16.	ACM Asia Conference on Computer and Communications Security	38	54
17.	Symposium On Usable Privacy and Security	37	55
18.	IEEE Security & Privacy	37	51
19.	Computer Security Applications Conference	36	61
20.	IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)	35	55

Choosing the third paper

*Last 5 years is preferrable

Categories > Engineering & Computer Science > **Artificial Intelligence**

	Publication	h5-index	h5-median
1.	Neural Information Processing Systems	309	503
2.	International Conference on Learning Representations	303	563
3.	International Conference on Machine Learning	254	463
4.	AAAI Conference on Artificial Intelligence	212	344
5.	Expert Systems with Applications	148	221
6.	IEEE Transactions on Neural Networks and Learning Systems	145	223
7.	IEEE Transactions On Systems, Man And Cybernetics Part B, Cybernetics	144	199
8.	Neurocomputing	135	202
9.	International Joint Conference on Artificial Intelligence (IJCAI)	133	197
10.	Applied Soft Computing	126	174
11.	Knowledge-Based Systems	120	169
12.	Neural Computing and Applications	117	157
13.	IEEE Transactions on Fuzzy Systems	113	172
14.	The Journal of Machine Learning Research	106	181
15.	Artificial Intelligence Review	91	152
16.	International Conference on Artificial Intelligence and Statistics	91	148
17.	Neural Networks	91	139
18.	Engineering Applications of Artificial Intelligence	87	132
19.	Applied Intelligence	78	119
20.	Conference on Robot Learning	76	133

Choosing the third paper

*Last 5 years is preferrable

Categories > Engineering & Computer Science > **Databases & Information Systems**

	Publication	h5-index	h5-median
1.	International World Wide Web Conferences (WWW)	106	169
2.	IEEE Transactions on Knowledge and Data Engineering	99	180
3.	ACM SIGIR Conference on Research and Development in Information Retrieval	90	138
4.	Information Processing & Management	83	135
5.	ACM International Conference on Information and Knowledge Management	79	122
6.	International Conference on Very Large Databases	79	104
7.	ACM International Conference on Web Search and Data Mining	75	129
8.	ACM SIGMOD International Conference on Management of Data	71	101
9.	Journal of Big Data	70	147
10.	International Conference on Data Engineering	62	93
11.	International Conference on Web and Social Media (ICWSM)	59	82
12.	IEEE International Conference on Big Data	53	94
13.	ACM Conference on Recommender Systems	49	91
14.	Knowledge and Information Systems	48	85
15.	Workshop of Cross-Language Evaluation Forum	44	64
16.	World Wide Web	44	58
17.	ACM Transactions on Intelligent Systems and Technology (TIST)	42	73
18.	IEEE Transactions on Big Data	42	72
19.	Information Systems	42	65
20.	Semantic Web	41	68

Study the works - 3 pass approach

1. The first pass is a quick scan to get a bird's-eye view of the paper.

2. In the second pass, read the paper with greater care, but ignore details such as proofs. It helps to write down the key points, or to make comments in the margins, as you read.

3. The key to the third pass is to attempt to virtually re-implement the paper: that is, making the same assumptions as the authors, re-create the work. By comparing this re-creation with the actual paper, you can easily identify not only a paper's innovations, but also its hidden failings and assumptions.

Detailed guideline Link

Presentation

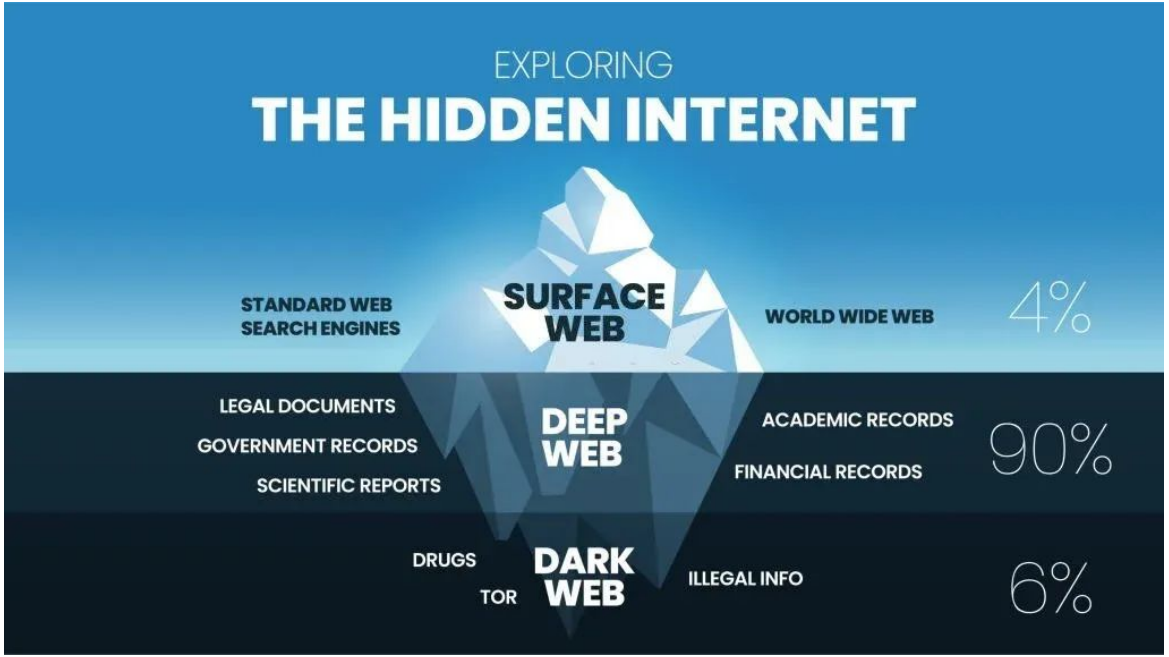
- Prepare a 20 minute presentation (divide your work equally and efficiently)
- 10 minutes of questions (from your colleagues, this counts on their 10% of active participation also)

Outline for today

- Literature Analysis Info
- Recap last lecture
- Digital forensics

Layers of the web

- The surface web
- The deep web
- The dark web



Surface web

- Part of the internet that most of use use every day. (socials etc)
- Accessible through regular browsers (firefox, chrome, safari)
- Anytime/anywhere as long as you have internet access

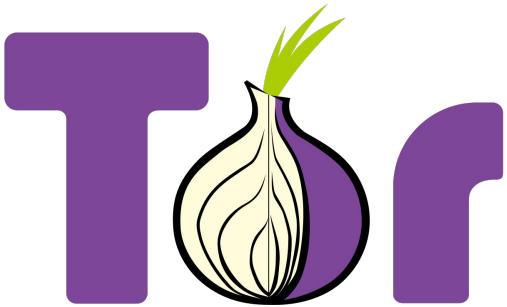
Deep web

- Refers to part of the internet that is behind “closed doors”
- Accessible only from a group of people within an organization.

