Network traffic classification based on periodic behavior detection

Josef Koumar, Tomáš Čejka

CTU FIT & CESNET a.l.e.
Prague, Czech republic

24. října 2022

Motivation and goal

Motivation

The possible existence of a side channel of encrypted network communication is based on the assumption that many processes communicate over the network infrastructure periodically and thereby reveal information about themselves by which they are identifiable.

Goal

1. Detection of periodic communication
   - Create time series from network traffic
   - Detection of periodic behavior on time series using mathematic model
   - Description of periodic behavior using set of features
2. Classification of applications, services and operating systems using features of periodic behavior

Create time series from network traffic

Network dependency

Network dependency is a long-term relationship between two IP addresses (devices), where one of them provides some service to the seconde one. For example a network dependency *192.168.1.1(53)–92.168.1.110* where IP *192.168.1.1* provides a DNS service to IP *192.168.1.110*.

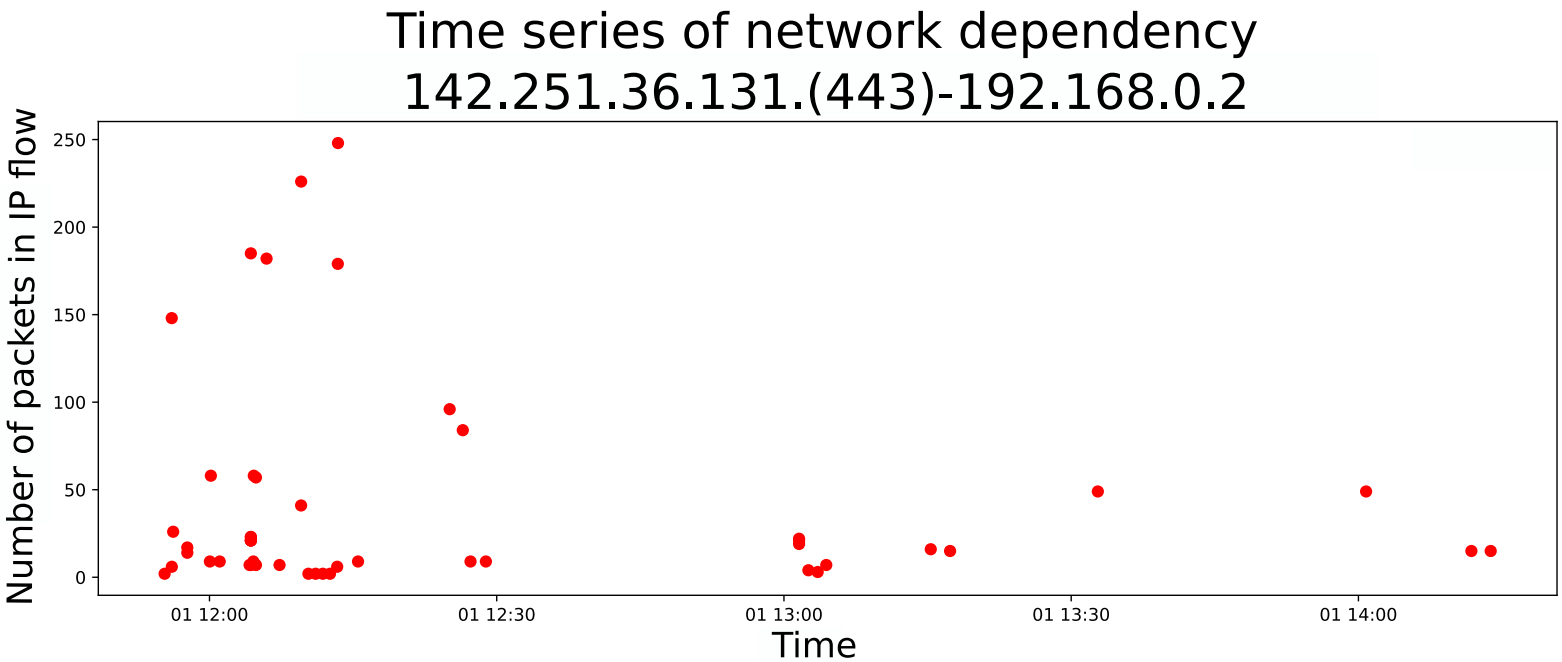
Time series from network traffic

Time series from network traffic is set of IP flows that was observed in particular network dependency.

