Visual Representation of Network Forensic Data

Ondřej Ryšavý
Jiří Hynek

Technical Report no. FIT-TR-2020-07
Faculty of Information Technology, Brno University of Technology

Abstract

Network forensics plays a critical role in incident handling and investigation. As there may be a huge amount of data that needs to be analyzed it is necessary to apply techniques that provide the easy to understand view on the data. The visual representation abstracts from the unnecessary details and offers various methods to observe the data from different angles. In this report, we discuss the possible visualization of selected network forensic data and use the Grafana framework for the demonstration of this approach.

This report was created with the financial support of Ministry of the Interior of the Czech Republic in frame of project ”Integrated platform for analysis of digital data from security incidents”, VI20172020062.

1 Introduction

Network forensics can provide information relevant to the system security and plays a critical role in incident handling and response. The incident forensic tool produces various data objects that can be presented to the user using various views. This document provides an overview of types of data objects generated by our Traffix tool and provides a set of possible visual representations driven by the proposed use cases. The purpose of the tool is to provide a quick overview of the data interested for digital forensics. The tool is able to store full packet capture at massive scale and provides a fast access to captured data. Also, the tool understand to various protocols so it performs packet header analysis, file extraction, session viewing and application protocol analysis. The visualization frontend of the tool enables the user to identify the timeframe or location of the data worth for further analysis. Of course there may be a lot of different scenarios depending on the particular case. In this report, we focus on several general use cases and for each we demonstrate the possibility of creating a visual information based on the extracted data to aid the process of data analysis and evidence identification.

1.1 Incident Investigation

Incident handling is related to a specific incident or based on some suspicious activities. The investigation is initiated by raised alarm or required as a supporting activity during the investigation. Packet captures are one of the richest from of source data for the investigation though because of the possible huge size of captured data also quite difficult to analyse. Traditionally network traffic analysis tools, such as Wireshark have been applied. The specialized tools, e.g, NetworkMiner, can aid with the investigation to automatically extract and provide information with the highest forensic value. Such information may consists of identified hosts, users, network applications, services used, and content transferred. Identified events and communication together with extracted artifacts can be further analyzed using various methods, e.g, temporal analysis, relation analysis, etc.

1.2 Network Traffic Monitoring

To identify the suspicious activities in the network, traffic monitoring approach is adopted. Usually this can be done automatically by various NIDS tools and other network monitoring applications. Usually, live monitoring is deployed and it is commonly the part of enterprise security architecture. Once the laert is identified some response may be taken. Usually, also packet capture is created containing data related to the occurence of the event. The capture data can be further used in the investigation by analysis of relevant events and identifying artifacts. As the first step, the overview of the communication at flow level is generated and relevant flows and events are then further investigated.

1.3 Data Exploration

In general terms the data exploration means to analyse the data without a prior indication of the suspicious activity nor any other information about compromise. As there are a lot of data that needs to be explored it is necessary to automate the process and reduce the amount of information to be analyzed. Today, most of the data transfers are encrypted using TLS protocol. Thus, to identify the activity we need to understand and identify the encrypted communication and use other channels providing useful information, e.g., domain name system, log files, control protocol communication, etc.

2 Data Sources

Several data and views are considered for the visualization. Based on the identified use cases we provide and overview of data source generated by the Traffix console is provided. The console exports data in InfluxDB format to be imported into the InfluxDB database, which is then used as a data source by Grafana.

2.1 Network Conversations

Network conversations provide the basic view on the network communication data. The overview of the captured packet dump should be provided to the investigator to help her with the understanding of the character of data to be analyzed.
It provides a different level and type of information:

- Flows - network conversation consists of a pair of flows. Each flow is identified by the flow key and has assigned statistical information. Flow key consists of transport protocol type, source and destination endpoint information.
- Hosts - based on conversation data, it is possible to compute aggregation of network communication for individual hosts or group of hosts.
- Applications - the conversation has also assigned the identified application name. The application name usually corresponds to application protocol.

Conversations

The *conversations* collection provides a basic overview of the captured traffic organized as bidirectional flows.For each flow a basic set of values is provided, e.g., octets, packets, etc.

Field	Kind	Description
start	Time	The timestamp of the first packet of the conversation.
application	Tag	Identified application using a combination of port numbers and DPI methods. For known application it is the name of the service otherwise it is represented as a port number.
protocol	Tag	The name of the protocol.
client	Tag	The source IP address of the flow.
server	Tag	The destination IP address of the port.
domains	Field	A list of domains associated with this flow.
duration	Field	Duration of the conversation.
flowKey	Field	The string representation of the conversation, e.g.: Udp$192.168.5.122:40080->209.85.51.222:53

fwdPackets	Field	The total number of packets of the forward flow.
fwdOctets	Field	The total bytes of the forward flow.
revPackets	Field	The total number of packets of the reverse flow.
revOctets	Field	The total bytes of the reverse flow.

2.2 Extracted Files

This case considers the possibility of extraction of content from network communication in case of unencrypted data exchange. The content is extracted on demand but it should be possible to list all available objects for extraction. It shall be possible to extract content from various plain text protocols, e.g., HTTP, FTP and SMB.

Files

The *files* table represents a meta information about the identified filed transferred by various plain text protocols.

Field	Kind	Description
timestamp	Time	The time of occurrence of the flow.
application	Tag	The name of the application protocol for which the content was identified and extracted, e.g., http.
client	Tag	The source IP address of the flow.
contentType	Tag	The content type using MIME representation, e.g., “text/plain”.
fileType	Tag	The type of the file as observed in file extension.
server	Tag	The destination IP address of the flow.
contentLength	Field	The length of the content’s data.
contentName	Field	The name of the content, e.g., the file name or full URL for HTTP objects.
exportedPath	Field	Path to a container file with the exported object. This information can be used for localize the file itself.
fileTypeMismatch	Field	An indication whether content type agrees with the file type.
flowKey	Field	Flow information related to the event. Events are identified in packets of a flow. The flow information helps to create a context of the issue.

2.3 Encrypted Traffic

The Traffix engine also collects information about encrypted communication. Currently, the TLS communication is recognized and meta information extracted from the TLS handshake, e.g, Server Name Indicator, JA3 values, associated domain names, certificate information, etc.

Encrypted
The *encrypted* collection provides information on each TLS secure channel established in the source data.

Field	Kind	Description
timestamp	Time	Corresponds to FirstSeen field.
client	Tag	The IP address of the client.
server	Tag	The IP address of the server.
alpn	Field	The application protocol used.
cipherSuite	Field	The name of cipher suite selected for the channel.
duration	Field	Duration of the conversation.
flowKey	Field	The key string of the forward flow.
fwdRecords	Field	The total number of TLS records in the forward (client to server) flow.
ja3Client	Field	The fingerprint of the TLS client.
ja3Server	Field	The fingerprint of the TLS server.
revRecords	Field	The total number of TLS records in the reverse (server to client) flow.
serverName	Field	Value of SNI in TLS handshake.
version	Tag	The TLS version, e.g., “Tls12”, “Tls13”

2.4 Resolved Domains

The domains table summarizes DNS queries and responses and can provide a valuable information on the communication. For instance, it can be possible to apply various methods for detection malware activity or indices of phishing attacks.