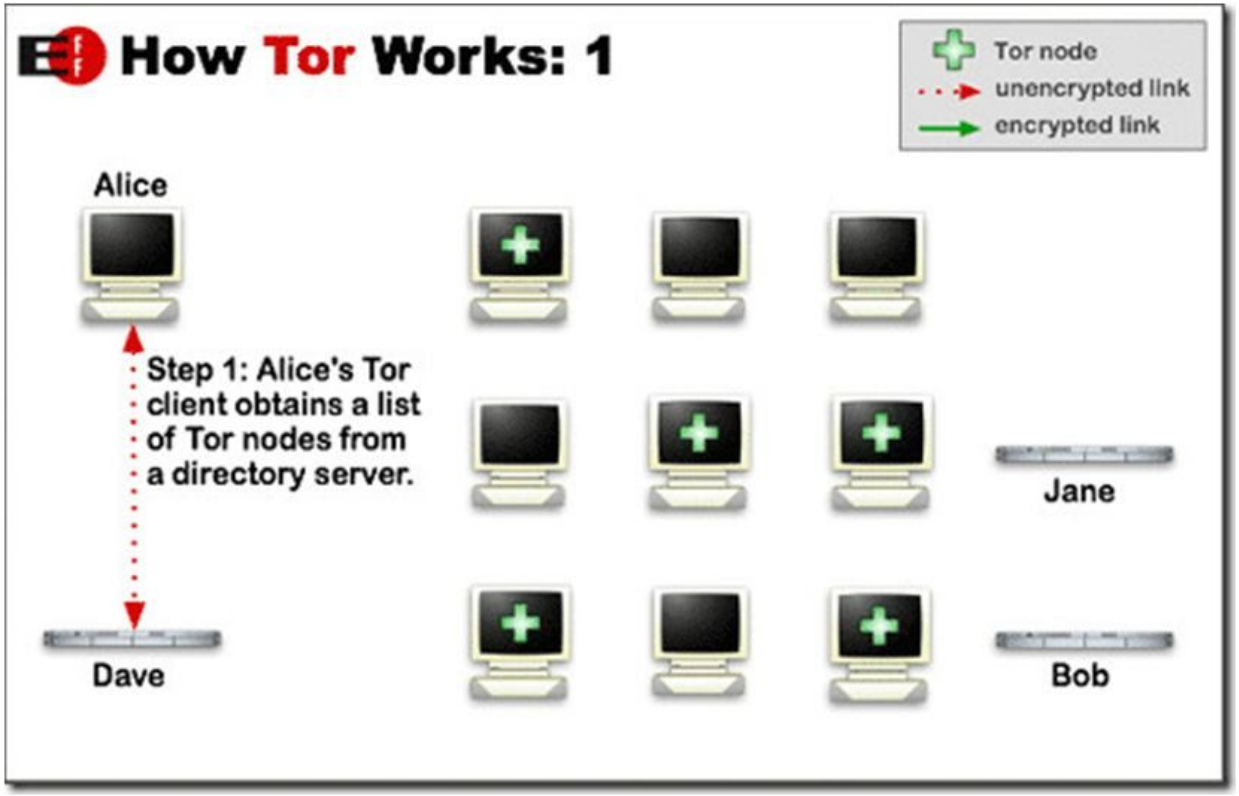
- Credentials and permissions needed to access.
- Information can not be found by search engines

Dark web

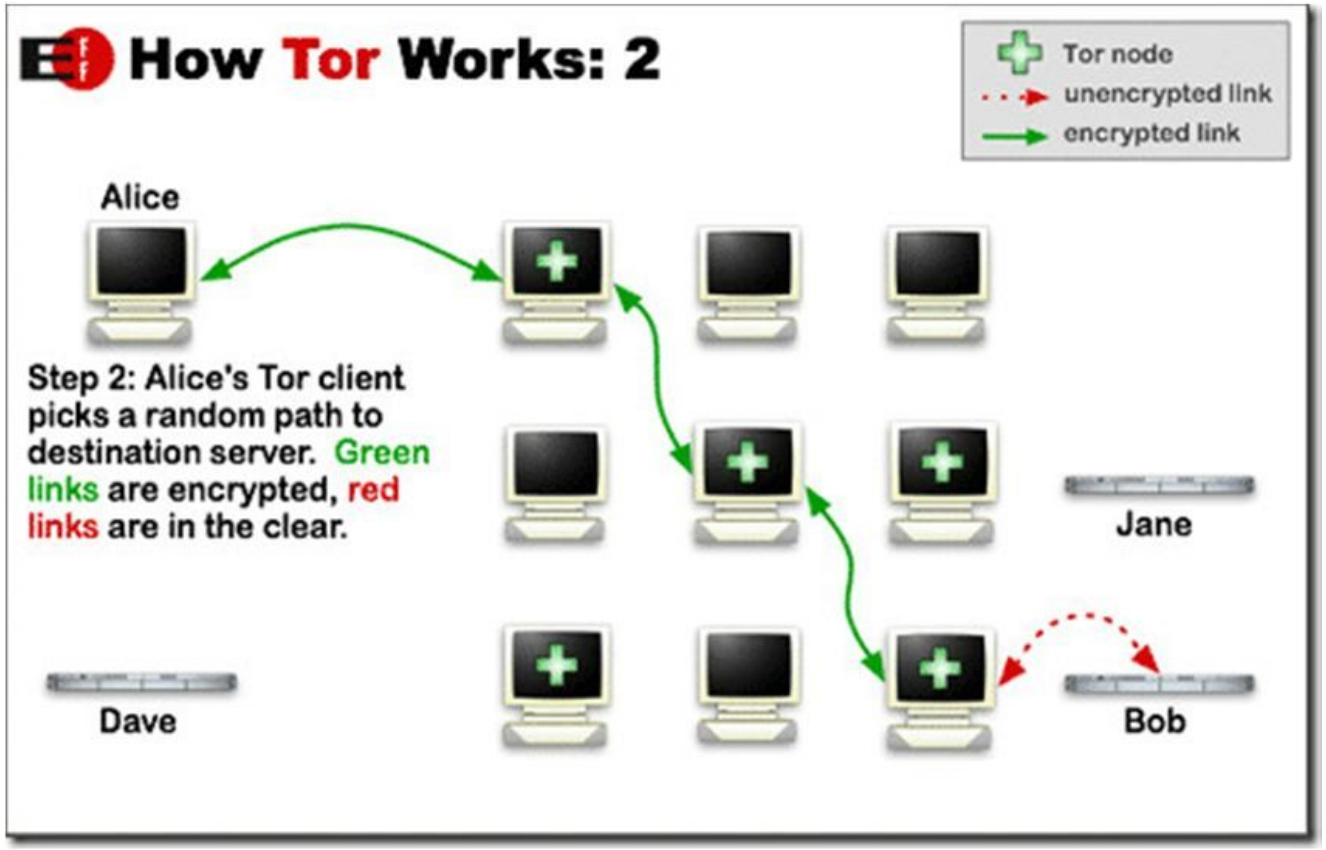
- Accessible through the use of special browsers (e.g. Tor)
- Unregulated part of the internet
- No organization, business or government is in charge of the dark web or is able to enforce rules/policies



