IEEE European Symposium on Security and Privacy

Genova – June 7th, 2022

SoK: The Impact of Unlabelled Data in Cyberthreat Detection

Giovanni Apruzzese, Pavel Laskov, Aliya Tastemirova

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Once upon a time…

- At the beginning of 2021, I was having a meeting with Prof. Pavel Laskov, brainstorming about new research directions on Machine Learning (ML)

- Pavel: “We should look at Semisupervised Learning, it’s very trendy now!”

UNIVERSITÄT
LIECHTENSTEIN

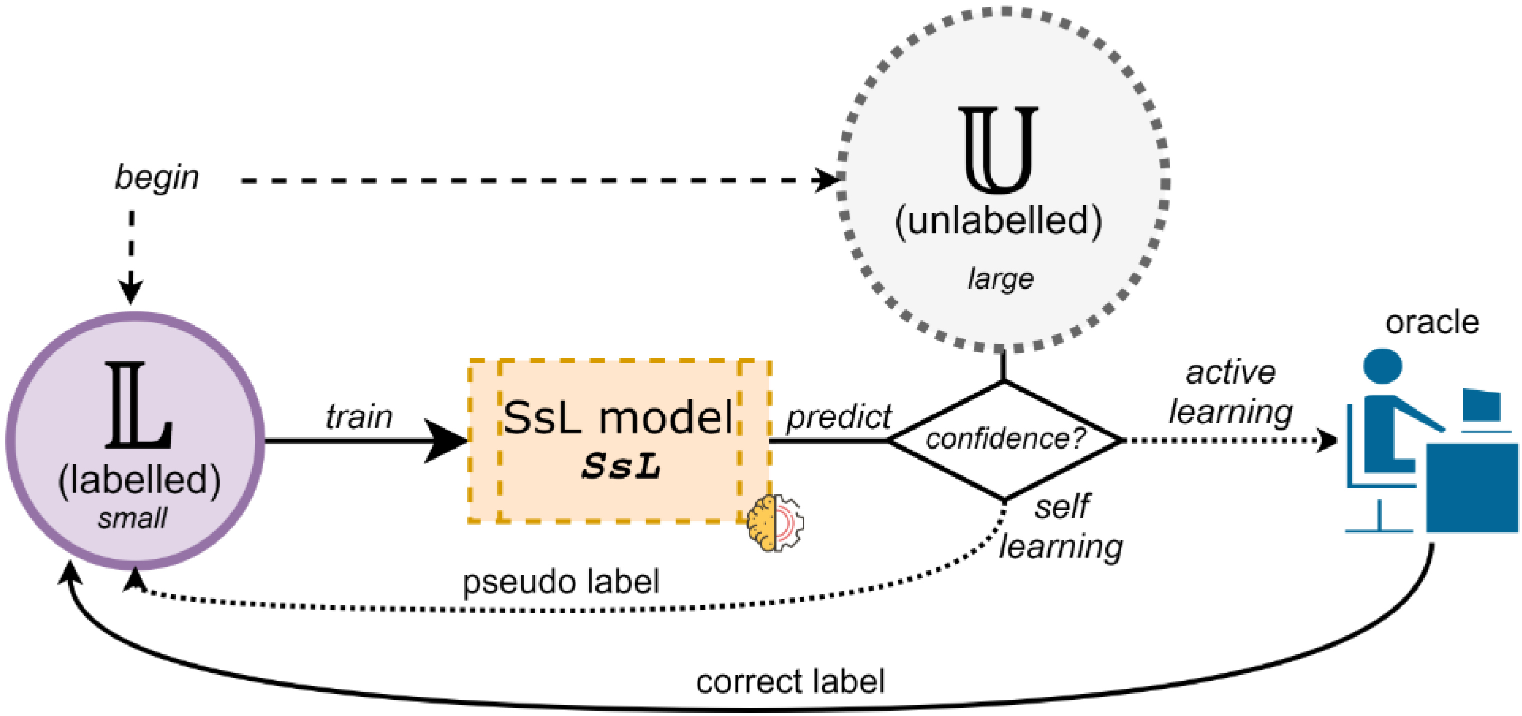
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Semisupervised Learning

- Labelled data is expensive, but *unlabelled* data is cheap(er)

→ Why not using unlabelled data to improve the proficiency of ML models?

