Teaching Network Security with IP Darkspace Data

Tanja Zseby, Felix Iglesias

Institute of Telecommunications
Faculty of Electrical Engineering and Information Technology
TU Wien

September 9, 2019

TU Wien Network Security Classes

- Two Security Courses for Master students
  - Theory Lectures (6 x 90 min) ➔ written exam
  - Lab Exercises (6 x180 min), Teams of 2 ➔ Report
  - Lab Review (oral exam)
  - Classes offered since 2014, continuously updated
  - 2014: 35 students ➔ 2019: 88 students
- 1. Network Security
  - Lab: IP Darkspace Analysis
  - Data: CAIDA IP darkspace data
- 2. Network Security Advanced
  - Lab: Network Steganography
  - Data: Modified MAWI Dataset (WIDE)

Educational Objectives

- Research-oriented teaching concept
  - Include current research in the classroom

- Class objectives:
  - Familiarize students with network data analysis methods
  - Provide students in-depth understanding of TCP/IP flow behavior
  - Deepen students' network security knowledge
  - Enable students' general scientific work skills
  - Increase exploratory and forensics analysis skill
  - Awaken the scientist in each student

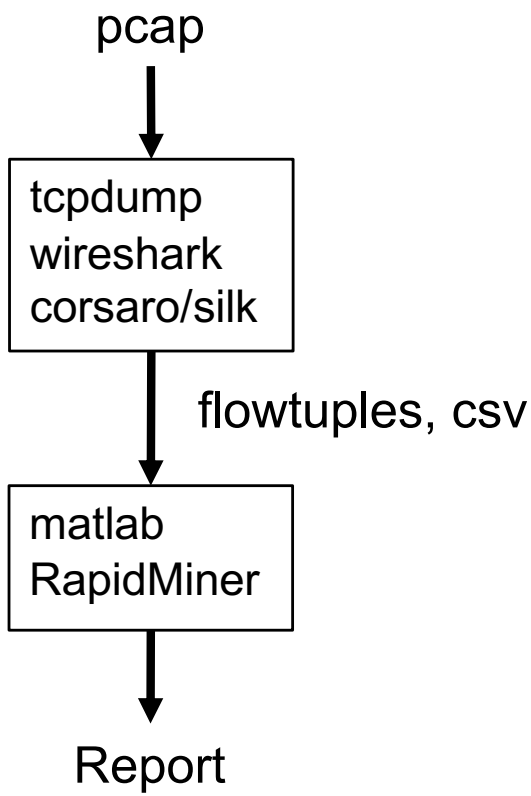
Students

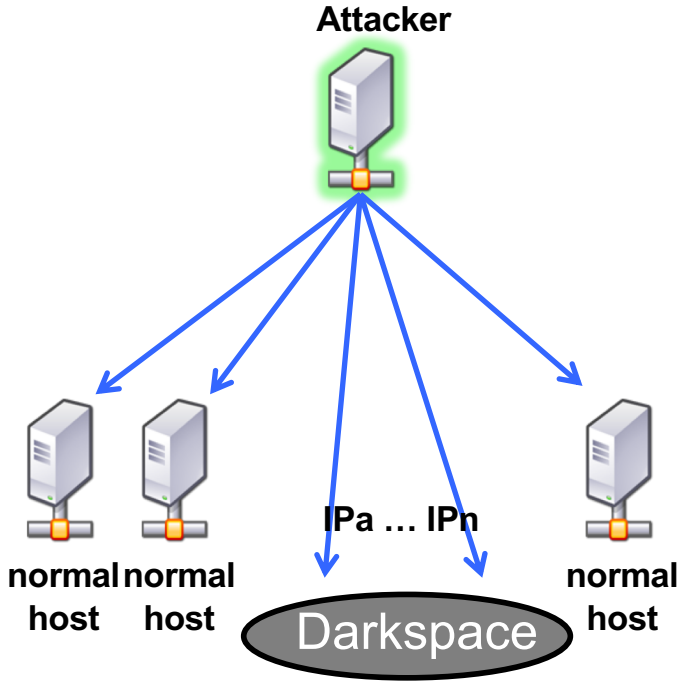
- International students
  - Different bachelor programs ➔ Different skills
- Different Masters programs
  - Electrical engineering
    - Telecommunications
    - Embedded Systems
  - Computer Science
  - Future: Data Science Master
- Ideal: if students from different programs team up
  - EE students with matlab, signals and systems experience
  - CS students with programming and Linux skills