Domains

The *domains* collection provides a list of all domain names requested to resolve and the resolved addresses.

Field	Kind	Description
timestamp	Time	The timestamp of the domain, i.e., the timestamp of DNS query message containing the domain.
alexa	Tag	Indication whether the domain name belongs to some of the alexa dataset.
client	Tag	The IP address of the client.
dga	Tag	A possible DGA used to generate the domain. This is computed by the DGA classifiers.
category	Tag	The category as identified using whitelist/blacklist methods.
domain	Tag	Domain name to resolve.
status	Tag	The status of the query can be NOERROR, FORMERR, SERVFAIL, NXDOMAIN, REFUSED, OTHERERROR
flowKey	Field	The flow key of the DNS communication.
resolvedTo	Field	The IP addresses resolved for the domain.
rtt	Field	The round trip time in seconds.
server	Field	The IP address of the server.

3 Architecture

We found the best option to present the data in the form of several dashboards where every dashboard is used for a particular use case described in Chapter 2. Dashboards are visualization tools that emphasize the most important information on a single screen so the user can perceive all the information at a glance[1]. It uses the advantage of graphical visual representation of data which saves the space and improves the preattentive processing of visual objects. Viewers are able to quickly understand relationships between data in contrast to textual representation[2].

There are three options to design and implement dashboards. Either the designer can create an in-house solution which would fulfil all the requirements but takes resources, or he can use a 3rd party commercial service which would provide tools for quick prototypes of dashboards. Usually, such services does not share the source code and the dashboards are stored on the 3rd party servers. There is a compromise between those approaches—use one of the available open-source analytics tools. We choose Grafana system. It provides a set of charting widgets which can be placed on a canvas and connected with data. It works as a CMS (*content management system*) which can be deployed on own servers. On the other hand, users are limited by the possibilities of the system. There might be problems to implement some uncommon use cases. Details are described in Section 3.2.

Since Grafana tool works with time series, every item in a database should have a timestamp. The best option is to connect Grafana with one of the *time series databases* (TSDB) which optimize storing and providing time series data. We choose InfluxDB system, described in section 3.1.

3.1 Database

We used InfluxDB, v. 1.8.0, under the MIT license, to store the data. InfluxDB is a non-relational database system which represents data in the form of measurements which store data in associated fields. Every item of the associated field is represented by a timestamp, tags (finite set of categorical data which can be used for aggregations) and fields (values of some data type). To store the data needed for visuzalition we created four measurement collections represented as InfluxDB tables:

- **Network Conversations** represent a bidirectional flow found in the packet captures. The data is stored in *conversations* table. The structure of the table is defined in Section 2.1.
- **Extracted Files** provide a list of files identified in plain text application protocols whose content can be extracted. The measurements are stored in *files* table. The structure of the table is defined in Section 2.2.
- **Encrypted Traffic** provide metadata about TLS flows stored in table *encrypted* that have the structure described in Section 2.3.

[1] Few, S. (2006). Information dashboard design: The effective visual communication of data. O'Reilly Media, Inc.

[2] Tufte, E. R. (2001). The visual display of quantitative information. Cheshire, CT: Graphics press.

- **Resolved Domains** track DNS messages in order to provide information on the resolved domain names. The data is stored in *domains* table which structure is described in Section 2.4..

One of the issues we had to deal with was performance of the database system. InfluxDB is limited by cardinality to 1 million items by default. Configuration of the system needs to be changed when a higher level of cardinality is needed. Performance of such database is however low. Hence, we had to optimize the number of items in measurements and characteristics of the items.

3.2 Front-end

Grafana, version 6.7.2, under the Apache 2.0 license, was used to create information system providing user interface to users. The system provide complete management of the content, as management of users, database sources, dashboards, and various settings. It provides two themes—dark and light—which increases usability of the system. Predefined palette of widgets is presented in Figure 3.1. Every widget can be customized via graphical editor and connected with the database via query of the particular DB system. Grafana provides numbers of database connectors.

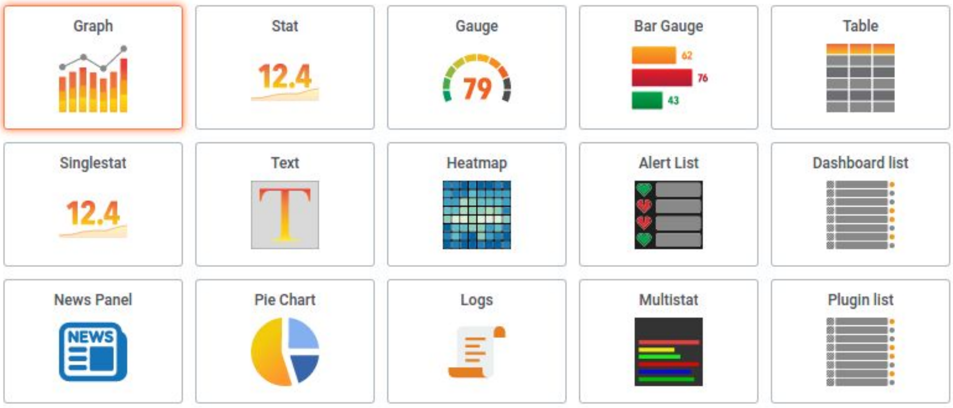


Figure 3.1: List of widgets provided by the Garfana system. We used two additional plug-ins for non-time series data: Multistat (bar chart) Pie Chart.

Grafana provides ability to prefetch data into Grafana's variables from the database systems and use the data multiple times in various widgets. This can save number of queries sent to database system, which can help with performance issues of the database system. Also it can provide ability to parametrize widgets. This functionality was used to provide ability to switch between time series (e. g., forward/backward packets) or change parameters (e. g., IP address).

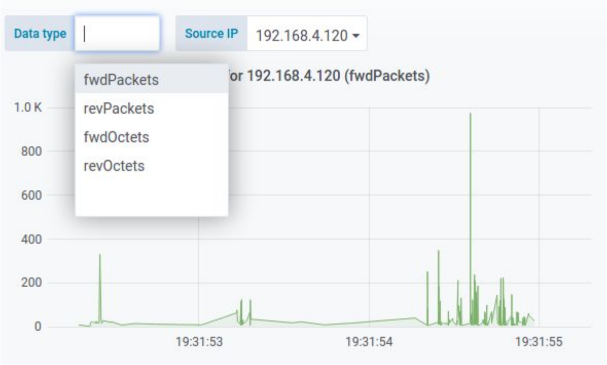


Figure 3.2: An example of variables settings. Variables can be applied in widgets' queries. Such widgets are updated when the values of variables are changed.