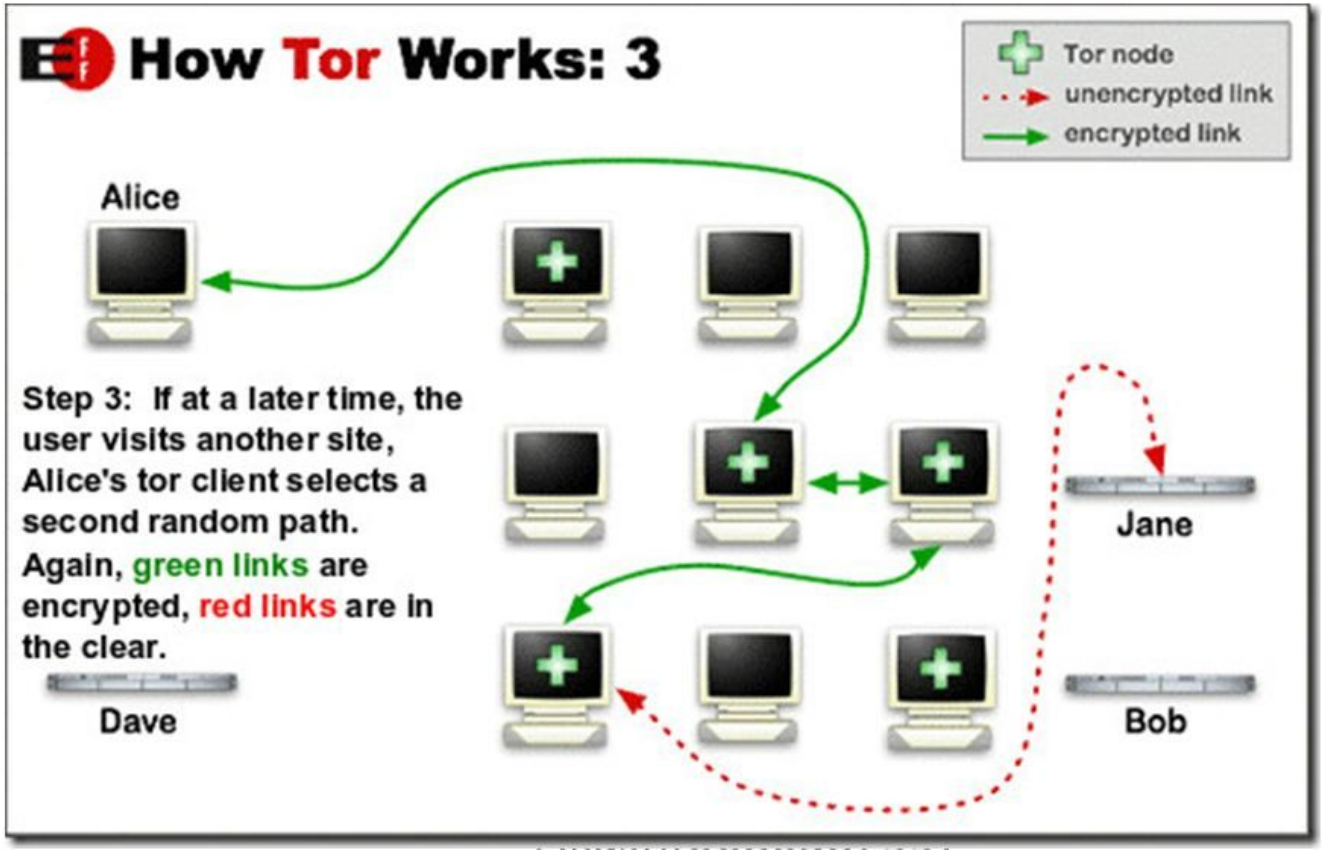
Tor in a nutshell



Tor in a nutshell



Tor in a nutshell



Geolocation-Based Profiling in the DarkWeb

End goal: To geolocate anonymous crowds of the DarkWeb forums

We will see:

- how to decompose the global profile of posting of the Dark Web forum into components that uncover the geographical origin of the crowd.

- detecting the native language of anonymous Dark Web users, starting from their posts in English

User profile

Profile P_u

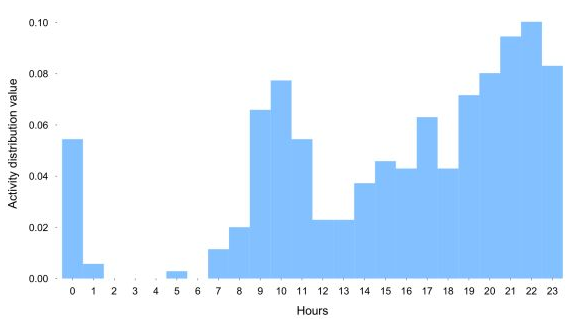
- Represented by an array of 24 elements, one per hour.
- $P_u[h]$, $h \in \{0, \ldots, 23\}$, is the fraction of daily online posting activity done by user u during hour h.

- $a_u(d,h)$ indicates whether user u has posted in the h^{th} hour of day d.

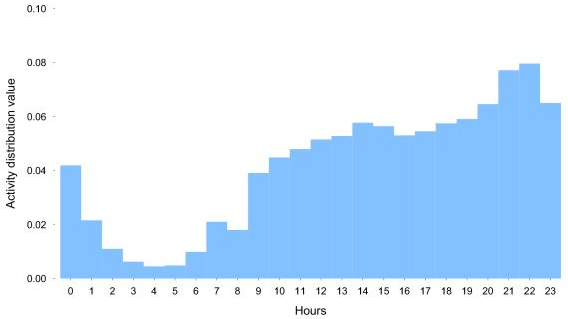
Profile P_u is the distribution of user u activity throughout the day on the target forum.

$$P_u = \{P_u[h] | h \in \{0, \ldots, 23\}, P_u[h] = \frac{\sum_d a_u(d,h)}{\sum_{d,h'} a_u(d,h')}\}$$

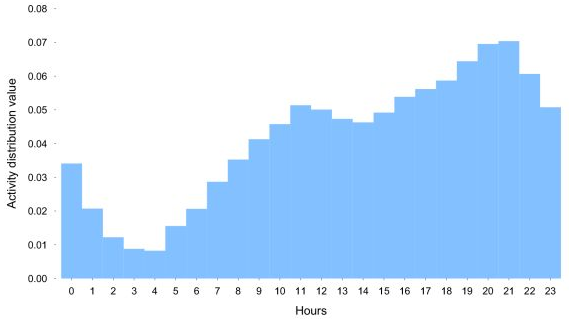
A users profile



(a) Profile of a German user.



(b) Twitter dataset of the German population (local time UTC + 1).

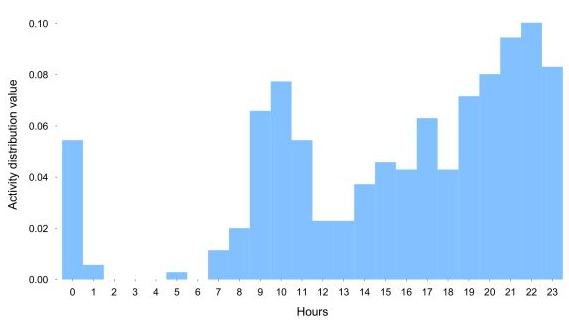


(c) Entire Twitter dataset (UTC).

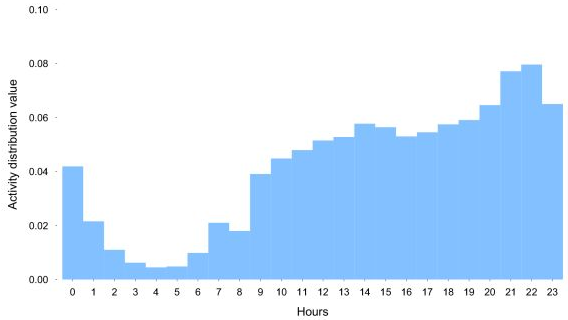
In both profiles we can easily distinguish the night as the hours of lower activity (the interval between 1:00 (1am) and 7:00 (7am)).

The activity of the German user has a first peak in the morning, drops during lunch time, and starts to grow again from the early afternoon to the evening, following a typical daily rhythm.

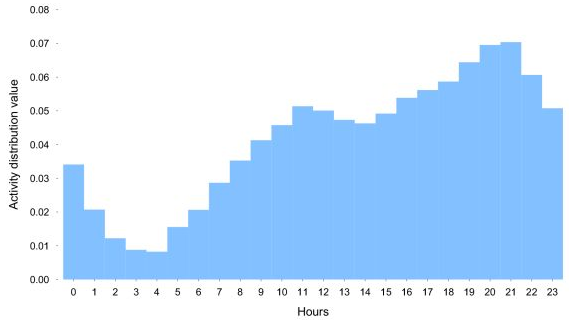
Correlation between countries/states



(a) Profile of a German user.



