ScuDo
Scuola di Dottorato – Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXVII Cycle

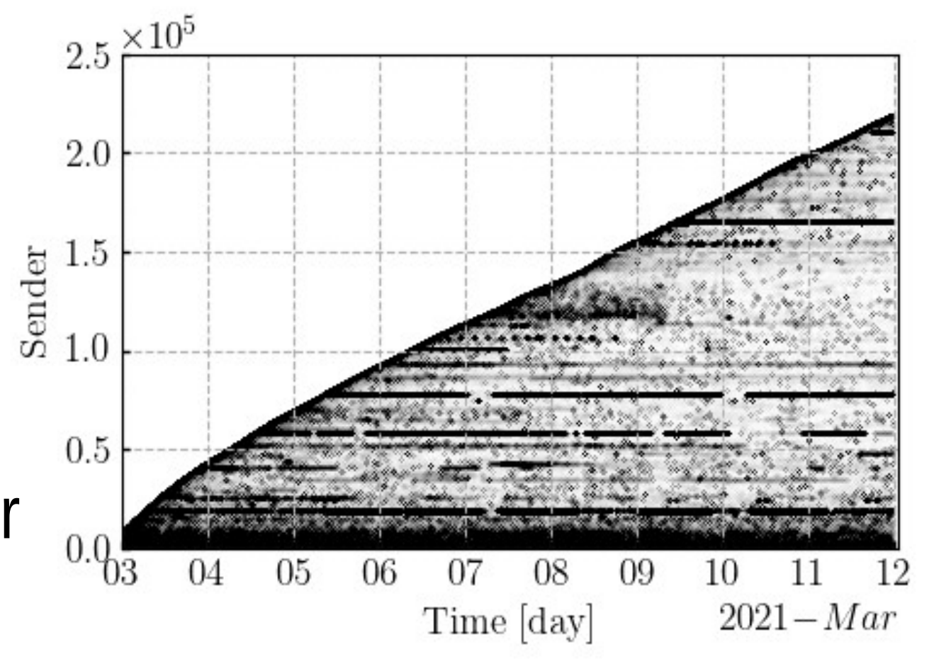
DarkVec: Automatic Analysis of Darknet Traffic with Word Embeddings

Luca Gioacchini

Supervisor: Prof. Marco Mellia

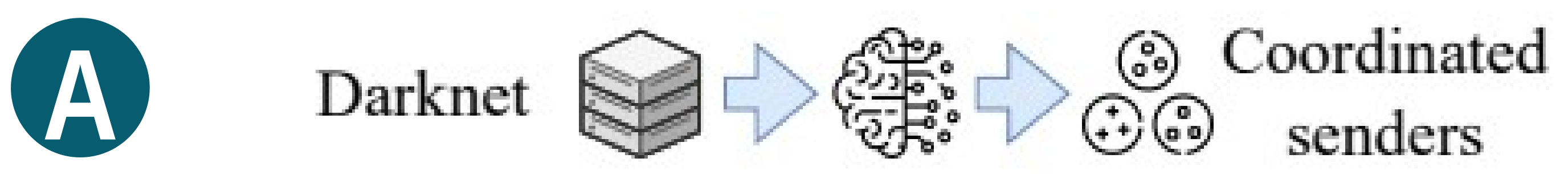
Research context and motivation

- Darknets are sets of **passive** IP addresses not hosting any service and receiving only unsolicited traffic.
- **Coordinated senders** (source IP addresses) targeting darknets may be a threat (e.g., botnets running distributed attacks).
- Need to **automatically** detect senders engaged in similar activities (coordinated).