NetSec Lab: IP Darkspace Analysis

- CAIDA IP Darkspace Data (Telescope Data)
  - Each Team gets different set of IP darkspace data
  - Students required to use recommended tools
- Exercises

1. preprocessing/ aggregation
2. Uni/bivariate analysis
3. Time Series
4. FFT





Example Exercises

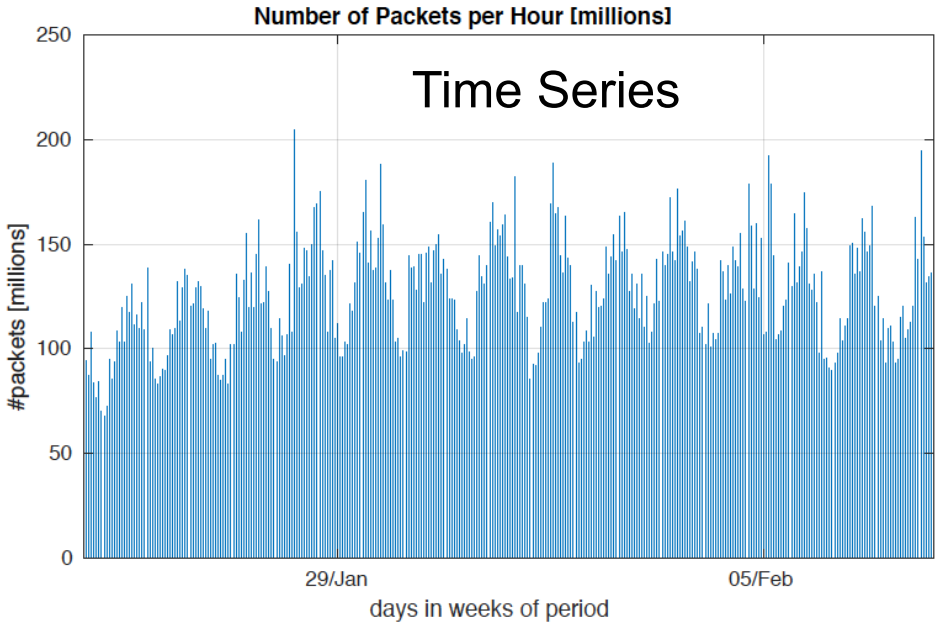
Ex. 3.3 – TCP port distribution

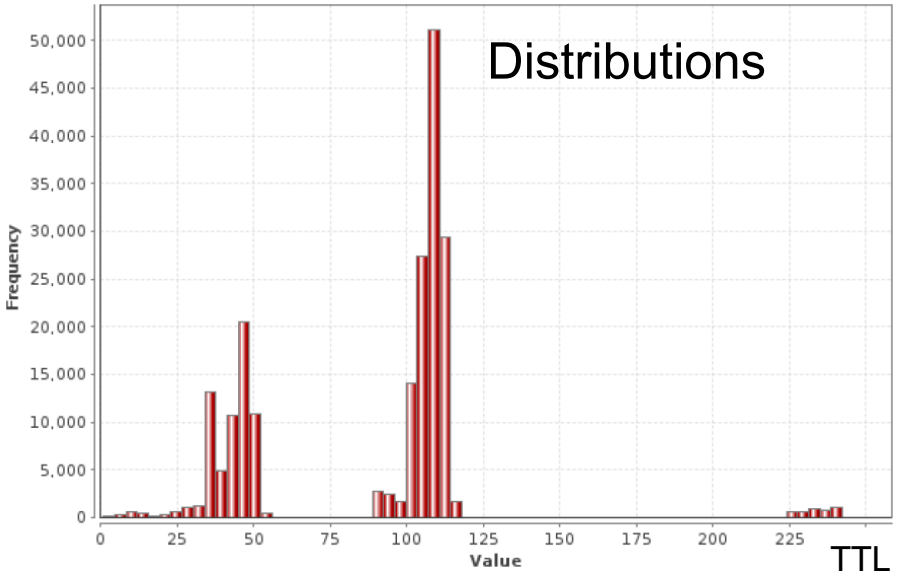
- **[rep-15:]** Write down in the report the number of TCP packets that corresponds to every one of the five most active ports for your whole time period (absolute value and percentage). MATLAB/Octave `sortrows` function could be helpful here.
- **[rep-16:]** Plot a frequency distribution showing the percentage of TCP packets (for your time period) for the different ports. It should have six bars, the first five bars correspond to the most active ports, whereas the last bar should account for the rest of the less active ports together (`bar` function).