(b) Twitter dataset of the German population (local time UTC + 1).

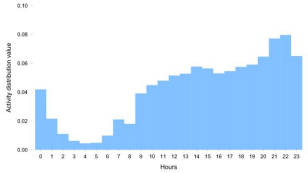


(c) Entire Twitter dataset (UTC).

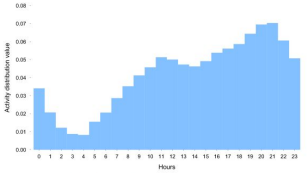
Profile of large crowds coming from different timezones are brought to the local Timezone (UTC), their profiles are almost identical.

- **use the general profile as the common baseline, properly shifted to the right timezone.**

Placing anonymous users to time zones



(b) Twitter dataset of the German population (local time UTC + 1).



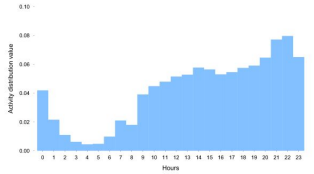
(c) Entire Twitter dataset (UTC).

Rationale: Users of the same region typically have a profile that is very close to that of the corresponding time zone crowd, and further away from crowds of different timezones.

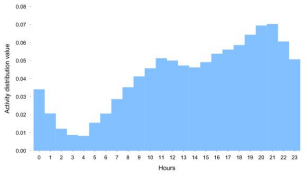
- For every member of an anonymous crowd, we compare his profile with that of all different timezone profiles
- geolocate that member to the timezone whose activity profile is **less distant**

Less Distant ->The one for which it takes less effort to transform the single user profile into by both shifting and moving probability mass.

Distance measure



(b) Twitter dataset of the German population (local time UTC + 1).



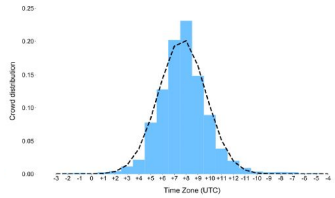
(c) Entire Twitter dataset (UTC).

Earth Mover's distance:

Given two distributions of earth mass spread on the same space, the EMD measures the least amount of work to move earth around so that the first distribution matches the second.

Less Distant ->The one for which it takes less effort to transform the single user profile into by both shifting and moving probability mass.

Single Country placement - Conclusion



(c) Malaysian

Observation:

- The previously mentioned behaviour is seen in all countries considered in the twitter dataset

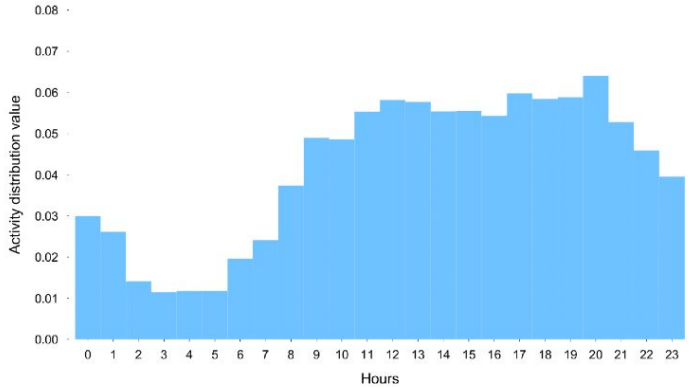
Conclusion:

To geolocate a given crowd of people from the same, unknown region, it is enough to build the corresponding activity profiles placement through the EMD distance and curve-fit the resulting distribution with a Gaussian.

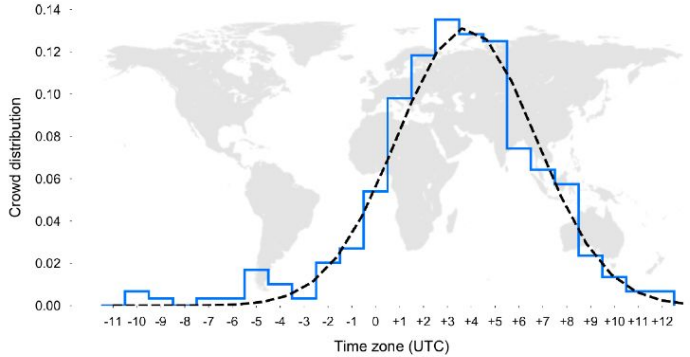
The center of the Gaussian will uncover the timezone of the unknown region and thus the geolocation of the crowd.

Geolocation in action - The Darkweb

- CRD Club, is mostly in Russian,
- Italian DarkNet Community (IDC) is the forum of the homonymous Italian marketplace in the Dark Web.



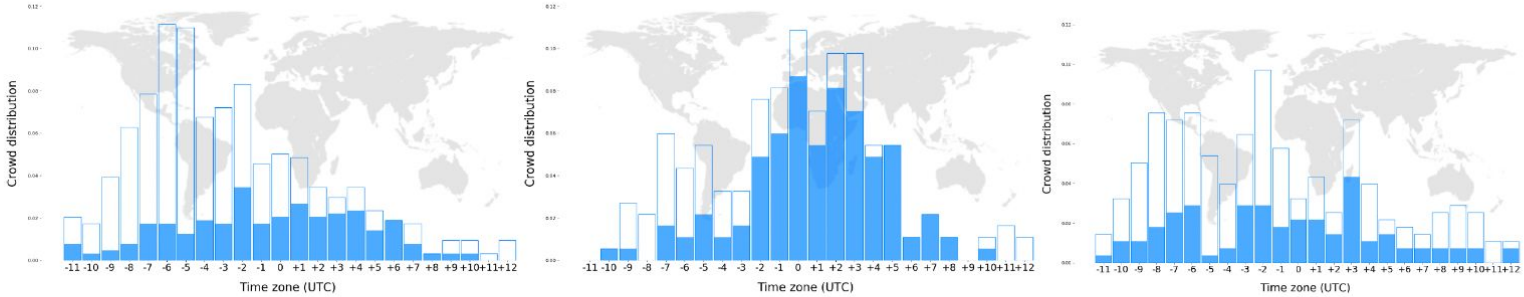
(a) Regional profile (UTC + 3).



(b) Gaussian distribution.

Geolocating unknown DarkWeb communities

- Not native vs native english speakers distribution?



(a) The Majestic Garden forum. (b) The Dream Market forum. (c) The PedoSupport Community forum.

Outline for today

- Literature Analysis Info
- Recap last lecture
- Digital forensics

Digital Forensics

Definition: branch of forensic science that deals with the identification, preservation, examination, analysis, and presentation of digital evidence derived from electronic devices and digital media.

Goal: To explain current state of a digital artifact

Digital Forensics - Why?