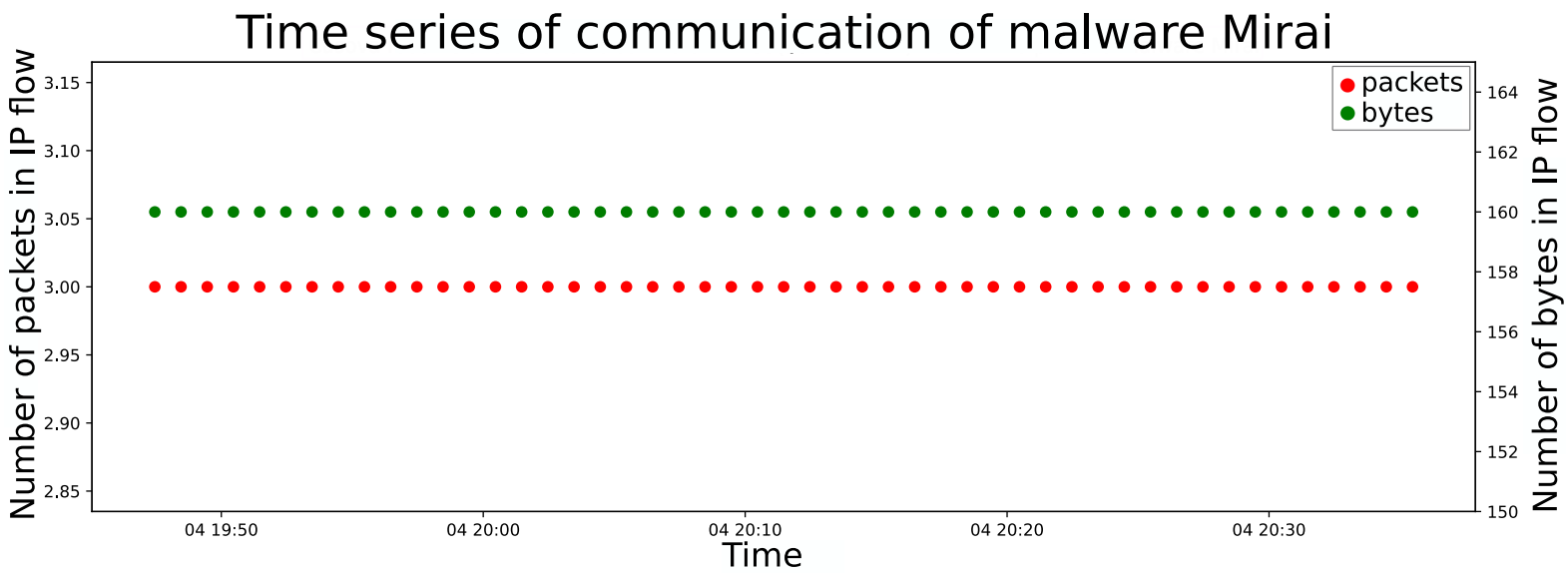
Why this type of time series?