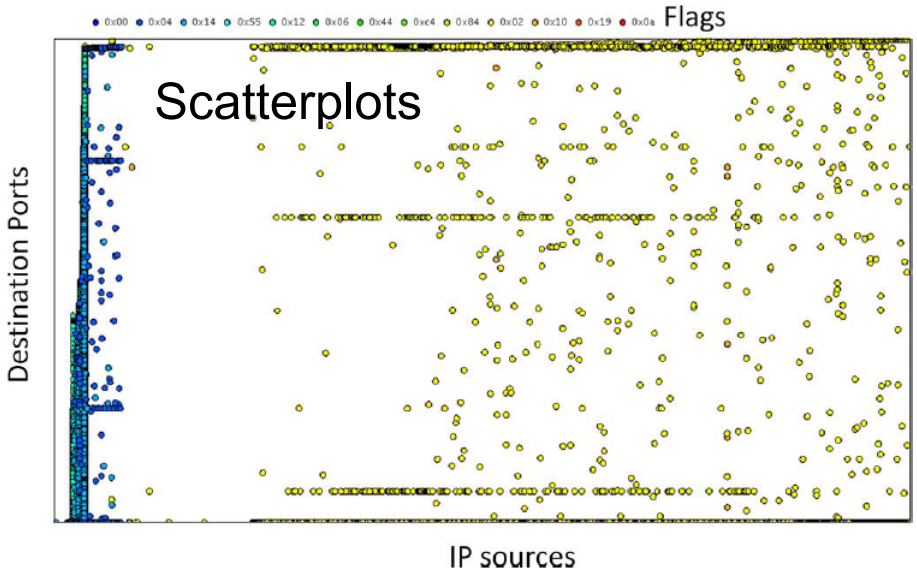
Ex. 3.4a – Temporal patterns of IP sources

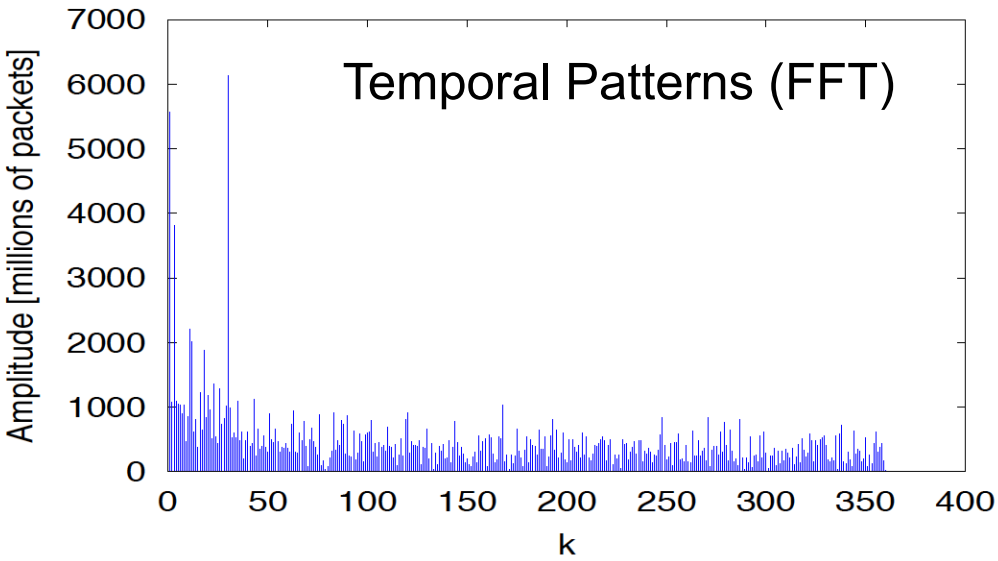
- **[rep-17:]** Perform listing 7 for IP sources and packets per hour. Add both FFT plots to your report. (remember to remove the offset in the plots).
- **[rep-18:]** Find the k values corresponding to the most noticeable peaks for every case and copy them in the report.
- Answer the following questions:
  - **[rep-19:]** To which time periods do the k values correspond to (p_k =?, consider that the unit is *hours*)?
  - **[rep-20:]** What do these peaks mean?
  - **[rep-21:]** Do you find any remarkable difference between the FFT graph of the number of packets per hour and the FFT graph of the number of unique IP sources per hour? Provide an explanation.