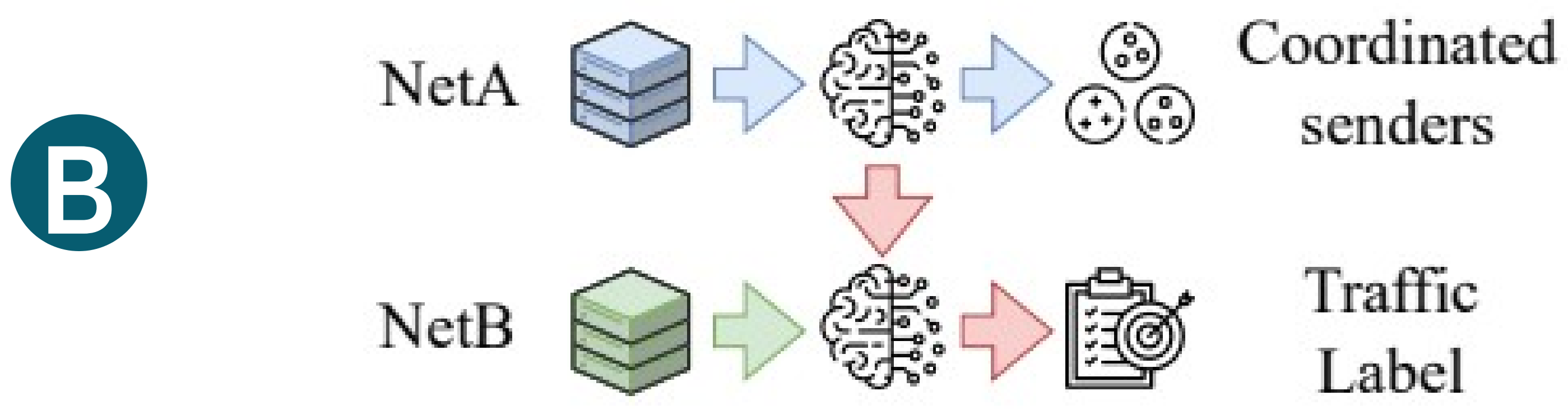


Addressed research questions/problems

- How can we **represent** similar behaviors among senders? *(Representation learning)*
- How can we **evaluate** the representations? *(Semi-supervised classification)*
- Without any prior knowledge, can we **group** senders engaged in similar activities? *(Unsupervised clustering)*

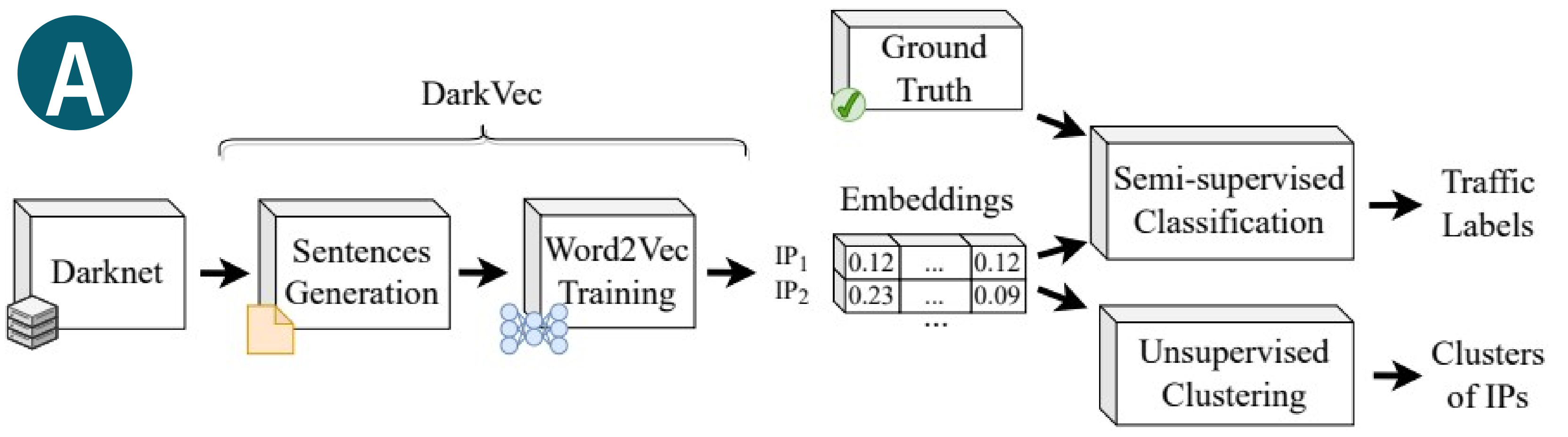


- Can we **transfer** the gained knowledge from one darknet to another? *(Transfer learning)*
- How can we **adapt** the representations between different darknets? *(Domain adaptation)*