Mixing *labelled* with *unlabelled* data is a ML approach denoted as **"Semisupervised Learning" (SsL)**



The assumptions of SsL appears to be enticing for Cyberthreat Detection (CTD)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Once upon a time… (cont'd)

- At the beginning of 2021, I was having a meeting with Prof. Laskov, brainstorming about new research directions on Machine Learning (ML)
- Pavel: "We should look at Semisupervised Learning, it's very trendy now!"
- It was the first time I directly tackled SsL, so I did what most researchers do when they start focusing on a new topic:
  - I looked into **existing literature** on SsL applications for CTD…
  - …and started to **replicate (basic) SsL methods** on public CTD datasets

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

All that glitters is not gold…

- My initial results portrayed SsL to be **bad**.
  - Like, really bad ☺

- As a sanity check, I asked a colleague of mine (Aliya Tastemirova) to:
  - **independently** replicate the SsL methods I developed
  - and evaluate their performance on **different CTD datasets**

- Her results confirmed my initial findings.

- We (Pavel, Aliya, and I) had a joint meeting, and we decided to dig deeper:
  - either all of **us were wrong**…
  - …or **something odd was going on** between the lines.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Bad performance?

- In some cases (e.g., Phishing Detection), SsL methods achieved 0.90 F1-score by using ~100 labels and thousands of unlabelled samples.

- One could claim such performance to be good…

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

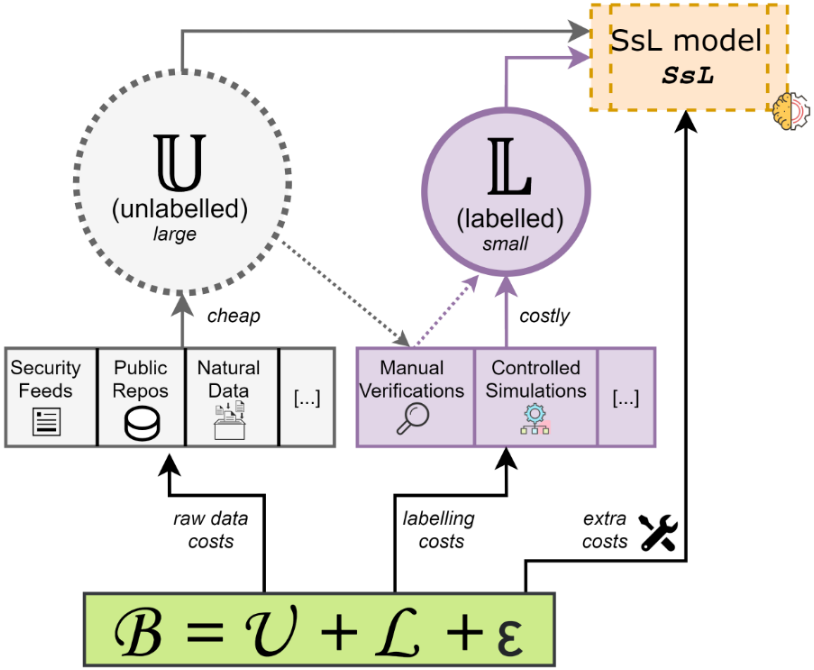
Bad performance? (cont'd)

- In some cases (e.g., Phishing Detection), SsL methods achieved 0.90 F1-score by using ~100 labels and thousands of unlabelled samples.

- One could claim such performance to be good…

- …unless a (traditional) supervised learning classifier using *only* 100 labels (without any unlabelled data) achieved an F1-score of **0.91**

- Our initial experiments showed that using unlabelled data provided "uncertain" improvement (if any).
  - In reality, unlabelled data may be cheaper to acquire than labels, but it is not **free**!



UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

If SsL is bad, then why is it so trendy in research?

- We investigated all (ttbook) existing literature on SsL for CTD, asking ourselves:

  "What are the benefits of unlabelled data in SsL?"

UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

If SsL is bad, then why is it so trendy in research?

- We investigated <u>all</u> (ttbook) existing literature on SsL for CTD, asking ourselves:

"What are the benefits of unlabelled data in SsL?"