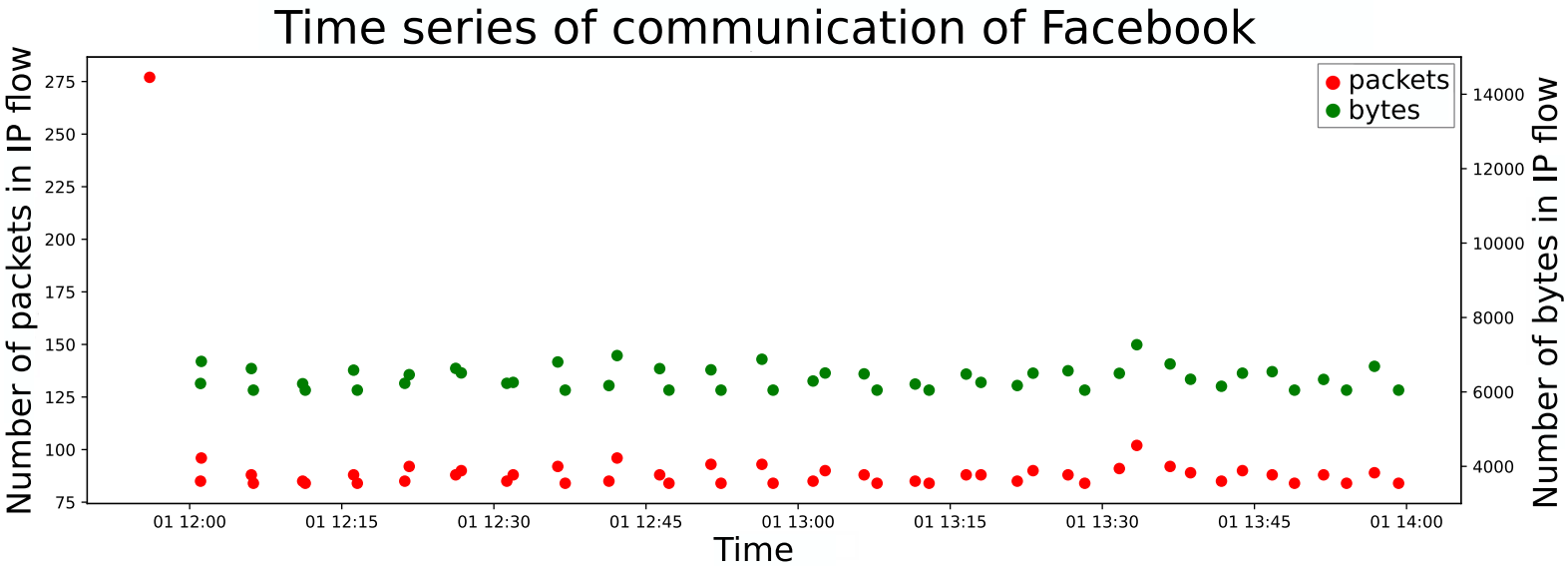
- We get traffic of process or set of processes that are (mostly) related to each other
- Each data point of time series is flow record
- We get time series with minimal noise