3.3 Deployment

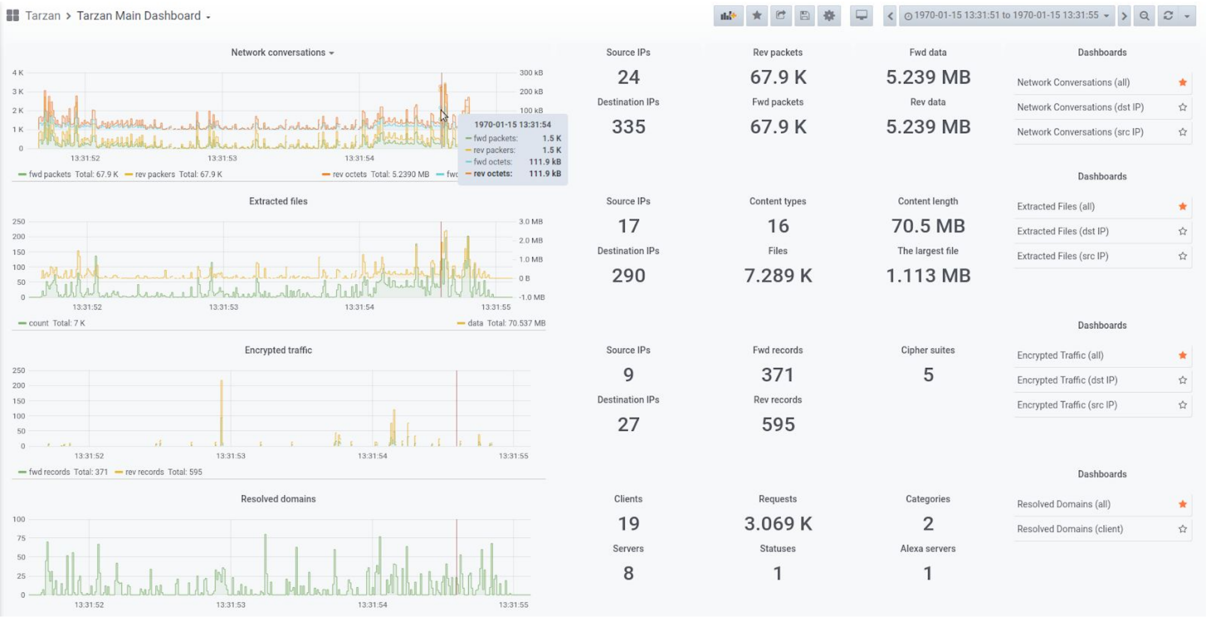
The dashboards designed in Grafana were exported in JSON format. This representation can be imported into another instance of Grafana running on a different machine. This instance of Grafana only needs to have an access to the same database server or a database server containing the same database structure. We used Docker to deploy the tools. We imported the database on InfluxDB server running on a Docker container. The database server is available on the port 8086. Then, another Docker container was created for Grafana server which is available on the port 3000. The dashboards' JSON files were imported to the server and re-connected with the database.

4 Data Visualization

We decided to display the data in dashboards. The typical characteristic of dashboards is that they display the data on a single screen, which is based on Few's definition of dashboard[3]. Grafana's dashboard layout is based on the responsive grid, which tries to use 100% of screen width and rearranges widgets automatically into one column if the resolution is low. However, we still needed to deal with the problem of limited height. Since multiple use cases using various kind of data were determined in the data analysis (Section 2), we created several dashboards fitting one screen used for specific use cases.

4.1 Main Dashboard

The first view the user should use is the main dashboard. Goal of the main dashboard is to summarize all the data across all the use cases. This perspective helps the user to get familiarized with the data basis and notice important parts (e.g. anomalies in the data traffic) which should be analyzed. An example of data visualized in the main dashboard is presented in Figure 4.1.



Source IPs 24
Destination IPs 335
Rev packets 67.9 K
Fwd packets 67.9 K
Fwd data 5.239 MB
Rev data 5.239 MB

Network Conversations (all) ★
Network Conversations (dst IP) ☆
Network Conversations (src IP) ☆

Source IPs 17
Destination IPs 290
Content types 16
Files 7.289 K
Content length 70.5 MB
The largest file 1.113 MB

Extracted Files (all) ★
Extracted Files (dst IP) ☆
Extracted Files (src IP) ☆

Source IPs 9
Destination IPs 27
Fwd records 371
Rev records 595
Cipher suites 5

Encrypted Traffic (all) ★
Encrypted Traffic (dst IP) ☆
Encrypted Traffic (src IP) ☆

Clients 19
Servers 8
Requests 3.069 K
Statuses 1
Categories 2
Alexa servers 1

Resolved Domains (all) ★
Resolved Domains (client) ☆

Figure 4.1: An example of the main dashboard composed of four types of data representing the uses cases described in Chapter 2. The time range filter is common for all widgets.

The dashboard is vertically divided into sections representing the four uses cases displaying the following measurements: network conversations, extracted files, encrypted conversations, and resolved domains. Every section composes of a stair-step chart

[3] Few, S. (2006). Information dashboard design: The effective visual communication of data. O'Reilly Media, Inc.

displaying traffic overview, numeric indicators presenting summary of traffic characteristics and a table of references to dashboards providing detailed perspectives of the use cases.

The stair-step chart displays time-series data for a selected time range. The data records are grouped by the predefined Grafana time interval. Widgets contains several series depending on the measurement (e.g. forward or reverse packets for the flows measurement). Every widget can be enlarged to the fullscreen mode and the time series contained by the widgets can be filtered as presented in Figure 4.2.

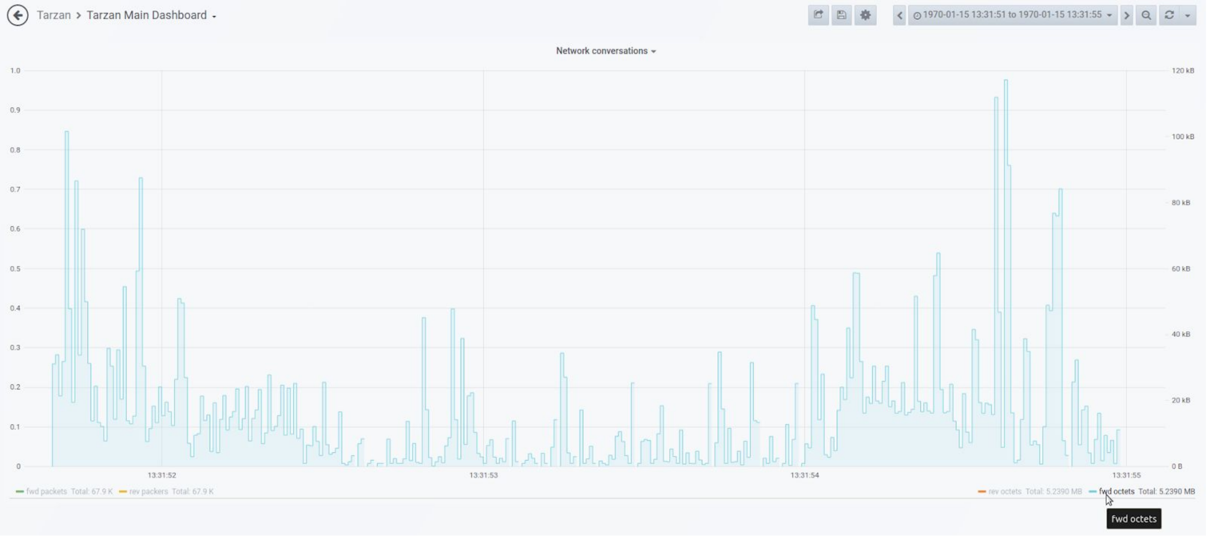

Tarzan > Tarzan Main Dashboard
1970-01-15 13:31:51 to 1970-01-15 13:31:55
Network conversations
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
120 kB
100 kB
80 kB
60 kB
40 kB
20 kB
0 B
13:31:52
13:31:53
13:31:54
13:31:55
fwd packets Total: 67.9 K
rev packers Total: 67.9 K
rev octets Total: 5.2390 MB
fwd octets Total: 5.2390 MB
fwd octets