- Legal cases
- Data recovery
- System analysis post-compromise
- Evidence gathering
- Debugging
- …

Identification of encrypted file fragments

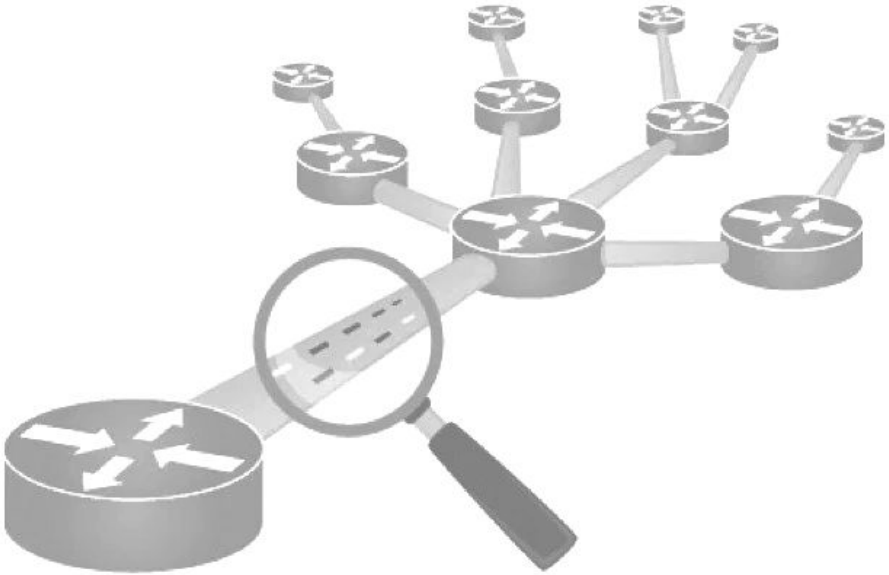
- **Ransomware detection**
- Digital forensics
- Network traffic analysis

Identification of encrypted file fragments

- Ransomware detection
- Digital forensics
- Network traffic analysis

Identification of encrypted file fragments

- Ransomware detection
- Digital forensics
- **Network traffic analysis**



Entropy estimation is not enough!

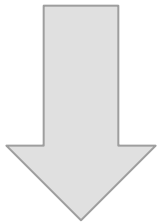
Most data types (e.g., text, images, audio)
are information-rich and highly structured

=> low entropy!

Entropy estimation is not enough!

~~Most data types (e.g., text, images, audio) are information-rich and highly structured~~

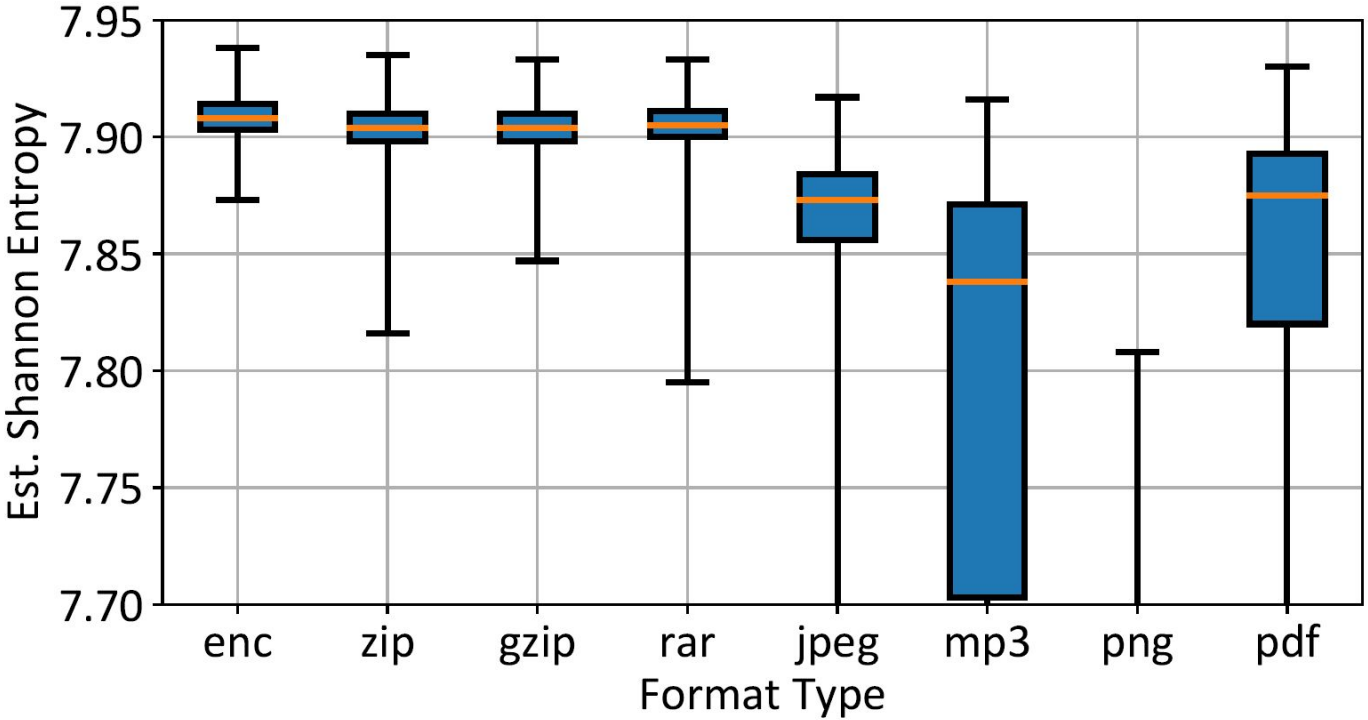
~~=> low entropy!~~



Modern CPUs can efficiently decompress data for processing, and compress it back for storage or transmission!

=> Most data formats use compression!

Entropy estimation is not enough!



NIST SP800-22 & χ^2 Test

- The NIST SP800-22 describes a suite of tests to evaluate the quality of random number generators:
  - ➢ 15 distinct tests, which analyze various structural aspects of a byte sequence.
  - ➢ Commonly employed as a benchmark for distinguishing compressed and encrypted content.

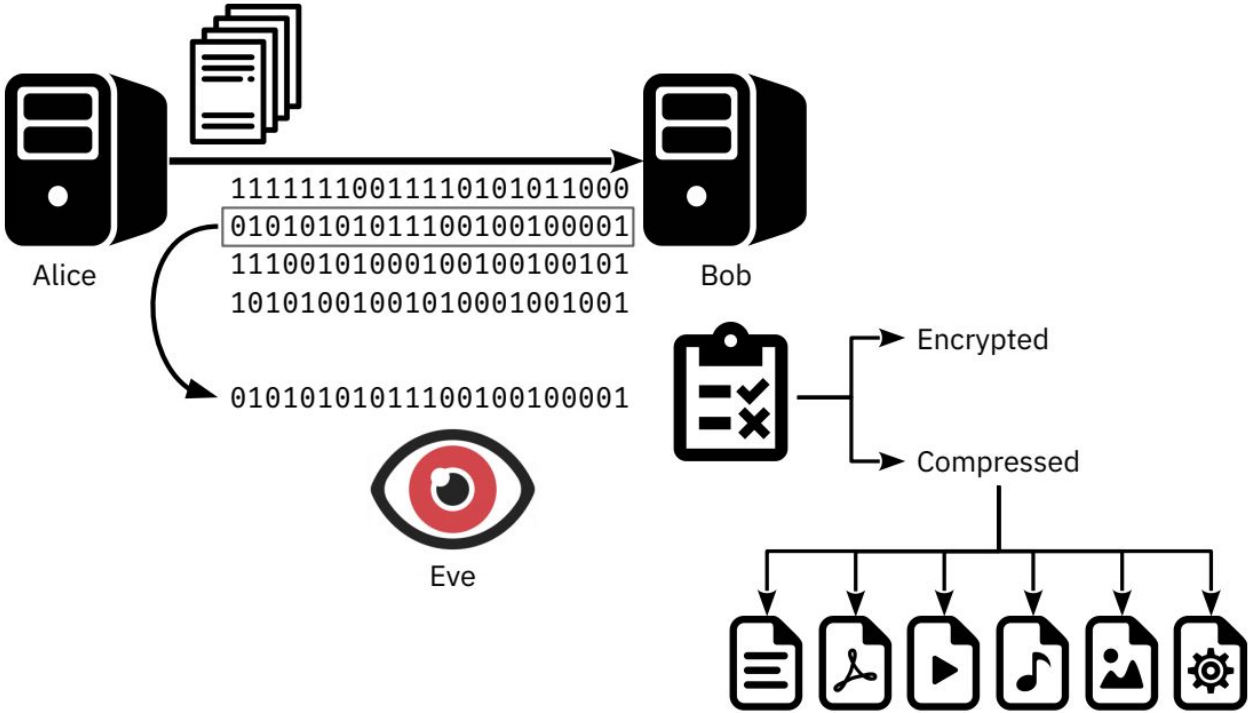
- The χ^2 Test is a simple statistical test to measure goodness of fit.
  - ➢ It has been widely applied to distinguish compressed and encrypted content.