Data Analysis (Examples from Reports)



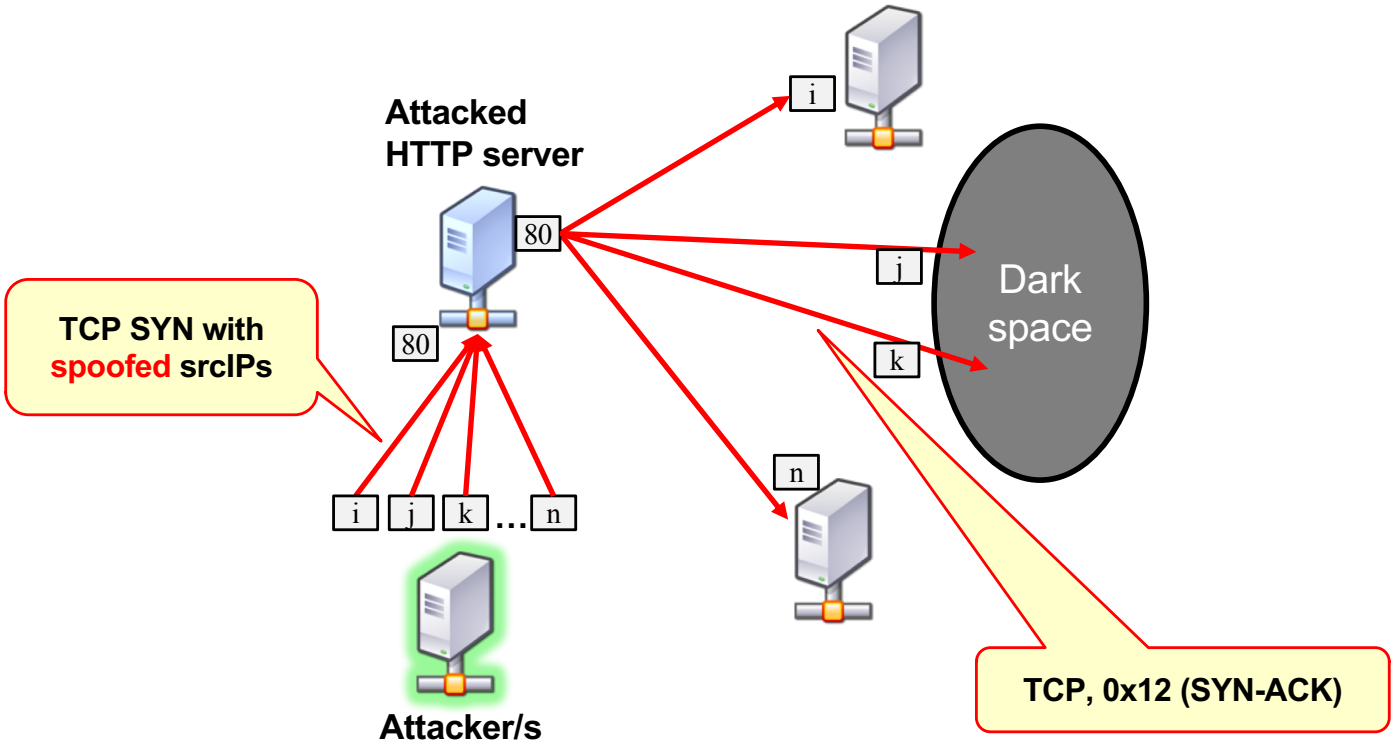






Identifying Backscatter

Name	Type	Statistics	Range
srcIP	nominal	mode = 191.191.71.228 (40052), least = 24	191.191.71.228 (40052), 0.11.59.95 (0), 0.119.105.126 (0), (
dstIP	nominal	mode = 0.196.51.3 (2), least = 0.76.243.80	0.128.198.153 (2), 0.129.230.94 (2), 0.130.14.70 (2), 0.130.1
srcPort	nominal	mode = 80 (40052), least = 3 (0)	80 (40052), 0 (0), 10023 (0), 10075 (0), 10081 (0), 10100 (0
dstPort	nominal	mode = 18285 (9), least = 0 (0)	18285 (9), 50218 (8), 17698 (7), 22878 (7), 39225 (7), 5792
Protocol	nominal	mode = 6 (40052), least = 1 (0)	1 (0), 6 (40052)
TTL	numeric	avg = 92.327 +/- 2.845	[38.000 ; 93.000]
flags	nominal	mode = 0x12 (40052), least = 0x00 (0)	0x00 (0), 0x04 (0), 0x14 (0), 0x12 (40052)
len	numeric	avg = 40 +/- 0	[40.000 ; 40.000]