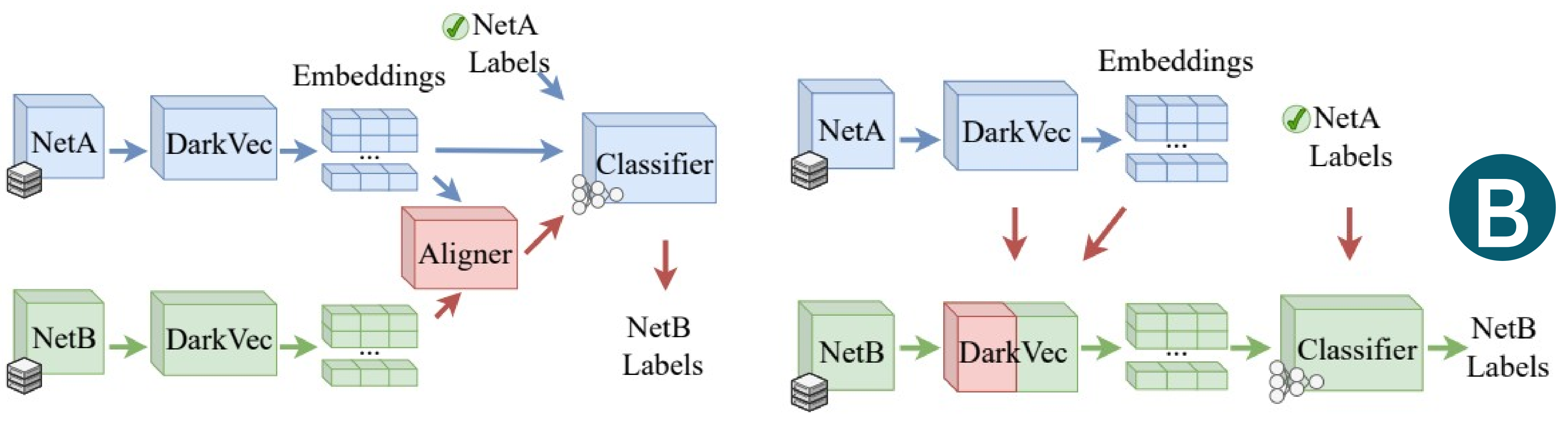


Novel contributions

- **DarkVec** – Methodology to represent senders engaged in similar activities on darknets.
- It relies on word embeddings (numeric representation of senders).



- Proposed **domain adaptation solutions**:



Explicit alignment (A*lign*)

Adaptation (A*dapt*)

Submitted and published works

Gioacchini, L., Spinsante, S. et al. "Sensors Characterization for a Calibration-Free Connected Smart Insole for Healthy Ageing", published in *International Conference on IoT Technologies for HealthCare,* vol. 360, 2021, pp.35-54

Gioacchini, L., Mellia, M. et al., "DarkVec: automatic analysis of darknet traffic with word embeddings", published in *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '21),* 2021, pp. 76-89

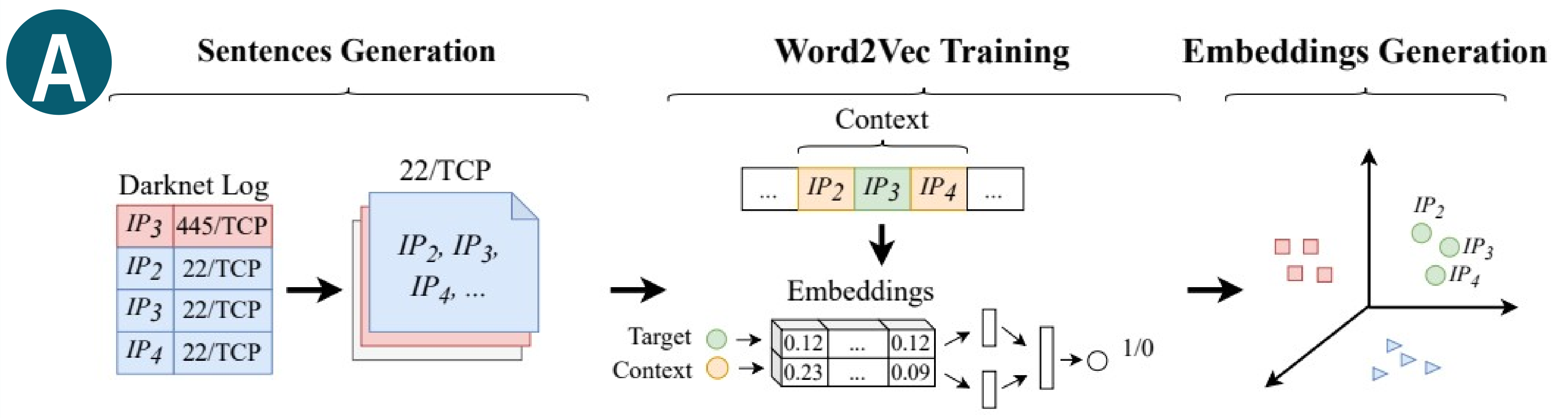
Gioacchini, L., Mellia, M. et al., "iDarkVec: incremental embeddings for darknet traffic analysis", submitted to *ACM Transactions On Internet Technologies, 2022*