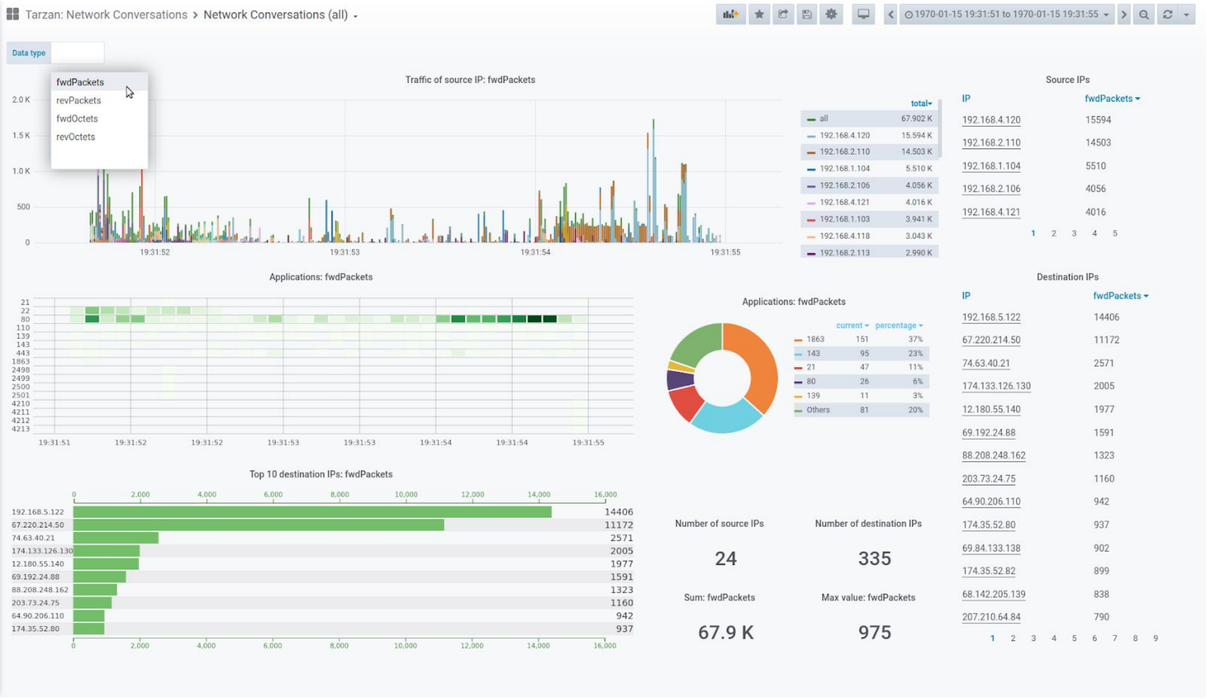
Figure 4.2: An example of a time series filtration. The user can select particular series in the legend location in the bottom of the chart.

The numeric indicators present values of specific characteristics (e.g. number of source and destination IPs, amount of transferred data, files, etc.). The indicators are synchronised with the global time range filter. Hence, change of time range in the stair-step charts affects presented values of all the numeric indicators.

Finally, there are the tables containing links to further dashboards which display data of the measurements on a deeper level of detail. Following sections discuss the meaning of particular charts and indicators for every measurement.

4.2 Network Conversations

Visualization of communication between IP addresses is an important data perspective. It is represented by the set of bidirectional (forward and reverse) flows (list of packets). It helps the user to understand who communicates with whom, which applications they use and how many packets or octets they transferred. Figure 4.3 represents the dashboard designed for this purpose. As every Grafana-based dashboard, it contains the implicit time range filter which can be used to focus on a subset of data, as displayed in Figure 4.4.


Tarzan: Network Conversations > Network Conversations (all)
1970-01-15 19:31:51 to 1970-01-15 19:31:55
Data type
fwdPackets
revPackets
fwdOctets
revOctets
Traffic of source IP: fwdPackets
Applications: fwdPackets
Top 10 destination IPs: fwdPackets

	total
all	67.902 K
192.168.4.120	15.594 K
192.168.2.110	14.503 K
192.168.1.104	5.510 K
192.168.2.106	4.056 K
192.168.4.121	4.016 K
192.168.1.103	3.941 K
192.168.4.118	3.043 K
192.168.2.113	2.990 K

Source IPs

IP	fwdPackets
192.168.4.120	15594
192.168.2.110	14503
192.168.1.104	5510
192.168.2.106	4056
192.168.4.121	4016

1 2 3 4 5

Applications: fwdPackets

	current	percentage
1863	151	37%
143	95	23%
21	47	11%
80	26	6%
139	11	3%
Others	81	20%

Destination IPs

IP	fwdPackets
192.168.5.122	14406
67.220.214.50	11172
74.63.40.21	2571
174.133.126.130	2005
12.180.55.140	1977
69.192.24.88	1591
88.208.248.162	1323
203.73.24.75	1160
64.90.206.110	942
174.35.52.80	937
69.84.133.138	902
174.35.52.82	899
68.142.205.139	838
207.210.64.84	790