Combine what we know?

Simultaneously incorporate *three methods* to distinguish between compressed and encrypted fragments:

- χ^2 test with a threshold.
- χ^2 confidence interval.
- Subset of NIST 800-22:
  - ➢ frequency within block test
  - ➢ cumulative sums test
  - ➢ approximate entropy test

Feature selection



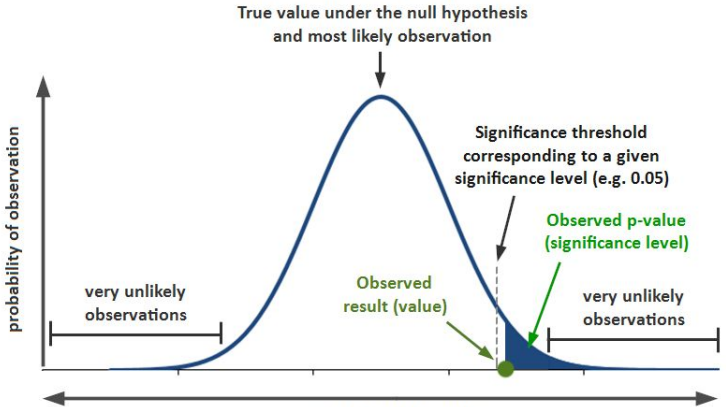
Feature selection

Method	Features
Chi square test	Absolute value
Chi square test	$\chi\%$ of confidence
NIST SP 800-22	Aggregate number of failed blocks

- A test of goodness of fit establishes whether an observed frequency distribution differs from a theoretical distribution.
- A test of homogeneity compares the distribution of counts for two or more groups using the same categorical variable
- A test of independence assesses whether observations consisting of measures on two variables are independent of each other.

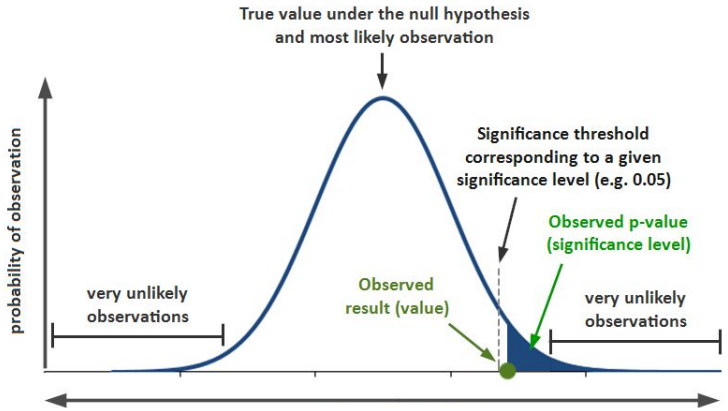
Feature selection

Method	Features
Chi square test	Absolute value
Chi square test	χ% of confidence
NIST SP 800-22	Aggregate number of failed blocks

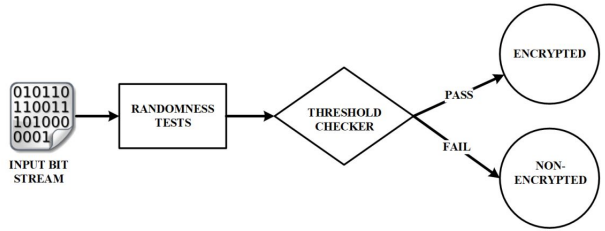


Feature selection

Chi abs. val.	$\chi\%$	NIST SP 800-22
$x \in AVG \pm \sigma$	$\chi\% > 99\% \parallel \chi\% < 1\%$	0 fails