Gioacchini, L., Mellia, M. et al., "Cross-network IP Embeddings Adaptation and Alignment", submitted to *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, 2023

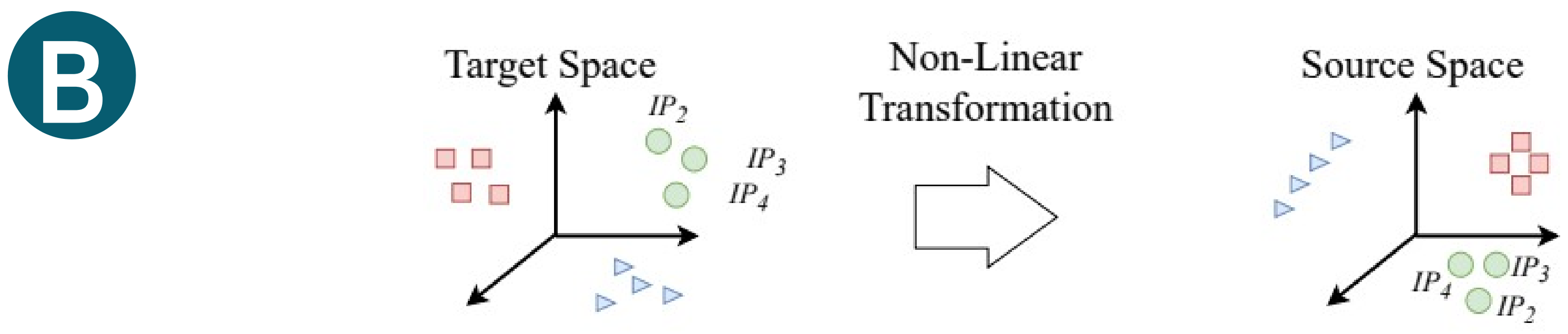
Adopted methodologies

Word2Vec – NLP technique applied to texts. It predicts the context of a word in a sentence.

- *Sentence*: **Sequence of IPs** as they reached the *top-x* darknet services.
- *Context*: **Temporal co-occurrences** of IPs targeting darknet.
- Generates embeddings such that words belonging to similar context are close in the embedding space.



Self-supervised domain aligner – Non-linear transformation. It **projects** target space embeddings onto source space ones using **anchors,** subsets of IPs active in both darknets.



Experimental results

Semi-supervised classification task

A1

	Samples	Training time	Accuracy
DANTE†	>7B	10 days	-
IP2Vec†	38M	60 min	0.67
DarkVec	4M	**18 sec**	**0.97**

† State-of-the-Art

Labels extension via knowledge transfer

B

	NetworkA→NetworkB	
	Shared Data	F1-Score
Baseline	-	0.93
Align	146 MB	0.91
Adapt	129 MB	**0.96** (+0.03)