1 2 3 4 5 6 7 8 9

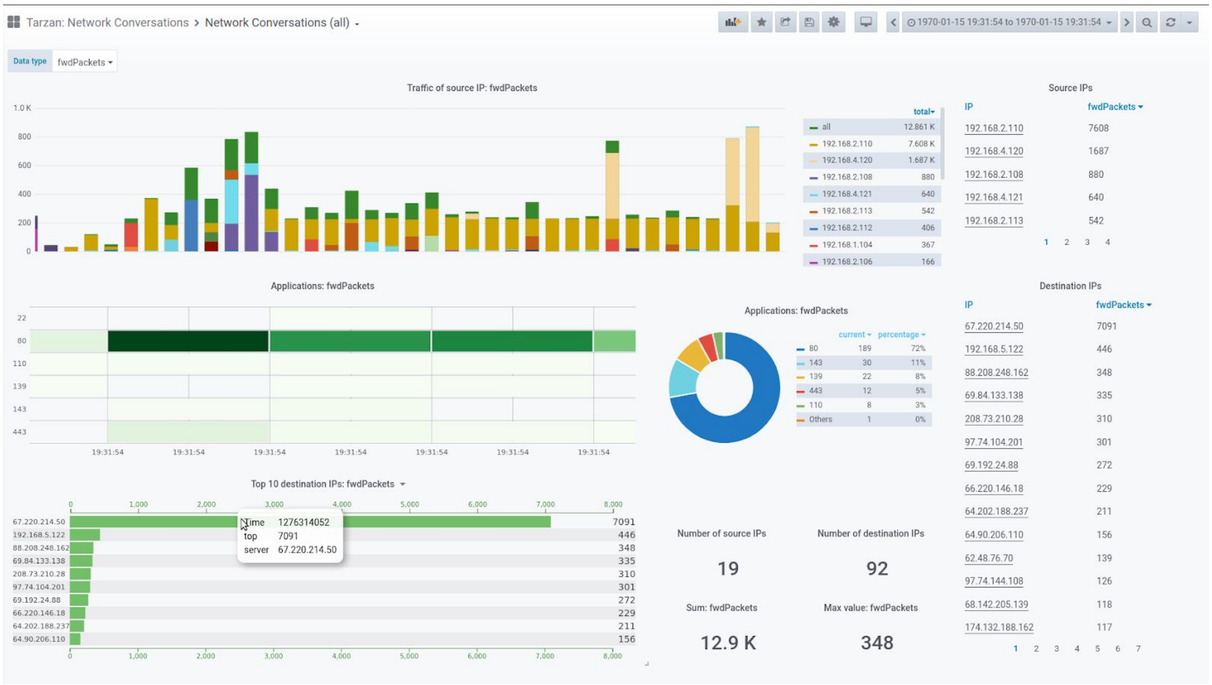
Number of source IPs: 24

Number of destination IPs: 335

Sum: fwdPackets: 67.9 K

Max value: fwdPackets: 975

Figure 4.3: The main dashboard used for visualization of overall communication between IP addresses. The user can choose between four time series: forward/reverse packets and forward/reverse octets using Grafana's variables (the top left part of the figure).


Tarzan: Network Conversations > Network Conversations (all)
1970-01-15 19:31:54 to 1970-01-15 19:31:54
Data type
fwdPackets
Traffic of source IP: fwdPackets
Applications: fwdPackets
Top 10 destination IPs: fwdPackets
Time 1276314052
top 7091
server 67.220.214.50

	total
all	12.861 K
192.168.2.110	7.608 K
192.168.4.120	1.687 K
192.168.2.108	880
192.168.4.121	640
192.168.2.113	542
192.168.2.112	406
192.168.1.104	367
192.168.2.106	166

Source IPs

IP	fwdPackets
192.168.2.110	7608
192.168.4.120	1687
192.168.2.108	880
192.168.4.121	640
192.168.2.113	542

1 2 3 4

Applications: fwdPackets

	current	percentage
80	189	72%
143	30	11%
139	22	8%
443	12	5%
110	8	3%
Others	1	0%

Destination IPs

IP	fwdPackets
67.220.214.50	7091
192.168.5.122	446
88.208.248.162	348
69.84.133.138	335
208.73.210.28	310
97.74.104.201	301
69.192.24.88	272
66.220.146.18	229
64.202.188.237	211
64.90.206.110	156
62.48.76.70	139
97.74.144.108	126
68.142.205.139	118
174.132.188.162	117

1 2 3 4 5 6 7

Number of source IPs: 19

Number of destination IPs: 92

Sum: fwdPackets: 12.9 K

Max value: fwdPackets: 348

Figure: 4.4: Application of the time range filter to zoom into a subset of data.

The dashboard contains the following widgets:

- **Stacked bar displaying overall traffic:** every bar of the chart displays the sum of the forward/reverse packets/octets depending on the user's selection for the Grafana's time unit. The users can see the overall traffic or filter specific source IP addresses (Figure 4.5). They can use this chart as an entry point of the dashboard. The chart helps to locate the important time range. The users can zoom into the time range and other charts and indicators are updated accordingly (Figure 4.4).

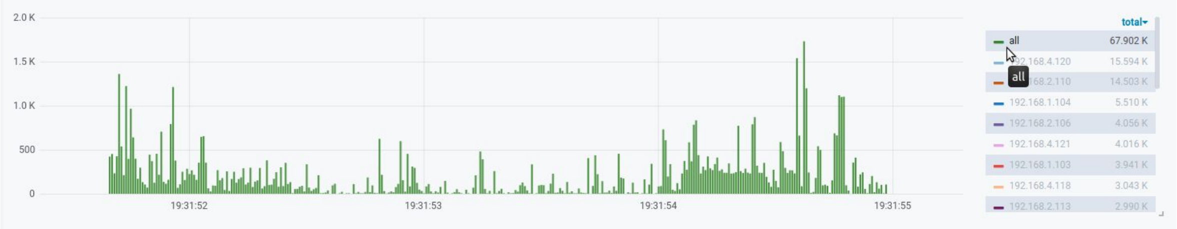


Figure 4.5: The chart visualizes traffic in time. Users can filter specific IP addresses by clicking the items of the legend. Also, the user can select multiple items using the CTRL key. Items can be sorted with respect to their values, so the users can focus on the dominant IP addresses only.

- **Heatmap displaying used applications (ports):** then, the users can distinguish dominant applications clients have used. Every item (bucket) of the heatmap represents sum of values for the Grafanas's time unit. Items of a higher sum are colored by a more intensive color. Hence, the users can locate the dominant applications in time very quickly (Figure 4.6).

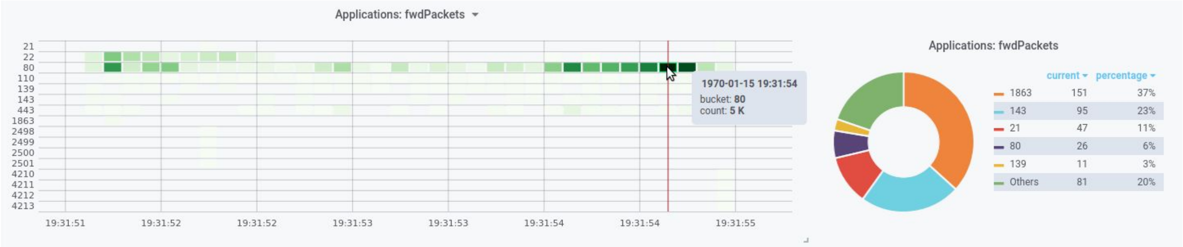


Figure 4.6: Visualization of used applications. Heatmap (left) emphasizes dominant applications in time. The donut chart (right) compares the dominant applications.

- **Donut chart displaying share of dominant applications (ports):** the users can compare the usage of dominant applications using the donut chart (Figure 4.6). Parts of the donut charts represent one of the applications with the share higher than 2%. Other applications are grouped in the "Others" item. The users can filter specific items similarly as in the stacked bar chart displaying overall traffic.[4]

[4] Using donut charts is usually not recommended by design guidelines since it can complicate understanding of data by users (Few, S., 2006. Information dashboard design: The effective visual communication of data. O'Reilly Media, Inc.). However, Grafana does not provide many options for visualization of non-time data series. Hence, we used Pie Chart plugin provided by Grafana Labs. This could be improved in the future.

- **Bar chart providing statistics of the top destination IPs:** one of the users' requirements is to distinguish dominant destination IPs. Bar chart provides quick access to the top 10 IPs.[5] The dominant IP addresses can be used for further analyses of the communication. Users are aware which IP addresses they should be focused on.