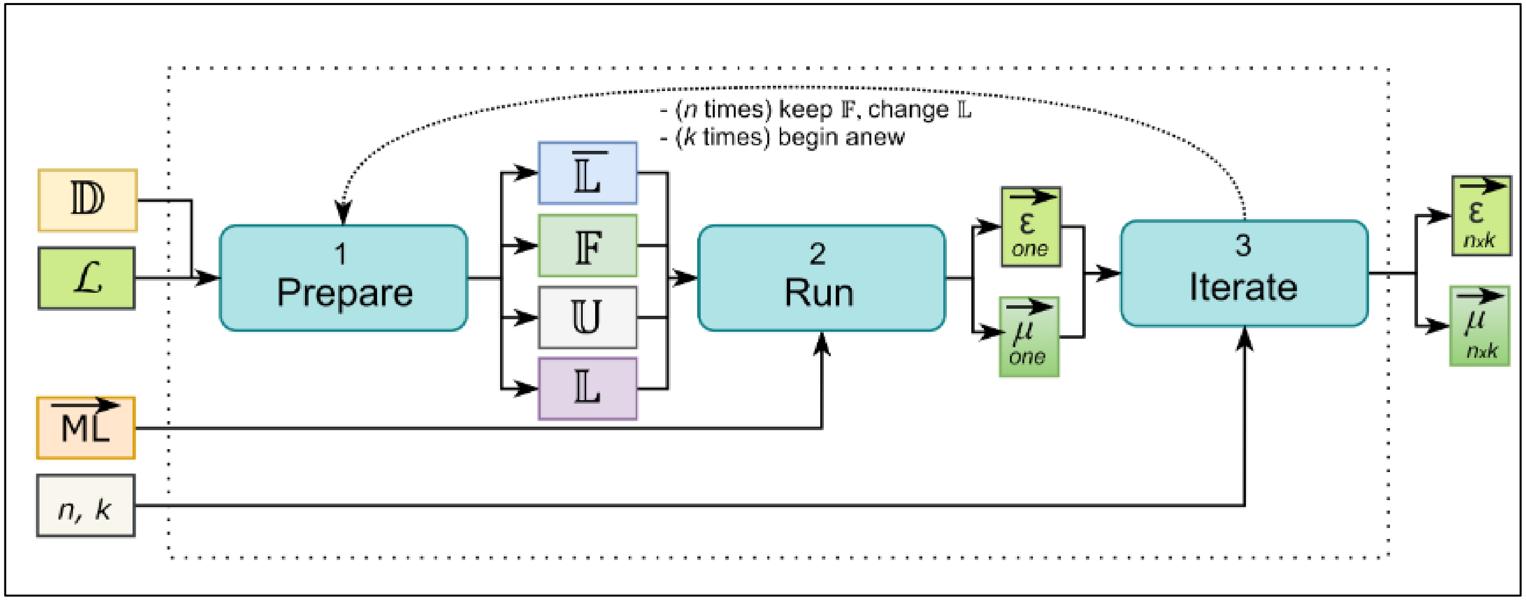
Task	Paper (1st Author)	Year	Lower Bound	Ablation Study	Upper Bound	Stat. Sign.	Transparency		Repr.	Dataset
							Labels	*Balance*		
Network Intrusion Detection	Li [93]	2007	✓	✓	✗	✗	✓	✓	◐	NSL-KDD
	Long [94]	2008	✓	✓	✗	◐	✓	✗	◐	NSL-KDD
	Görnitz [95]	2009	✓	✓	✗	◐	✓	✓	✗	Private
	Seliya [96]	2010	✓	✓	✗	✗	✓	✓	◐	NSL-KDD
	Symons [97]	2012	✗	✓	✓	◐	✓	✗	✗	Kyoto2006
	Wagh [98]	2014	✗	✗	✗	✗	✓	✓	◐	NSL-KDD
	Noorbehbahani [35]	2015	✗	◐	✓	✗	✓	✓	◐	NSL-KDD, Custom
	Ashfaq [99]	2017	✗	◐	✗	✗	✓	✗	◐	NSL-KDD
	Qiu [67]	2017	✗	◐	✓	✗	✓	✓	✗	Custom
	McElwee [100]	2017	✗	◐	✓	✗	✓	✗	◐	NSL-KDD
	Kumari [68]	2017	✓	◐	✗	✗	✓	✗	◐	NSL-KDD
	Yang [101]	2018	◐	✓	✓	✗	✓	✗	✗	NSL-KDD, AWID
	Gao [102]	2018	✓	◐	✗	✗	✓	✗	✗	NSL-KDD
	Shi [103]	2018	◐	◐	✗	✗	✓	✗	✗	NSL-KDD
	Yao [36]	2019	◐	◐	✓	✗	✓	✓	◐	NSL-KDD
	Yuan [104]	2019	✗	◐	✗	◐	✓	✓	◐	NSL-KDD
	Zhang [65]	2020	◐	✗	✓	◐	✓	✗	◐	NSL-KDD
	Hara [105]	2020	✗	◐	✓	✗	✗	✗	✗	NSL-KDD
	Ravi [106]	2020	✓	✗	✗	✗	✓	✗	✗	NSL-KDD
	Gao [107]	2020	✗	✓	✓	✓	✓	✓	✗	NSL-KDD
	Li [108]	2020	✗	◐	✓	✓	✓	✗	◐	NSL-KDD, Private
	Zhang [70]	2021	◐	◐	✗	◐	✗	✓	◐	CICIDS2017, CTU13
