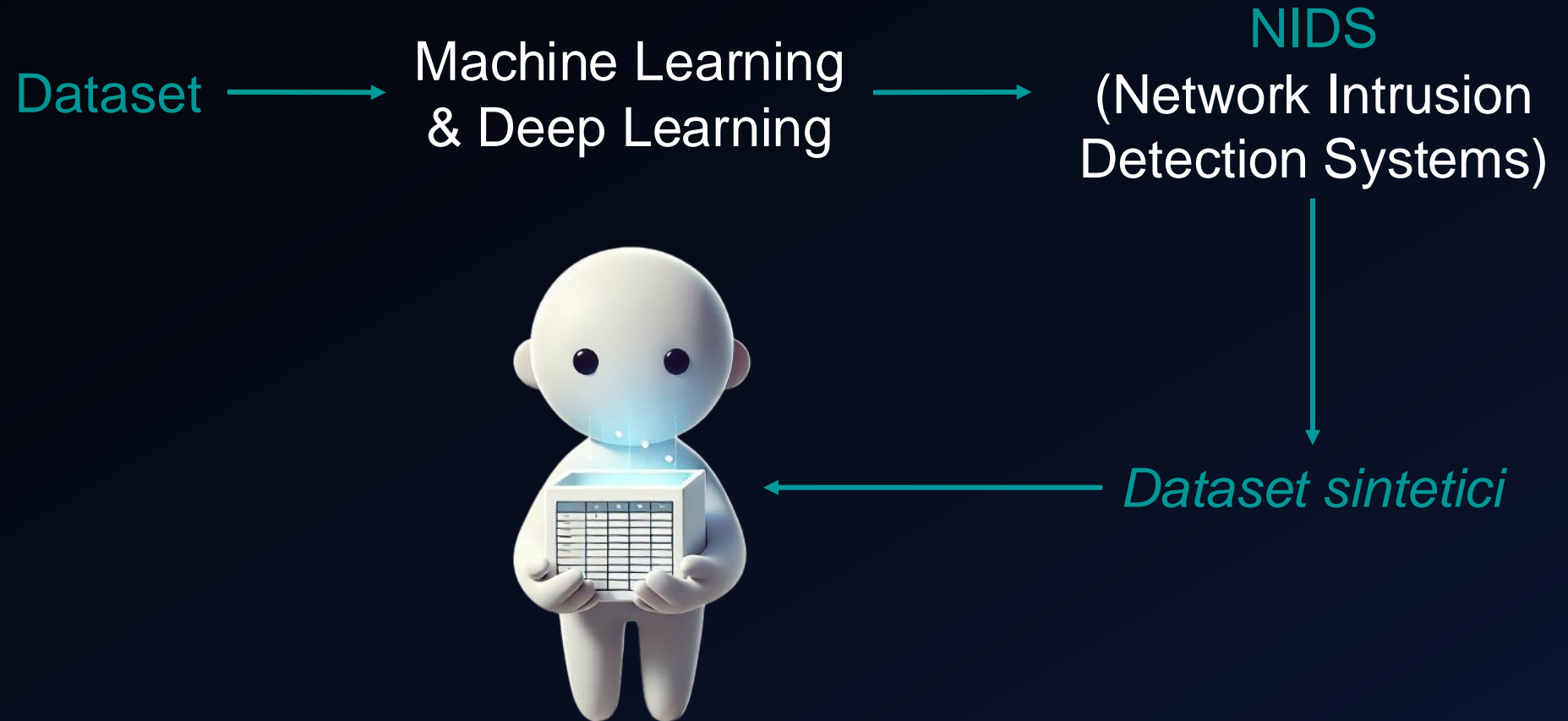


The Best Network Traffic Dataset

A comparative analysis

*Luca Corsetti
Federico Mancini
Samuele Mazziotti*

Panoramica sui Dataset di riferimento




DARPA 1998 *(Defense Advanced Research Projects Agency)*



- Novità: addestramento e test dei primi sistemi di rilevamento delle intrusioni
- Criticità: scarsa realistica del traffico benigno e il modo in cui venivano simulati gli attacchi