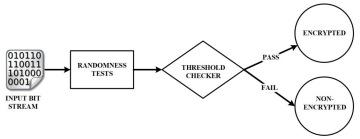


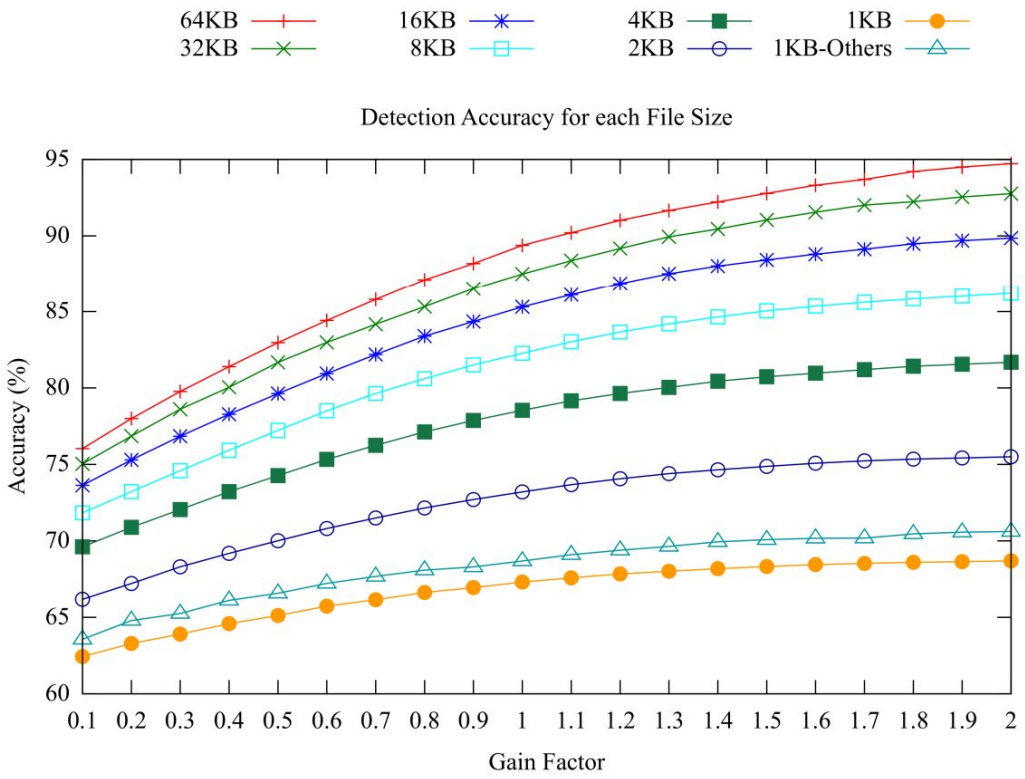
The interval values on different size



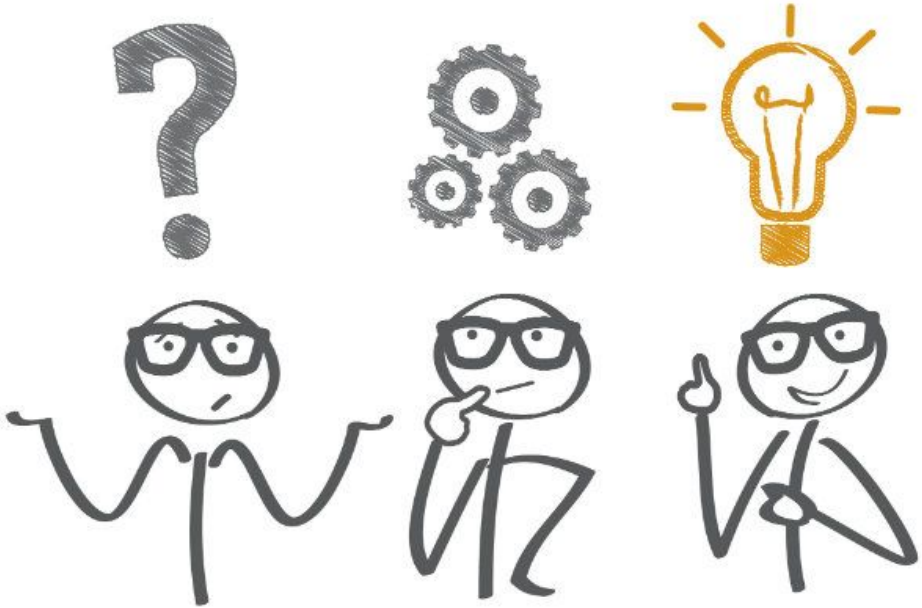
KB	Chi-square abs. val.	$\chi\%$	NIST SP 800-22
64	255.37 ± 22.82	$x > 99 \parallel x < 1$	0 fails
32	255.08 ± 22.68	$x > 99 \parallel x < 1$	0 fails
16	254.96 ± 22.76	$x > 99 \parallel x < 1$	0 fails
8	255.09 ± 22.54	$x > 99 \parallel x < 1$	0 fails
4	255.04 ± 22.60	$x > 99 \parallel x < 1$	0 fails
2	254.98 ± 22.57	$x > 99 \parallel x < 1$	0 fails
1	255.02 ± 22.57	$x > 99 \parallel x < 1$	0 fails

Performance





Issues?



Encryption/Compression Distinguisher

Tests based on byte-value distribution can distinguish some encrypted and compressed content, ***but have accuracy issues.***

Encryption/Compression Distinguisher

Tests based on byte-value distribution can distinguish some encrypted and compressed content, ***but have accuracy issues.***

WHY?

Encryption/Compression Distinguisher

Tests based on byte-value distribution can distinguish some encrypted and compressed content, ***but have accuracy issues.***

WHY?

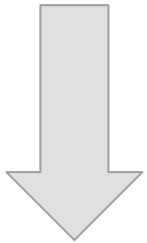
These tests lose information about the shape of the data *distribution*!

Encryption/Compression Distinguisher

Tests based on byte-value distribution can distinguish some encrypted and compressed content, ***but have accuracy issues.***

WHY?

These tests lose information about the shape of the data *distribution*!



Deep Neural Networks can consider the entire discrete distribution and can learn to recognize complex distributions!

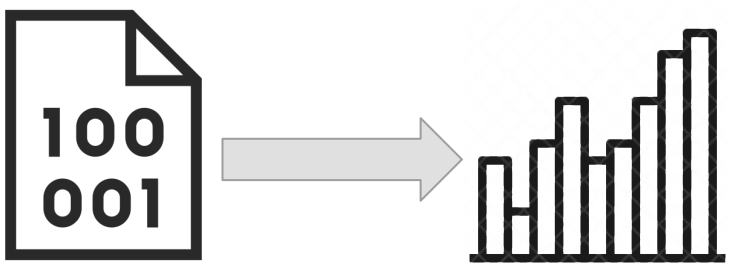
Dataset

- 200 millions encrypted and compressed fragments:
  - ➢ AES encrypted data (enc).
  - ➢ DEFLATE- and rar-compressed data (zip/gzip and rar).
  - ➢ png and jpeg images.
  - ➢ mp3 audio files.
  - ➢ pdf documents.
- Split each file into fragments of 512B, 1KB, 2KB, 4KB, and 8KB.
- Selected 5 millions fragments for each fragment size/data type.

How it works?

Data fragment

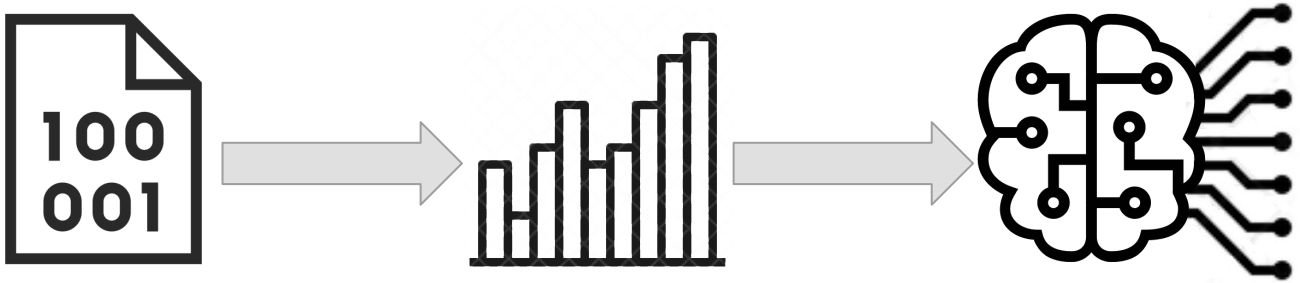
How it works?



Data fragment

Histogram

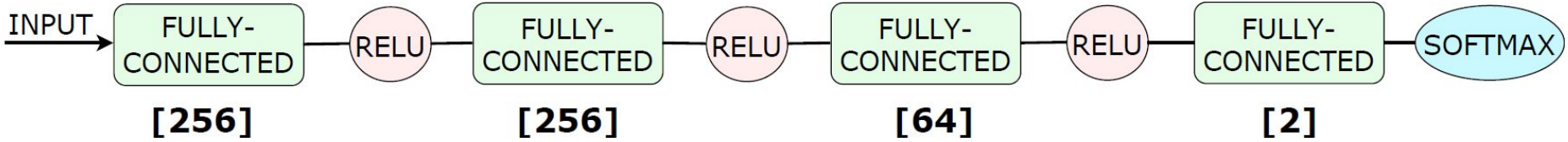
How it works?



Data fragment Histogram Classification

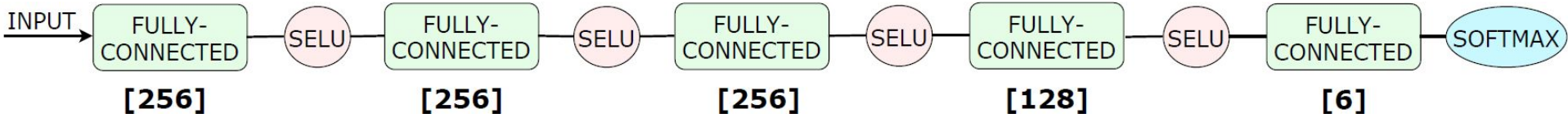
Binary Classifier

- 3M vectors from the encrypted class.
- 3M vectors from the data type we aim to distinguish.
  - ➢ 85% training,
  - ➢ 5% development,
  - ➢ 10% test.