	Liang [109]	2021	✓	◐	✓	◐	✓	✓	◐	NSL-KDD
Phishing Detection	Gyawali [110]	2011	✗	✓	✓	✗	✓	✓	◐	Private
	Zhao [111]	2013	✓	✓	✓	✓	✗	✓	✓*	DetMalURL
	Gabriel [15]	2017	◐	◐	✗	✗	✗	✗	◐	Private
	Yang [112]	2017	✓	◐	✗	✗	✓	✓	◐	Private
	Bhattacharjee [113]	2017	✗	✓	✗	◐	✗	✗	◐	Private
	Li [55]	2017	✓	✓	✓	◐	✓	✓	✗	Custom
Malware Detection	Moskovitch [114]	2008	✗	✓	✗	◐	✓	✓	✗	Custom
	Santos [115]	2011	✗	✗	✓	✗	✓	✓	◐	Custom
	Nissim [116]	2012	✗	◐	✓	◐	✗	✗	✗	Private
	Zhao [117]	2012	✗	✗	✗	✗	✓	✓	◐	Private
	Nissim [118]	2014	✓	✓	✗	◐	✓	✓	✗	Custom
	Zhang[119]	2015	◐	◐	✗	✗	✓	✓	✗	Private
	Nissim [120]	2016	✗	✓	✓	◐	✓	✓	◐	Custom
	Ni [121]	2016	✓	✓	✗	◐	✓	✓	◐	Private
	Chen [122]	2017	✓	✓	✗	◐	✗	✗	◐	Private
	Rashidi [66]	2017	✗	✓	✓	◐	✓	✓	✗	Drebin
	Fu [123]	2019	✓	✓	✗	✗	✓	✗	◐	Private
	Irofti [124]	2019	◐	◐	✗	◐	✗	✗	✓	DREBIN, EMBER
	Pendlebury [86]	2019	✗	✗	✓	◐	✓	✓	✓	AndroZoo
	Sharmeen [125]	2020	✓	◐	✗	◐	✓	✓	◐	Drebin, AndroZoo
	Chen [126]	2020	◐	◐	✓	✗	✓	✓	◐	MCC
	Koza [11]	2020	✓	◐	✓	◐	✓	✗	✓	Private
	Noorbehbahani [13]	2020	✓	✗	✗	◐	✓	✓	✗	AndMal17
	Li [127]	2021	✗	◐	✗	◐	✓	✗	◐	FalDroid, DREBIN, Genome
	Liang [109]	2021	✓	◐	✓	◐	✓	✓	◐	Custom