Time series from network traffic



Periodic behavior in time series from network traffic

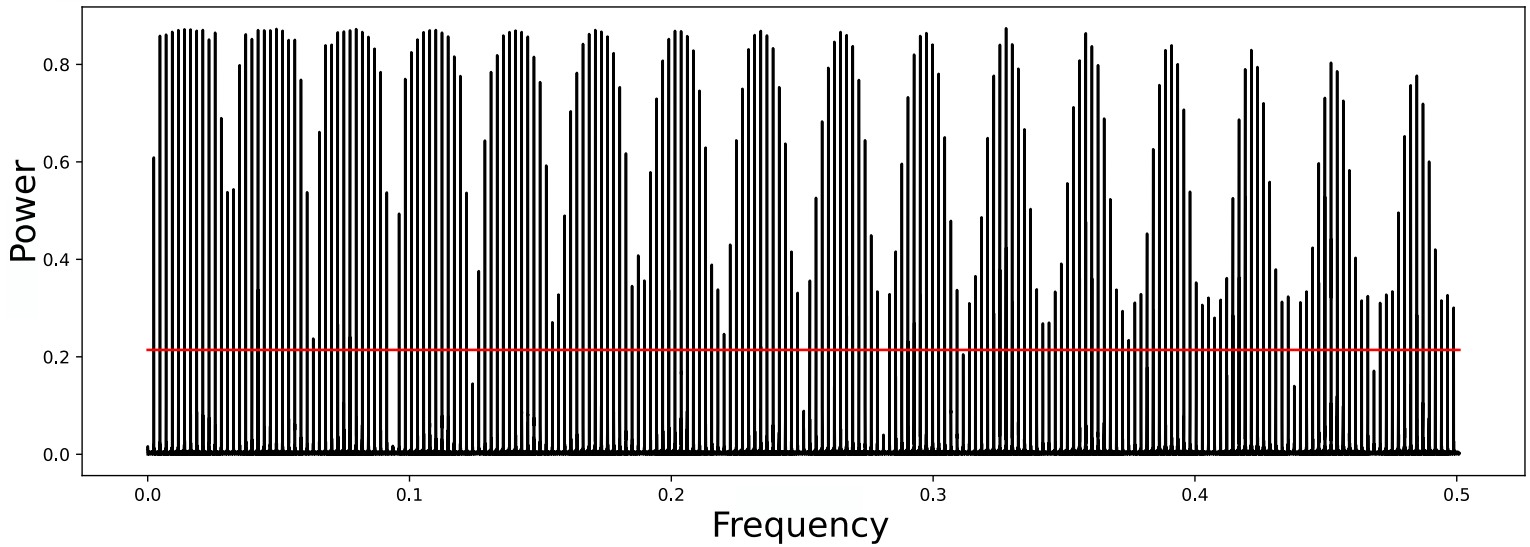




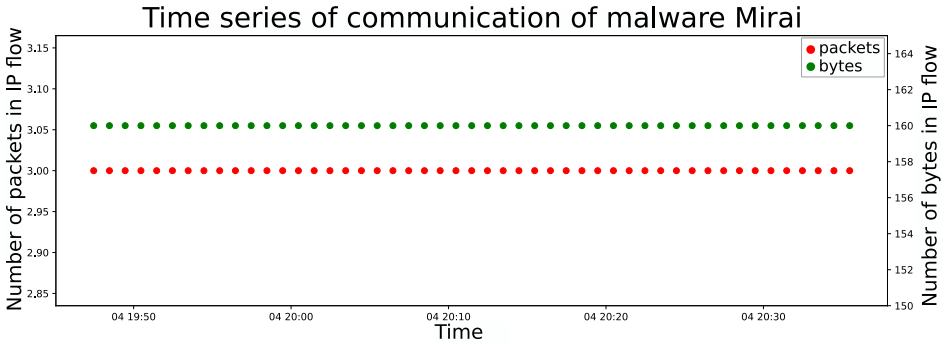
Detection of periodic behavior in network traffic time series

- The main component of our model is the Lomb-Scargle periodogram that allows the detection of periodic behavior in unevenly (or unequally or irregularly) spaced time series
- It is necessary to use statistical test of significance on periodogram, the best of tested was Scargle's Cumulative Distribution Function

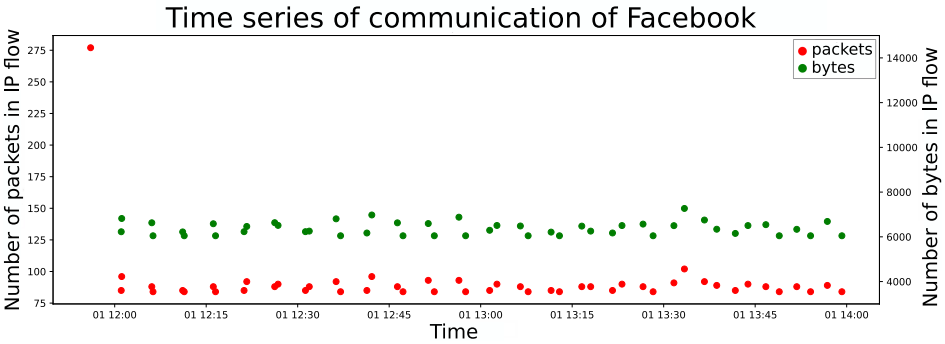
Lomb-Scargle periodogram of network traffic time series



Features of periodic behaviors



A natural description of periodic behavior in the plot above is a *number of flow records* that periodically repeat, a *time period*, a *number of packets*, and a *number of bytes in the flow*.



However, these features cannot describe a periodic behavior in the plot above, so we use *boundaries of the interval* of the data points.