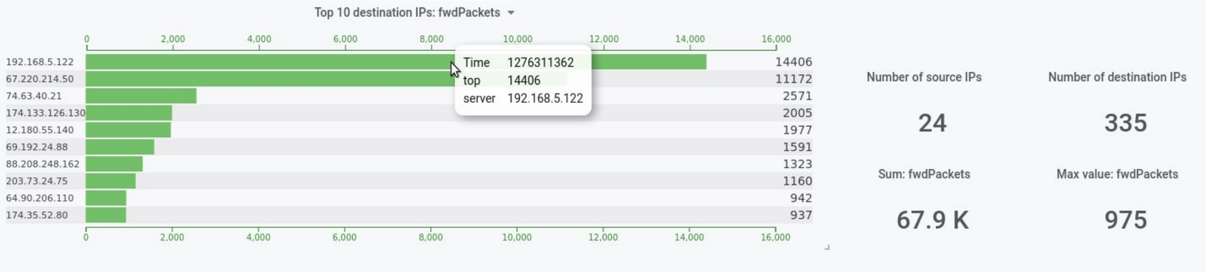
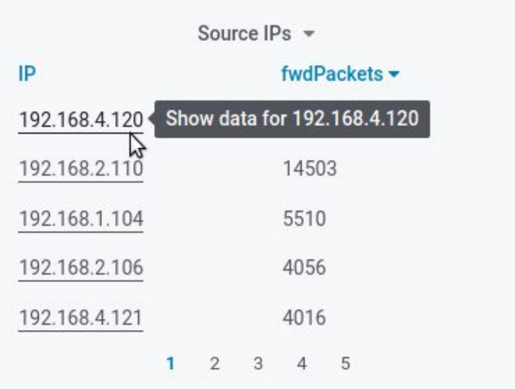


Figure 4.7: Example of the top statistics of destination IP addresses using bar chart (left) and overall indicators (right) representing numbers of (source and destination) IP addresses, summary and maximum value of the data series. Charts are synchronized with the time range filter.

- **Statistics indicators:** they can be used for monitoring overall information regarding the traffic of the selected time range. The users should know how many source and destination IP addresses participated the conversations. Also, the users can see how many packets or data have been transferred (depending on the selected time series: forward/reverse packets/octets) and see the maximum value of the time range. Examples of the indicators can be seen in Figure 4.7.
- **Tables displaying source and destination IP addresses:** finally, the users might be interested in a specific source or destination IP address. They can use the tables of the IP addresses (Figure 4.8) which provide ability to sort items according to their values. Users can click on an IP address and they are redirected to the dashboard presenting data for the IP address.



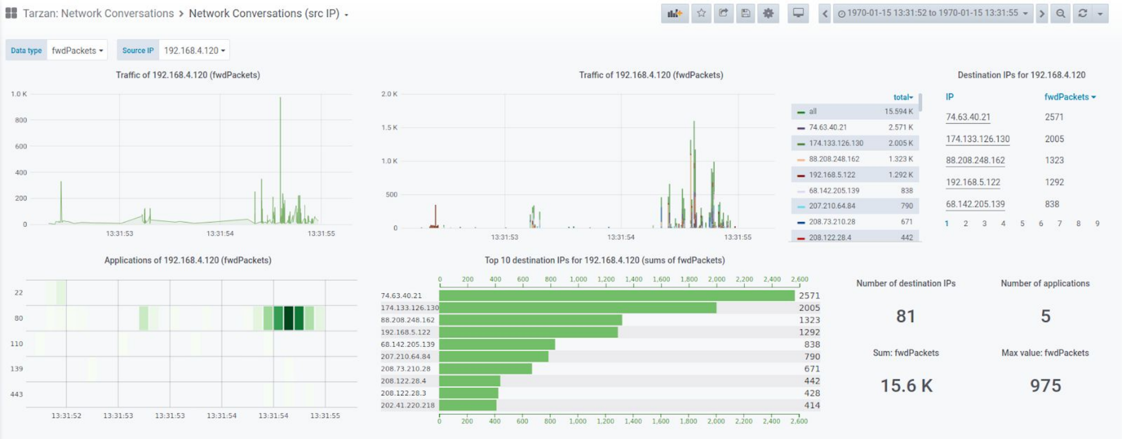
Source IPs

IP	fwdPackets
192.168.4.120	Show data for 192.168.4.120
192.168.2.110	14503
192.168.1.104	5510
192.168.2.106	4056
192.168.4.121	4016

Figure 4.8: The table of IP addresses providing hyperlinks to further data perspectives.

[5] Bar chart used for visualization top statistics represents another example non-time data series. We used michaeldmoore's Multistat plugin.

When the user clicks IP address hyperlink, a detailed view is opened. It is similar to the overall dashboard but it contains Grafana's variable allowing to change IP address (Figure 4.9). Users can analyze the data regarding the IP address. When a source IP address is selected, communication with corresponding destination IP addresses is presented and vice versa. The dashboard contains line chart displaying the overall traffic of the selected IP address, stacked bar chart displaying the distribution of the traffic among the corresponding IP addresses, heat map displaying usage of applications (ports), bar charts presenting top corresponding addresses, statistics indicators and the table of corresponding IP address hyperlinks which can be used to switch to their data perspective.



Tarzan: Network Conversations > Network Conversations (src IP)

1970-01-15 13:31:52 to 1970-01-15 13:31:55

Data type: fwdPackets

Source IP: 192.168.4.120

Traffic of 192.168.4.120 (fwdPackets)

	total
all	15.594 K
74.63.40.21	2.571 K
174.133.126.130	2.005 K
88.208.248.162	1.323 K
192.168.5.122	1.292 K
68.142.205.139	838
207.210.64.84	790
208.73.210.28	671
208.122.28.4	442

Destination IPs for 192.168.4.120

IP	fwdPackets
74.63.40.21	2571
174.133.126.130	2005
88.208.248.162	1323
192.168.5.122	1292
68.142.205.139	838

1 2 3 4 5 6 7 8 9

Number of destination IPs: 81

Number of applications: 5

Sum: fwdPackets: 15.6 K

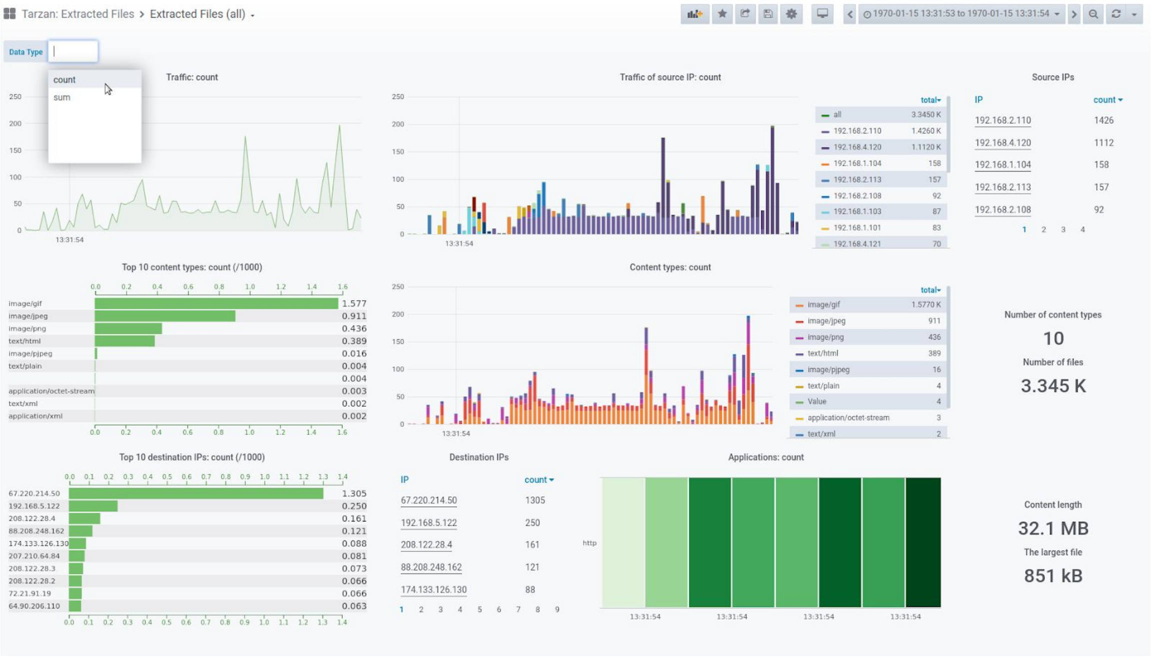
Max value: fwdPackets: 975

Figure 4.9: An example of the dashboard presenting data regarding overall traffic for the specific source IP address (192.168.2.120).