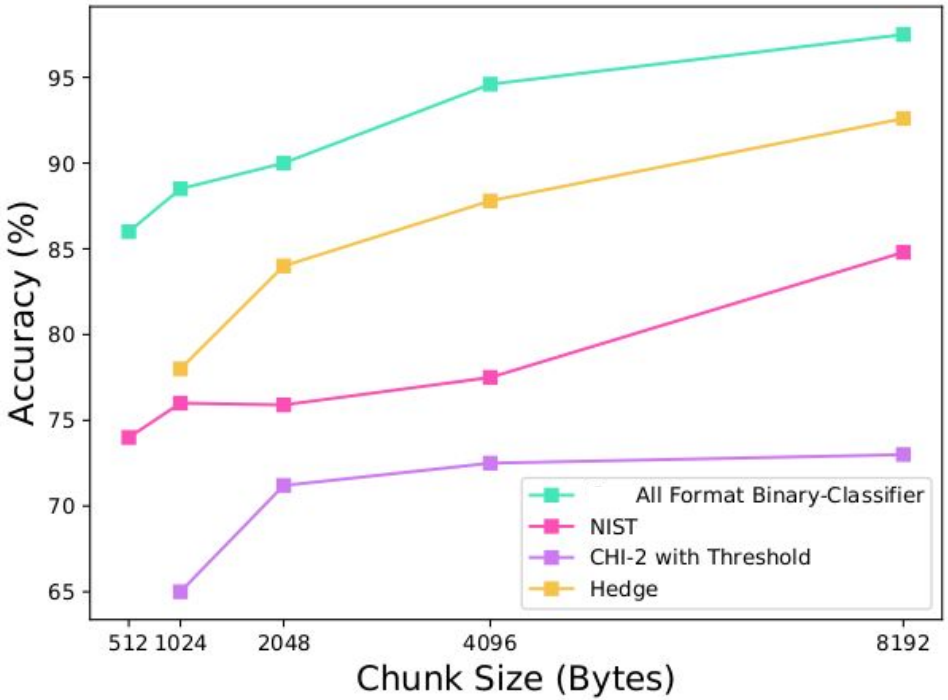
Content-Type Detectors

- 6M vectors from a mix of the file types considered.
  - ➢ 85% training,
  - ➢ 5% development,
  - ➢ 10% test.



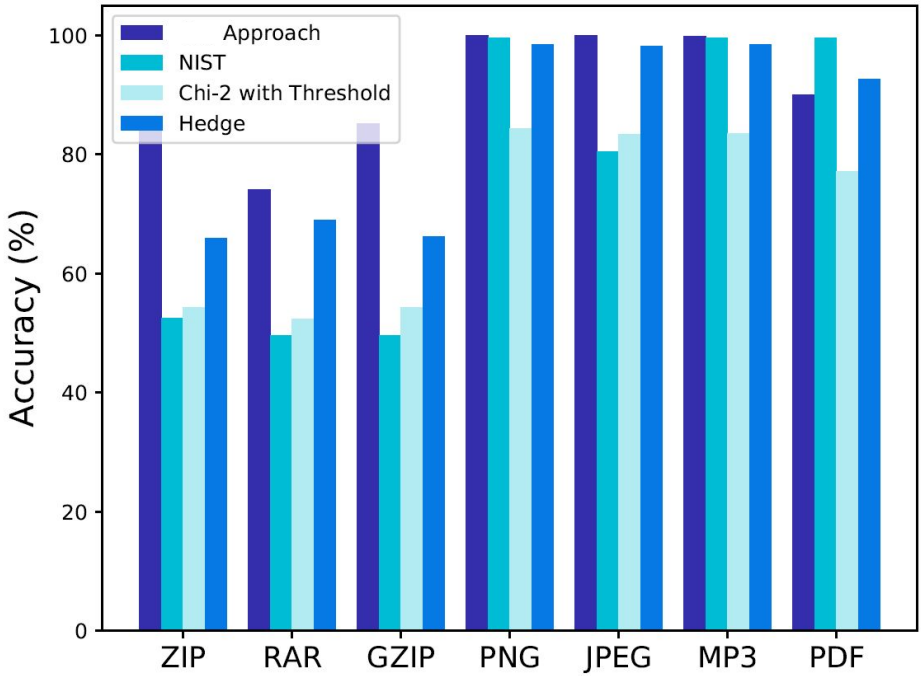
Evaluation

Q: *Is a given data fragment compressed or encrypted?*



Evaluation

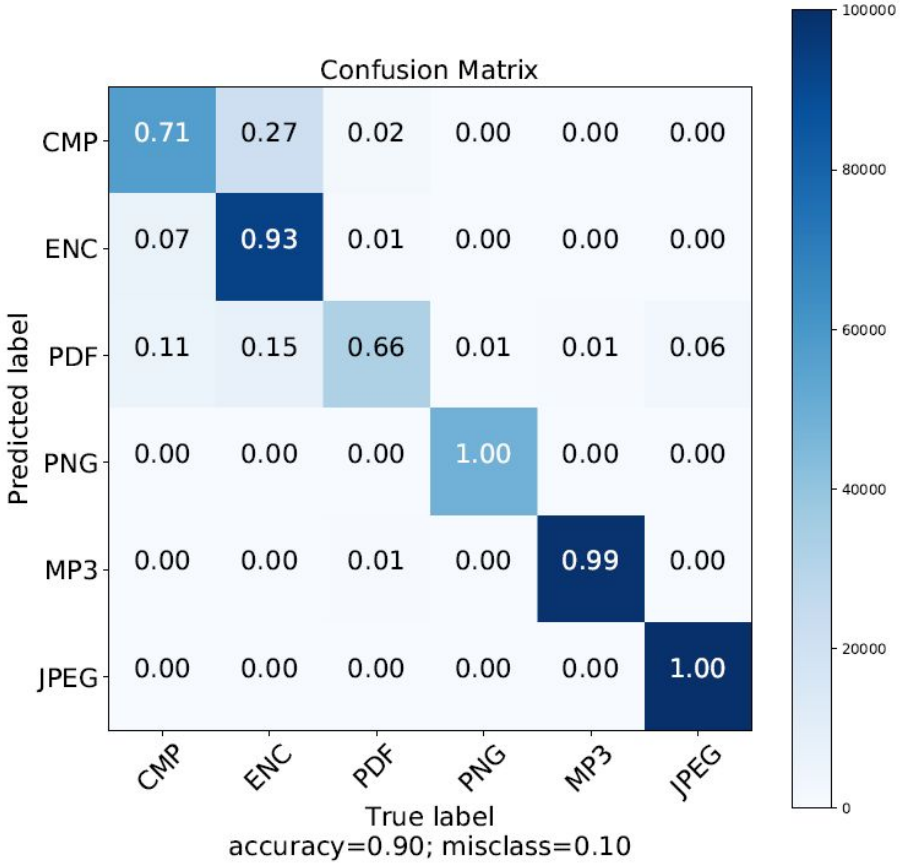
Q: *Are some compressed formats harder than others to distinguish from encrypted content?*



Evaluation

The content-type detector has the ability to:

- distinguish encrypted and compressed data.
- pinpoint the specific format compressed data belong to.



Overhead comparison

- 1000 randomly-selected compressed or encrypted samples.
- Run all three approaches on each sample, taking individual runtime and repeating the experiment 1000 times (results in seconds).

Approach	Mean	Median	Std.dev
NIST	0.1	0.1	0.004
HEDGE	0.44	0.43	0.008
Binary Classifier	0.00046	0.00044	0.00012

Findings

- Statistical tests are better than entropy measurement but they have limitations.
- Accuracy highly depends on fragment size.
- A learning based approach can overcome statistical tests limitations.

Reading Material

1. Digital Forensics (the Encryption/Compression detection use case) Link-1, Link-2, Link-3