Creating a dataset of periodic behaviors

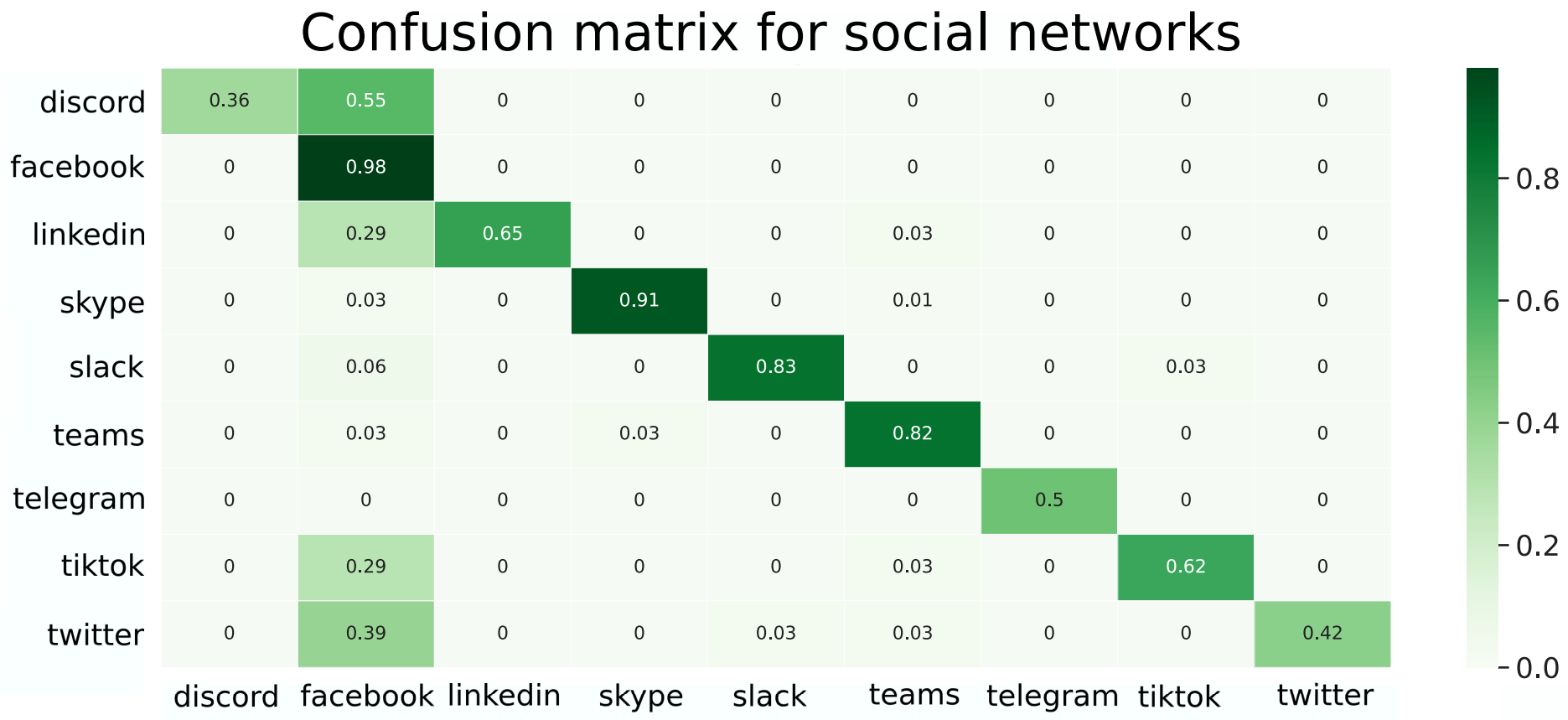
We use our model tu create dataset contains 30 thousands records tagged with 65 classification classes. For creation was used around 1 tb of network traffic. Communication was taken from multiple source - public available PCAPs, network CESNET2, NETMONLAB FIT CTU a home networks. Examples of classification groups and classes:

- network services and protocols (KeepAlive, HTTP2 Ping, ...)
- social networks (Facebook, Instagram, Twitter, ...)
- communication applications (Teams, Messenger, Discord, ...)
- version systems and clouds (Github, Google Drive, Dropbox, ...)
- email services and clients
- game clients (Steam, Epic Games Store, ...)
- antivirus software (Eset, Avast, Kaspersky, ...)