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Revealing the impact of unlabelled data in CTD

The state-of-the-art does not allow to determine whether using unlabelled data is *truly* beneficial in CTD

- As a constructive step, in our paper we:
  - Provide a set of requirements to estimate the benefits (if any) of using unlabelled data in CTD
  - Propose a framework, CEF-SsL, that allows to meet all such requirements in research
  - We experimentally evaluate CEF-SsL on 9 CTD datasets by considering 9 SsL methods.



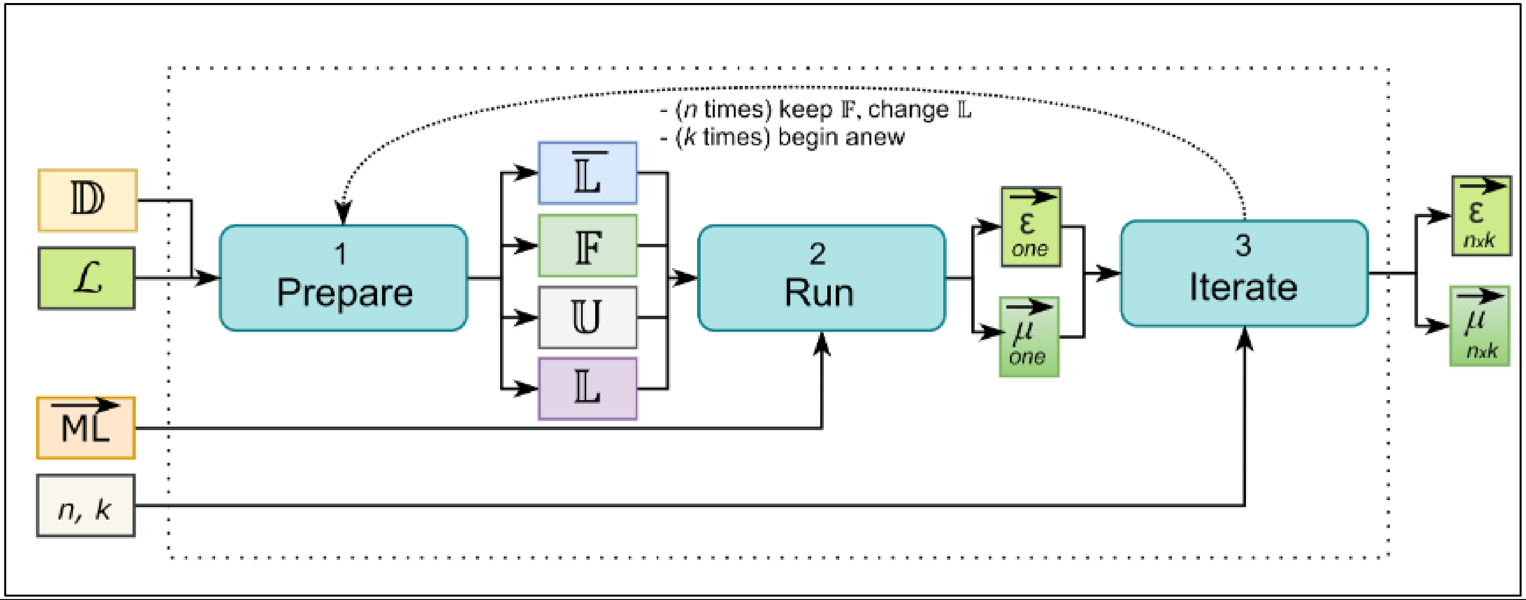
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Revealing the impact of unlabelled data in CTD

The state-of-the-art does not allow to determine whether using unlabelled data is *truly* beneficial in CTD

- As a constructive step, in our paper we:
  - Provide a set of requirements to estimate the benefits (if any) of using unlabelled data in CTD
  - Propose a framework, CEF-SsL, that allows to meet all such requirements in research
  - We experimentally evaluate CEF-SsL on 9 CTD datasets by considering 9 SsL methods.



Let me show you some hard numbers on the "troubleshooted" version of CICIDS17 [1]…

UNIVERSITÄT LIECHTENSTEIN

[1] Engelen, Gints, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." 2021 IEEE Security and Privacy Workshops.

IEEE European Symposium on Security and Privacy

Genova – June 7th, 2022

SoK: The Impact of Unlabelled Data in Cyberthreat Detection

Giovanni Apruzzese, Pavel Laskov, Aliya Tastemirova

UNIVERSITÄT LIECHTENSTEIN