KDD99

- Sviluppato dal MIT Lincoln Laboratory nel 1999
- Novità: ha migliorato la struttura e la rappresentazione dei flussi di rete
- Criticità: ridondanza dei dati e scarsa varietà degli attacchi



***L'evoluzione dei dataset
ha realmente migliorato
l'affidabilità dei NIDS?***

7 Dataset e 6 "*bad design smell*"



- ISCX 2012
- CTU-13
- UNSW-NB15
- ***CIC 17 & CIC 18***
- TON IoT
- Bot-IoT

ISCX 2012 (*International Security Cybersecurity eXperiment 2012*)

sviluppato dall'Università del New Brunswick - 2012

Novità

Attacchi in 4 scenari
multistadio sovrapposti

Criticità

Mancanza di una chiara
etichettatura e limitata variabilità
del traffico generato

- Include 5 gg di traffico di rete
- Si compone di 25 host

CTU-13

sviluppato dall'Università Tecnica della Repubblica Ceca - 2014

Novità

13 scenari distinti, ognuno caratterizzato dalla registrazione di attività reali di botnet

Criticità

Presenza di traffico misto che rendere difficile l'addestramento dei modelli

- Alta granularità delle etichette
- 20 mln di connessioni e numerosi host

UNSW-NB15

sviluppato dal Cyber Range and Security Research Group dell'Università del New South Wales (UNSW) + Australian Centre for Cyber Security (ACCS) - 2015

Novità

10 categorie di attacco & set di feature:

- _ caratteristiche *di base*
- _ caratteristiche *di contenuto*
- _ caratteristiche *temporali*
- _ *flag di protocollo*

Criticità

Feature altamente dipendenti tra loro

- Il dataset include 2,5 milioni di flow records
- Comprende 45 host

CIC-IDS-2017 & CSE-CIC-IDS-2018

sviluppati dal Canadian Institute for Cybersecurity (CIC) nel 2017 e 2018

CIC 17

- Novità: 14 classi di attacco e uso di modelli comportamentali per generare traffico benigno realistico
- Criticità: problemi di etichettatura e limitata diversità del traffico
- Dimensione: 80 GB

CIC 18

- Novità: include 500 host e un dataset più esteso
- Criticità: problemi nella separazione tra alcune classi di attacco
- Dimensione: 16 milioni di flow records

TON IoT *(Telemetry Operational and Network data for the Internet of Things)*

sviluppato dal Cybersecurity Research Group - 2019

Novità

12 dispositivi e analizza
gli attacchi simulati

Criticità

Limitata eterogeneità delle
feature e l'uso predominante di
attacchi volumetrici

- Include oltre 22 milioni di record tra i vari tipi di dati

Bot-IoT

sviluppato dal Canadian Institute for Cybersecurity - 2021

Novità

10 dispositivi e affronta
le minacce specifiche
delle botnet nelle reti

Criticità

Limitata eterogeneità delle
feature e l'uso predominante di
attacchi volumetrici

- Come TON IoT, include milioni di record



Analisi dei dataset NIDS

Data Design Smells

Le 6 principali problematiche nei dataset NIDS:

- **Wrong labels** → errori di etichettatura dei dati
- **Unclear ground truth** → discrepanze nell'aggiornamento dei dataset
- **Highly dependent features** → caratteristiche troppo influenti nella classificazione dei dati
- **Poor data diversity** → bassa diversità nel tipo di dati
- **Artificial diversity** → presenza di dati sintetici
- **Traffic collapse** → generazione di traffico ad informazione limitata

Effetti collaterali nel mondo accademico

1.

Assenza di controllo sulla qualità dei dataset

2.

Dati distorti che compromettono l'accuratezza dei risultati

3.

Assenza di trasferibilità dei modelli

4.