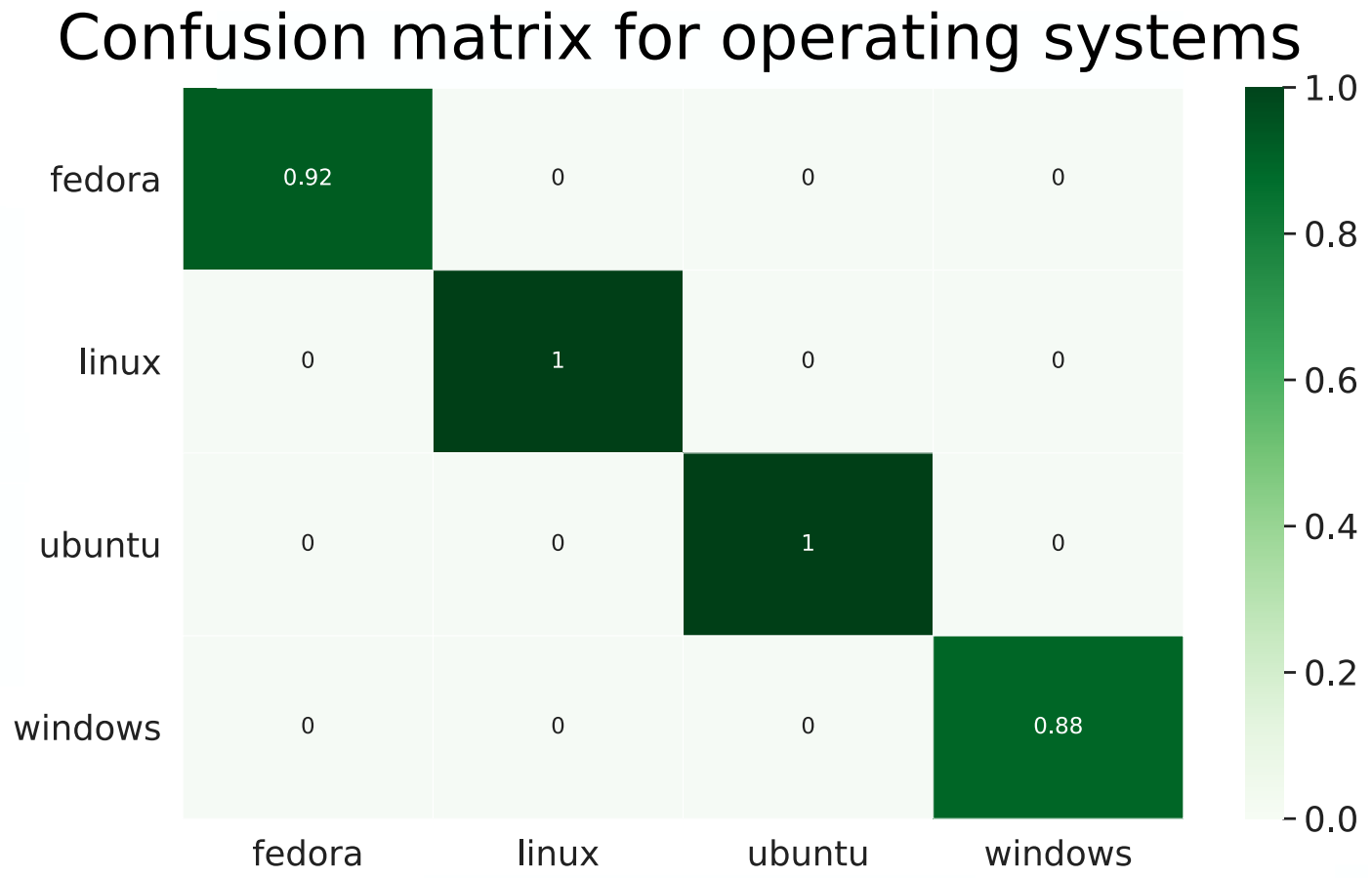
Results of trained classifiers

		Accuracy	Precision	Recall	F1-score
Naive Bayes	*macro avg*	8	16	20	10
	weighted avg	8	25	8	4
Logistic Regression	*macro avg*	42	1	2	1
	weighted avg	42	19	42	26
kNN	*macro avg*	62	50	40	41
	weighted avg	62	62	62	61
Extra Tree	*macro avg*	65	61	59	58
	weighted avg	65	66	66	66
Decision Tree	*macro avg*	77	45	46	45
	weighted avg	77	77	77	77
Random Forest	*macro avg*	89	87	64	72
	weighted avg	89	90	89	88
XGBoost	*macro avg*	90	82	66	71
	weighted avg	90	89	89	89