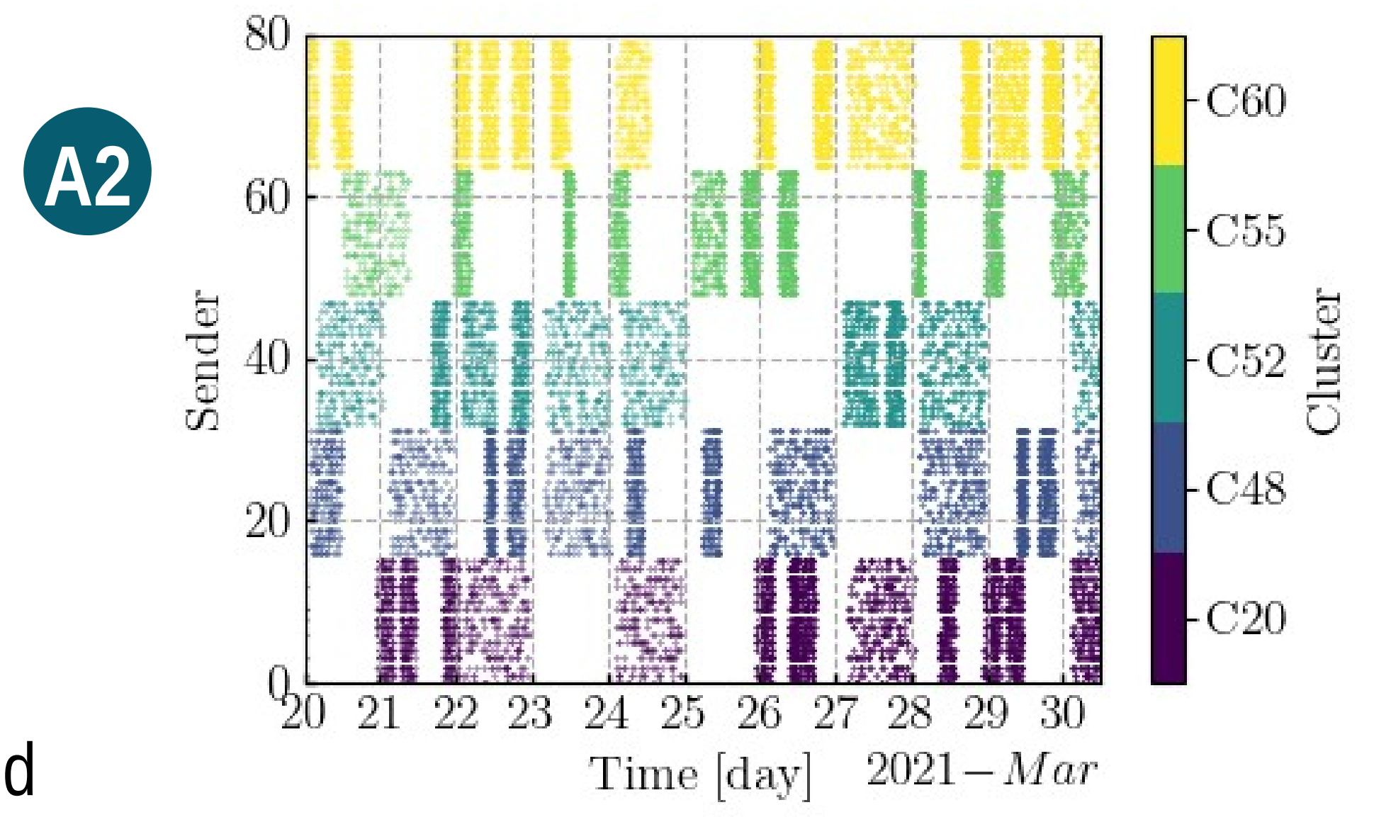
Unsupervised clustering

DarkVec embeddings allow to:

1) Detect **sub-clusters** in GT classes

2) **Extend Ground Truth** classes (334 IPs in 3 new classes)

3) Identify 13 clusters (>2k IPs) acting **suspiciously**. They were never reported in security databases



Example of sub-clusters activity patterns

Future work

- **Enriching** the embeddings through additional traffic-related information
- Study and investigate the **temporal evolution** of clusters
- Collaborative embeddings generation through **federated learning**

List of attended classes

01DNMIU – Adversarial training of neural networks (6/6/2022, 3)
01TRARV – Big data processing and programming (1/3/2022, 4)
01QTEIU – Data mining concepts and algorithms (3/2/2022, 4)
01SCSIU – Machine learning for pattern recognition (22/7/2022, 4)
01DNMIU – Optimized execution of neural networks at the edge (2/8/2022, 5)
02QUBRS – Statistical data processing (4/2/2022, 4)
02LWHRV – Communication (3/12/2021, 1)

...

Soft skills hours: 42/40
Hard skills score: 209/200

Politecnico di Torino
1859

PhD program in
Electrical, Electronics and Communications Engineering