Documentazione scarsa o poco chiara

Valutazione di CIC 17 e CIC 18

1. Attacchi mancanti

Attack:

categoria / classe del flusso

Count:

totale dichiarato & flussi
aggiunti

% Gain:

percentuale di flussi aggiunti
rispetto al totale dichiarato

ATTACK LABELS MISSED IN PUBLISHED VERSION

Attack	2017		2018	
	Count*	%Gain*	Count*	%Gain*
DDoS LOIC	[128,025] +31,339	24.48%	-	-
DDoS HOIC	-	-	[1,246,034] +918,543	73.72%
DoS GoldenEye	[10,293] +203	1.97%	-	-
DoS Hulk	[230,124] +3,680	1.60%	[923,824] +884,408	95.73%
DoS Slow HTTP	[5,499] +163	2.96%	-	-
DoS Slowloris	[5796] +10	0.17%	-	-
FTP Patator	[7,935] +19	0.24%	-	-
Heartbleed	[11] +1	9.09%	-	-
Infiltration	[36] +4	11.11%	[160,639] +63,381	39.46%
Port Scan	[158,804] +61,003	38.41%	-	-
SSH Patator	[5,897] +1	0.02%	[187,589] +389	0.21%
Web Brute Force	[1,507] +1	0.07%	-	-
Web SQL Injection	[21] +3	14.29%	[87] +1	1.15%
Web XSS	[652] +2	0.31%	[230] +2	0.87%

* Values with reference to label counts in the published dataset.

Valutazione di CIC 17 e CIC 18

2. Errori d'etichettatura

Principali cause:

- Payload etichettati erroneamente come maligni
- Artefatti di orchestrazione dei flussi
- Flussi di traffico interrotti

LABEL REASSIGNMENTS: CORRUPTIONS > 5%

Year	Published Label	Revised Label	% Corruption*	Remarks*
2017	Bot	Botnet - Attempted	25.05%	Port/System Closed. Continued C&C connection attempts by victim after attack terminated at the published time, and C&C is no longer reachable (presumably shut down).
	DoS GoldenEye	DoS GoldenEye - Attempted	25.06%	Empty Payload
	DoS Hulk	DoS Hulk - Attempted	32.70%	No malicious payload (4.47%) Empty Payload (28.23%) Attack artifacts (0.001%)
	DoS Slow HTTP	DoS Slow HTTP - Attempted	56.34%	Empty Payload (5.13%) Target Unresponsive (51.19%) Attack startup/ tear down artifact (0.02%)
	DoS Slowloris	DoS Slowloris - Attempted	29.65%	Empty Payload (29.55%) Attack startup/ tear down artifact (0.10%)
	FTP-Patator	FTP-Patator - Attempted	49.93%	No malicious payload (49.69%) Empty Payload (0.20%) Attack startup/ tear down artifact (0.04%)
	SSH-Patator	SSH-Patator - Attempted	49.80%	No malicious payload
	Web - Brute Force	Web - Brute Force - Attempted	95.16%	Empty Payload (89.85%) Attack startup/ teardown artifact (0.60%) Attack artifact (4.71%)
	Web Attack - XSS	Web Attack - XSS - Attempted	94.48%	Attack startup/tear down artifact (0.61%) Empty Payload (93.87%)

Valutazione di CIC 17 e CIC 18

3. Sovrapposizione delle classi

Attack:

categoria / classe del flusso

Overlap Class:

classe diversa da quella dell'attacco, alla quale appartengono flussi con le stesse caratteristiche

Overlap:

numero di flussi condivisi tra le due classi

AMBIGUOUS LABEL COUNTS

Year	Attack	Overlap Class	# Overlap
2017	DDoS	Benign	6
	DoS Hulk	Benign	4,872
	DoS Slowloris	Benign	1
	Portscan	Benign	1,053
2018	Brute Force - Web	Benign	1
	Web - Brute Force	Infiltration	7
	Web Attack - XSS	Benign	1
	Web Attack - XSS	SQL Injection	1
	DoS Slow HTTP	FTP-Patator	100,760
	DoS Slow HTTP	SSH-Patator	89,438
	FTP-Patator	SSH-Patator	169,745
	Infiltration	Benign	36,889