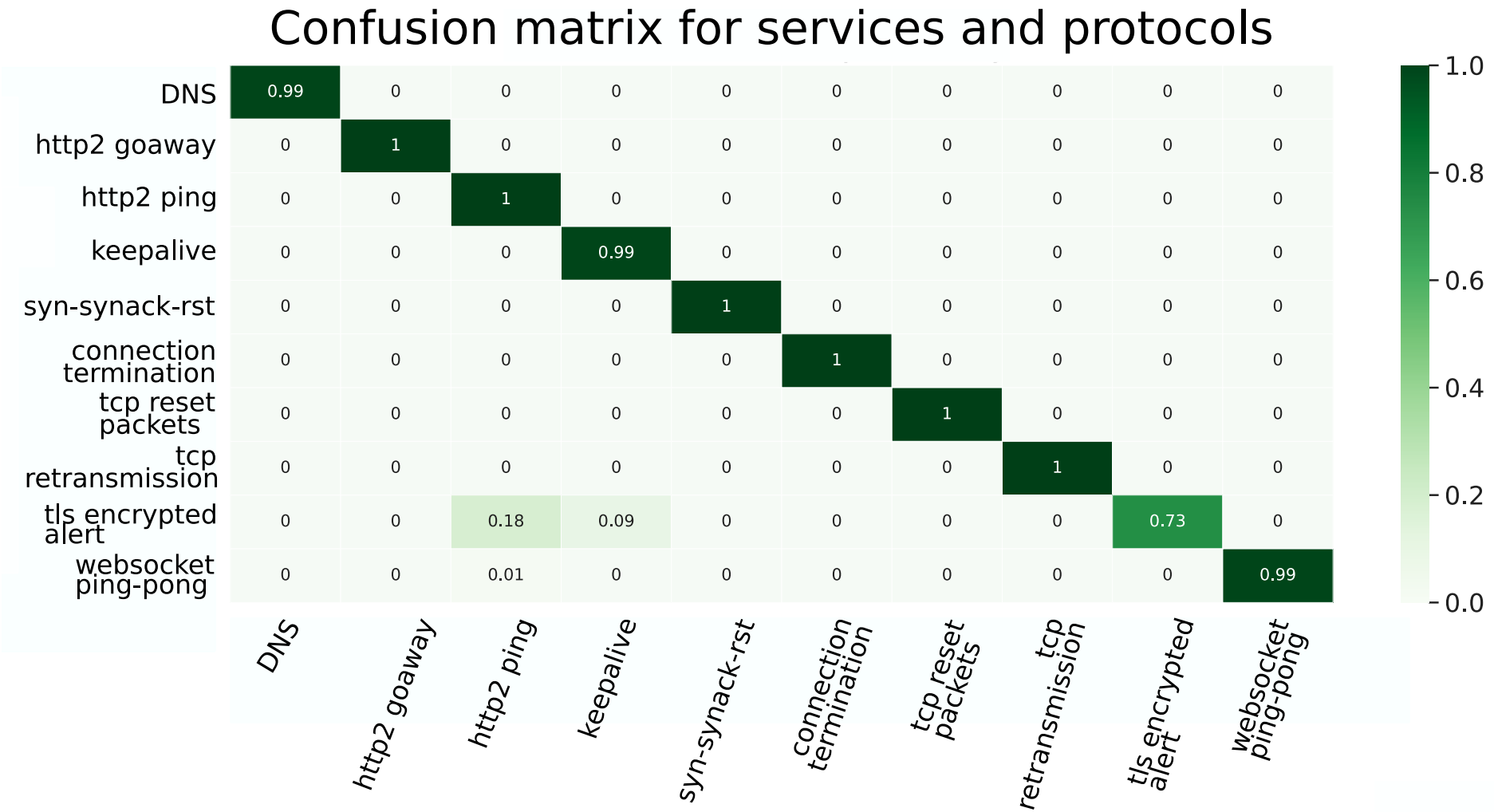
Results for classes group of social networks



Results for classes group of operating systems



Results for classes group of services



Conclusion

- We have found suitable mathematical tools from astrophysics to discover periodicity in the network traffic

Conclusion

- We have found suitable mathematical tools from astrophysics to discover periodicity in the network traffic
- We are able to train a classifier based on machine learning that exploits traffic statistic and reveal information about periodic behavior to recognize particular application that originated it

Conclusion

- We have found suitable mathematical tools from astrophysics to discover periodicity in the network traffic
- We are able to train a classifier based on machine learning that exploits traffic statistic and reveal information about periodic behavior to recognize particular application that originated it
- In total out best classification model can classify 61 types of traffic with F1-score 90%

Conclusion

- We have found suitable mathematical tools from astrophysics to discover periodicity in the network traffic
- We are able to train a classifier based on machine learning that exploits traffic statistic and reveal information about periodic behavior to recognize particular application that originated it
- In total out best classification model can classify 61 types of traffic with F1-score 90%
- In future work, we will focus on the applicability of model in high-speed networks, and also we will focus on increase number of classification classes and problems, for example botnet detection.