The disadvantage of the chosen approach which works with stand-alone dashboard for overall traffic and source/destination IP address is that the structure of these three dashboards is redundant. All three dashboards could be joined into one which would allow filter either source or destination IP address optionally. User could focus on specific IP when needed. This approach was, however, limited by capabilities of Grafana and InfluxDB. We could consider such change of the design as one of possible improvements in future.

4.3 Extracted Files

Besides overall traffic, it is important to analyze the information about files extracted from communication. We can monitor either the number of files or the amount of data. For this purpose, a single dashboard has been created. The dashboard uses Grafana's variable to let the users choose between two functions: "count" (the number of files) and "sum" (the amount of data). Then, the users can analyze the traffic, IP adresses (source and destination) and applications similarly as described in Section 4.2. Also, the dashboard provides information about the content types which can help understanding characteristics of the communication (what kind of files it was transferred the most). The data are presented in the form of bar chart providing quick statistics of the most occured content types, and in the form of stacked bar chart which can filter specific content types in time (similarly as it is done for source IP addresses). An example of the dashboard is presented in Figure 4.12.



Tarzan: Extracted Files > Extracted Files (all)
1970-01-15 13:31:53 to 1970-01-15 13:31:54
Data Type
count
sum
Traffic: count
Traffic of source IP: count
Content types: count
Top 10 content types: count (/1000)
Top 10 destination IPs: count (/1000)
Applications: count
http
13:31:54

	total
all	3.3450 K
192.168.2.110	1.4260 K
192.168.4.120	1.1120 K
192.168.1.104	158
192.168.2.113	157
192.168.2.108	92
192.168.1.103	87
192.168.1.101	83
192.168.4.121	70

Source IPs

IP	count
192.168.2.110	1426
192.168.4.120	1112
192.168.1.104	158
192.168.2.113	157
192.168.2.108	92

1 2 3 4

	total
image/gif	1.5770 K
image/jpeg	911
image/png	436
text/html	389
image/pjpeg	16
text/plain	4
Value	4
application/octet-stream	3
text/xml	2

Top 10 content types: count (/1000)

image/gif	1.577
image/jpeg	0.911
image/png	0.436
text/html	0.389
image/pjpeg	0.016
text/plain	0.004
	0.004
application/octet-stream	0.003
text/xml	0.002
application/xml	0.002

Number of content types

10

Number of files

3.345 K

Top 10 destination IPs: count (/1000)

67.220.214.50	1.305
192.168.5.122	0.250
208.122.28.4	0.161
88.208.248.162	0.121
174.133.126.130	0.088
207.210.64.84	0.081
208.122.28.3	0.073
208.122.28.2	0.066
72.21.91.19	0.066
64.90.206.110	0.063

Destination IPs

IP	count
67.220.214.50	1305
192.168.5.122	250
208.122.28.4	161
88.208.248.162	121
174.133.126.130	88

1 2 3 4 5 6 7 8 9

Content length

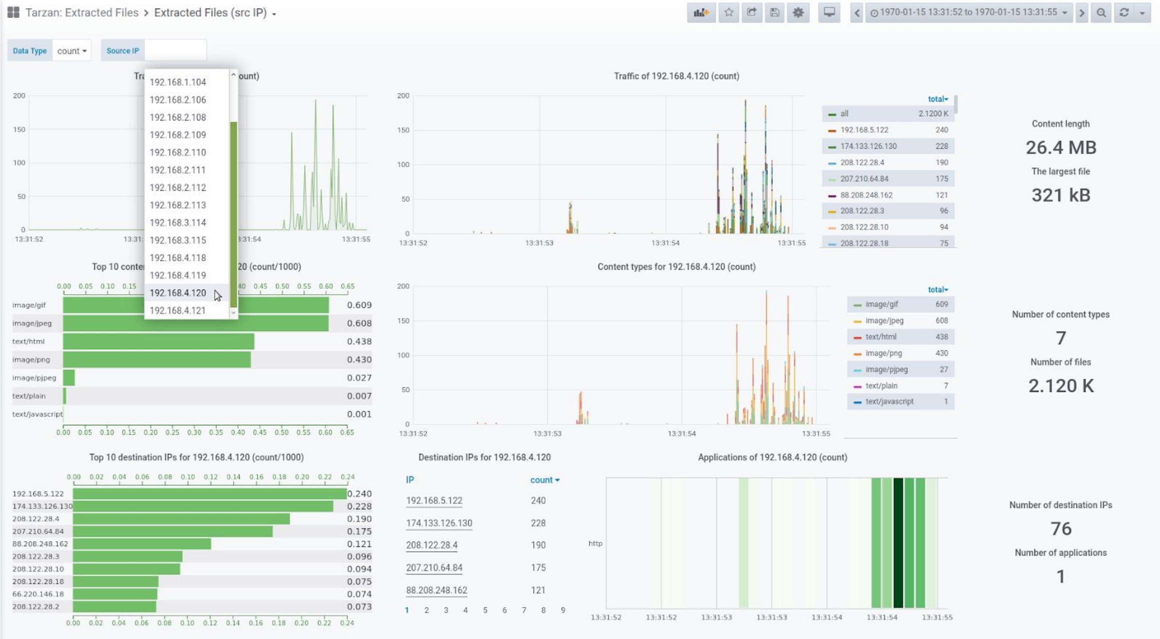
32.1 MB

The largest file

851 kB

Figure 4.12: An example of the dashboard providing information about the transferred file. Users can switch between the number of files or amount of data perspectives using Grafana's variable (the top left part).

Similar statistics is presented for specific IP addresses. The users can use the table of source or destination IP addresses to open dashboard which provides filtered data regarding the chosen IP address. An example of such dashboard is presented in Figure 4.13. The users can use the dashboard, for example, to perform a deeper analysis of specific clients, understand their behaviour and see what kind of data they download.