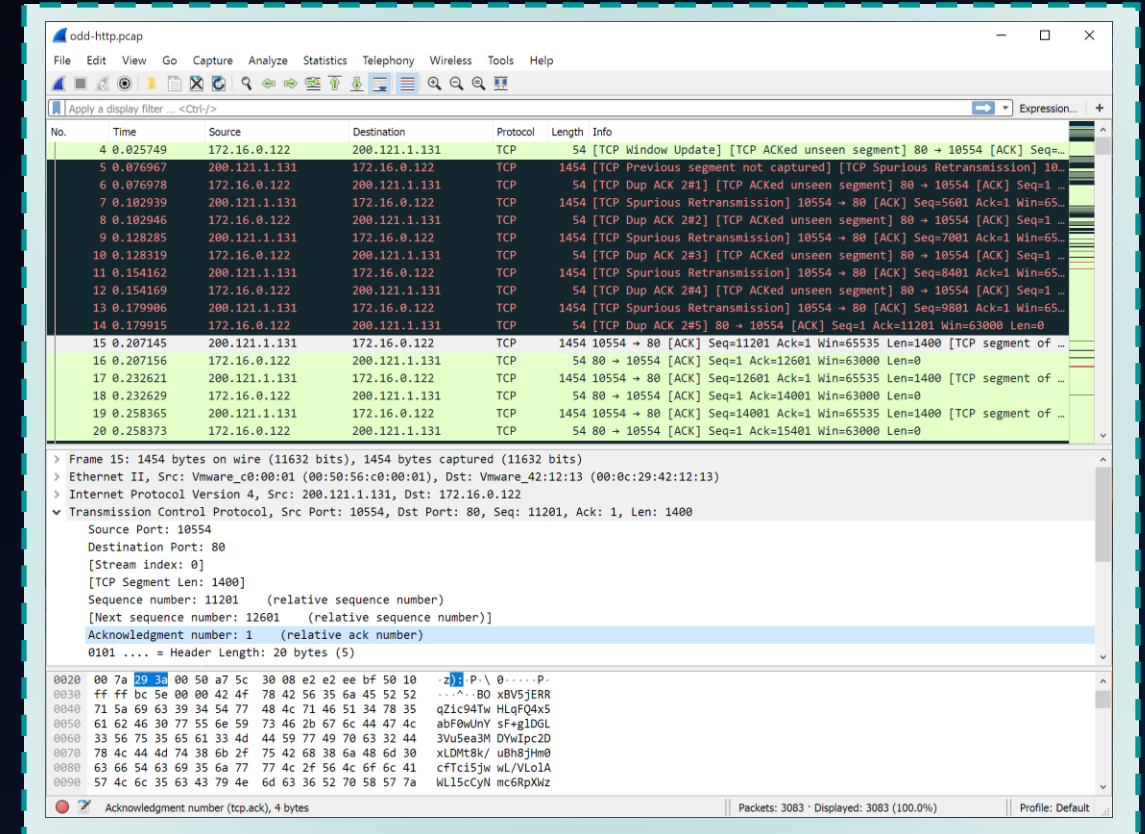
Come si può valutare un dataset?

Metodo Manuale

1. Analisi dei PCAP originali
2. Raggruppamenti dei flussi in clusters
3. Dai cluster si identificano i flow significativi
4. Si ricercano i Data Design Smells



Time Consuming





Soluzione?