Student Feedback

STUDENTS SELF-ASSESSMENT OF ACQUIRED SKILLS

1-strongly agree (positive), 5-strongly disagree (negative)

Question	max-min	mean
The course raised my interest in exploring the topic further.	1 - 2	1.29
Information was provided during the course on how I will be able to use the contents in the future.	1 - 3	1.50
The course increased my knowledge.	1 - 3	1.14
I am capable of using the knowledge I gained from the course.	1 - 3	1.21

[feedback provided by 14 students]

what did you enjoy most?

“labs were fun and engaging ”

“the moments: when you successfully finish an exercise”

What could be improved?

“tool-tutorials before the class ”

“more free exploration exercises”

“more exercises!; to be honest, I could have done another three exercises, it was fun!”

(Some) Lessons Learned

- Working with real measurement data
  - Boosts motivation, triggers research spirit
  - Encourages to check theory vs. reality
  - Teaches responsible handling of data
  - Unique data set per team ➔ cheating detection
- But: A lot of effort
  - Maintaining lab environment
  - Correcting reports
  - Unexpected effects ➔ need to check data before
- Enforce pre-requisites
- Form heterogeneous teams
- Introduce variety of tools, then allow free choice
- “Keep it Fun!” (story, easter eggs)

Benefits

- Students work with real data
  - A lot of positive responses
- Students learn about attacks
  - Scanning
  - Backscatter
  - But: only some attacks visible and mainly missed attack attempts, attack consequences not the attack itself
- Plenty of data available
  - Every team can get own data set
  - Teams may discover new things

Limitations/Challenges

- General limitations of darkspace traffic
  - No bi-directional flows, no connections
  - No Labels (not suitable for testing algorithms)
- Operational limitations
  - Huge files, huge effort for getting most recent data
  - No filter options
  - Data needs to stay in lab (students sign CAIDA agreement)
- Anonymization
  - Limits analysis options (e.g., geolocation)
- ➔ 2019 lab used only aggregated DS data
  - Now students do own attacks, preprocessing exercises with captured data

Whish List 1: Providing Data

- Offer customized data files
  - Data in different formats and sizes (file sizes, time intervals)
  - Different aggregation schemes
  - Filtered data (e.g., removing repetitive instances, 445 scans, etc.)
  - Pre-processed data (10-min captures, flows, time-series, etc.)
  - Ideal: flexible filter/aggregation options (different flow keys, time series,…)
- Provide Labels
  - Automatized analysis
  - Provide classification tools, scripts

Wish List 2: Remote Data Analysis

- Possibilities for students to work on data remotely
  - Remote work environment for multiple teams
  - Working on most recent data
- Provide standard analysis environments
  - Standard tools and programming environments
  - e.g., matlab, python, scikit-learn, Rapidminer?
  - Repeatability
- Still provide the possibility to download parts
  - User friendly query options ("time period", "signals", "sampling time", "filtering options")
- Provide info material, examples, tutorials
  - Possibility to share/discuss findings with others (CAIDA researchers, other groups?

Available Material

- IP Darkspace Data ➔ available at CAIDA
- MAWI Data: http://mawi.wide.ad.jp/mawi/

- Teaching material ➔ available to other teachers
  - Exercise Sheets
  - Solver scripts
  - Report templates
  - Evaluation and Grading Scheme

  http://www.tc.tuwien.ac.at/netsec-lab

Zseby, Iglesias, King, Claffy: "*Teaching Network Security With IP Darkspace Data*"; IEEE Transactions on Education, **59** (2015), 1; 1 - 7.

Zseby, Iglesias, Bernhardt, Frkat, Annessi: "*A Network Steganography Lab on Detecting TCP/IP Covert Channels*"; IEEE Transactions on Education, **59** (2016), 3; 224 - 232.

Thank you!

tanja.zseby@tuwien.ac.at