Metodo Automatico



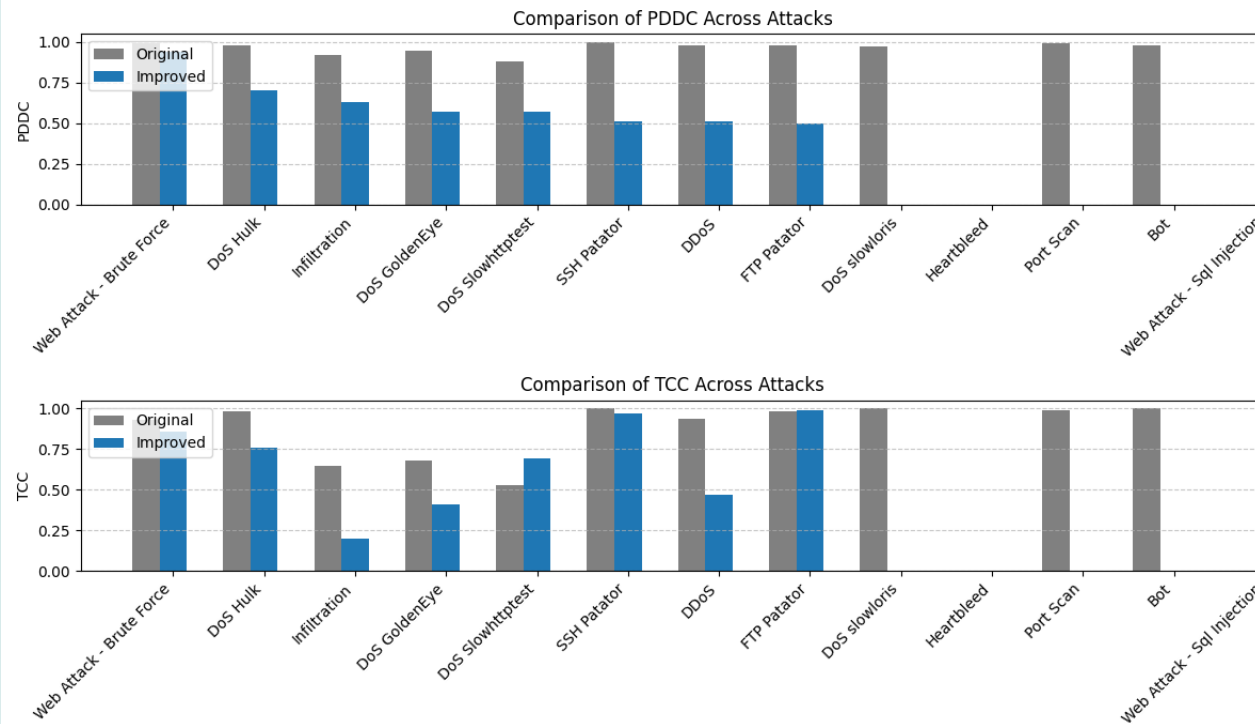
Euristiche

Le 6 principali problematiche nei dataset NIDS:

- **Wrong labels:** $WL_c = \frac{|\hat{C}|}{|C|}$ where $x \in \hat{C}$ if $ENN(x) \neq C$, $WL_c \in [0, 1]$
- **Unclear ground truth:** $UGT_c = \frac{F_{Dst\ Port \in BG\ Ports}}{|C|}$ $UGT_c \in [0, 1]$
- **Highly dependent features:** $HDF_c = \max(F1(F_i))_i$ $HDF_c \in [0, 1]$
- **Poor data diversity:** $PDD_c = \sum_{i < N} \sum_{j < M} \frac{|C_i|}{|C|} \frac{CS_{C_i}}{MN}$, $PDD_c \in [0, 1]$
- **Traffic collapse:** $TC_c = \max_i \left(\sum_{j < M} \frac{[CS_{C_i} > 0.95]}{M} \right)$ $TC_c \in [0, 1]$

Esperimenti

Valutare le euristiche sul dataset migliorato **CIC 17**



PDDC
migliorato

TCC parzialmente
migliorato

Conclusioni

I dataset non sempre garantiscono dati affidabili e rappresentativi

Non esiste un dataset perfetto

*Cosa fare per
migliorare?*

*Approccio più
rigoroso nella
selezione e
validazione dei
dataset*

Grazie per l'attenzione