Tarzan: Extracted Files > Extracted Files (src IP)
1970-01-15 13:31:52 to 1970-01-15 13:31:55
Data Type
count
Source IP
192.168.1.104
192.168.2.106
192.168.2.108
192.168.2.109
192.168.2.110
192.168.2.111
192.168.2.112
192.168.2.113
192.168.3.114
192.168.3.115
192.168.4.118
192.168.4.119
192.168.4.120
192.168.4.121
Traffic of 192.168.4.120 (count)
Content types for 192.168.4.120 (count)
Top 10 content types for 192.168.4.120 (count/1000)
Top 10 destination IPs for 192.168.4.120 (count/1000)
Applications of 192.168.4.120 (count)
http
13:31:52
13:31:53
13:31:54
13:31:55

	total
all	2.1200 K
192.168.5.122	240
174.133.126.130	228
208.122.28.4	190
207.210.64.84	175
88.208.248.162	121
208.122.28.3	96
208.122.28.10	94
208.122.28.18	75

Content length

26.4 MB

The largest file

321 kB

	total
image/gif	609
image/jpeg	608
text/html	438
image/png	430
image/pjpeg	27
text/plain	7
text/javascript	1

image/gif	0.609
image/jpeg	0.608
text/html	0.438
image/png	0.430
image/pjpeg	0.027
text/plain	0.007
text/javascript	0.001

Number of content types

7

Number of files

2.120 K

Top 10 destination IPs for 192.168.4.120 (count/1000)

192.168.5.122	0.240
174.133.126.130	0.228
208.122.28.4	0.190
207.210.64.84	0.175
88.208.248.162	0.121
208.122.28.3	0.096
208.122.28.10	0.094
208.122.28.18	0.075
66.220.146.18	0.074
208.122.28.2	0.073

Destination IPs for 192.168.4.120

IP	count
192.168.5.122	240
174.133.126.130	228
208.122.28.4	190
207.210.64.84	175
88.208.248.162	121

1 2 3 4 5 6 7 8 9

Number of destination IPs

76

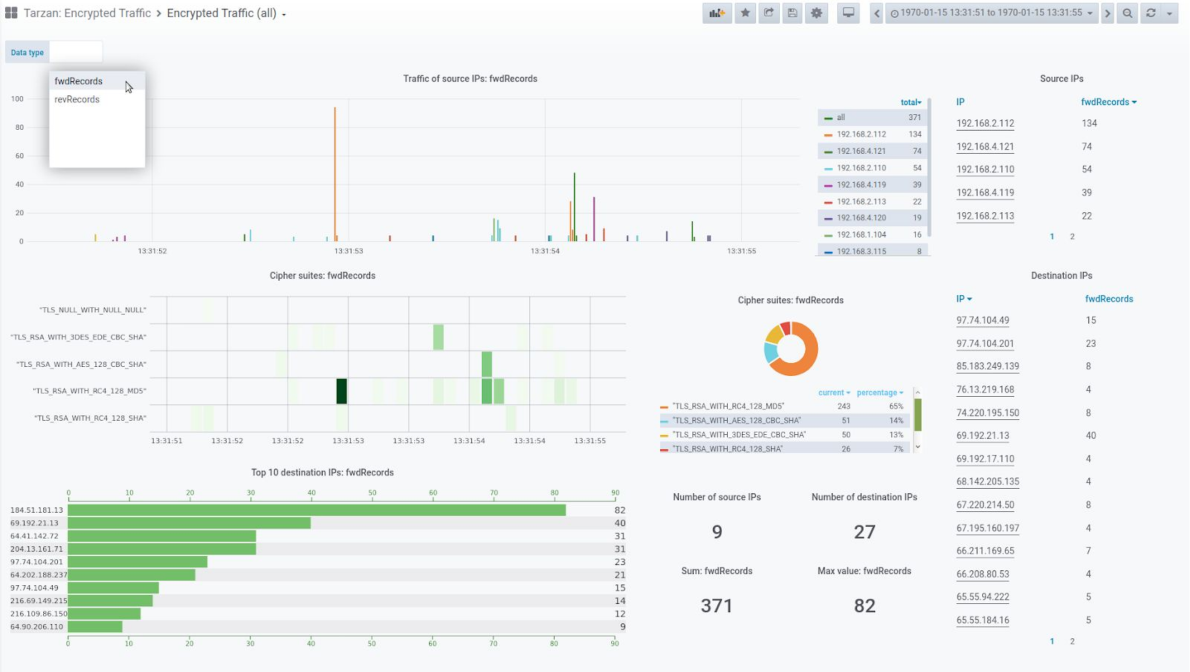
Number of applications

1

Figure 4.13: An example of the dashboard presenting data regarding extracted files for the specific source IP address (192.168.2.120).

4.4 Encrypted Traffic

The encrypted traffic data represented by TLS flows is another measurements of the use cases presented in Chapter 2. The dashboard design is very similar to design of Network Conversations dashboard (Section 4.2). It is composed of the stacked bar chart displaying the traffic in time grouped by particular source IP addresses. The users can analyze relations between source and destination IP addresses, overall statistics and be redirected to detailed perspectives displaying data of particular IP addresses. The users can choose between two data series: forward and reverse records. On the contrary to the basic traffic, the dashboard does not provide any information about used applications. Instead, it provides information about used cypher suites.


Tarzan: Encrypted Traffic › Encrypted Traffic (all)
1970-01-15 13:31:51 to 1970-01-15 13:31:55
Data type
fwdRecords
revRecords
Traffic of source IPs: fwdRecords
100
80
60
40
20
0
13:31:52
13:31:53
13:31:54
13:31:55
Cipher suites: fwdRecords
"TLS_NULL_WITH_NULL_NULL"
"TLS_RSA_WITH_3DES_EDE_CBC_SHA"
"TLS_RSA_WITH_AES_128_CBC_SHA"
"TLS_RSA_WITH_RC4_128_MD5"
"TLS_RSA_WITH_RC4_128_SHA"
13:31:51
13:31:52
13:31:52
13:31:53
13:31:53
13:31:54
13:31:54
13:31:55
Top 10 destination IPs: fwdRecords
0
10
20
30
40
50
60
70
80
90
184.51.181.13 82
69.192.21.13 40
64.41.142.72 31
204.13.161.71 31
97.74.104.201 23
64.202.188.237 21
97.74.104.49 15
216.69.149.215 14
216.109.86.150 12
64.90.206.110 9
Cipher suites: fwdRecords

	total
all	371
192.168.2.112	134
192.168.4.121	74
192.168.2.110	54
192.168.4.119	39
192.168.2.113	22
192.168.4.120	19
192.168.1.104	16
192.168.3.115	8

Source IPs

IP	fwdRecords
192.168.2.112	134
192.168.4.121	74
192.168.2.110	54
192.168.4.119	39
192.168.2.113	22

1 2

	current	percentage
"TLS_RSA_WITH_RC4_128_MD5"	243	65%
"TLS_RSA_WITH_AES_128_CBC_SHA"	51	14%
"TLS_RSA_WITH_3DES_EDE_CBC_SHA"	50	13%
"TLS_RSA_WITH_RC4_128_SHA"	26	7%

Destination IPs

IP	fwdRecords
97.74.104.49	15
97.74.104.201	23
85.183.249.139	8
76.13.219.168	4
74.220.195.150	8
69.192.21.13	40
69.192.17.110	4
68.142.205.135	4
67.220.214.50	8
67.195.160.197	4
66.211.169.65	7
66.208.80.53	4
65.55.94.222	5
65.55.184.16	5

1 2

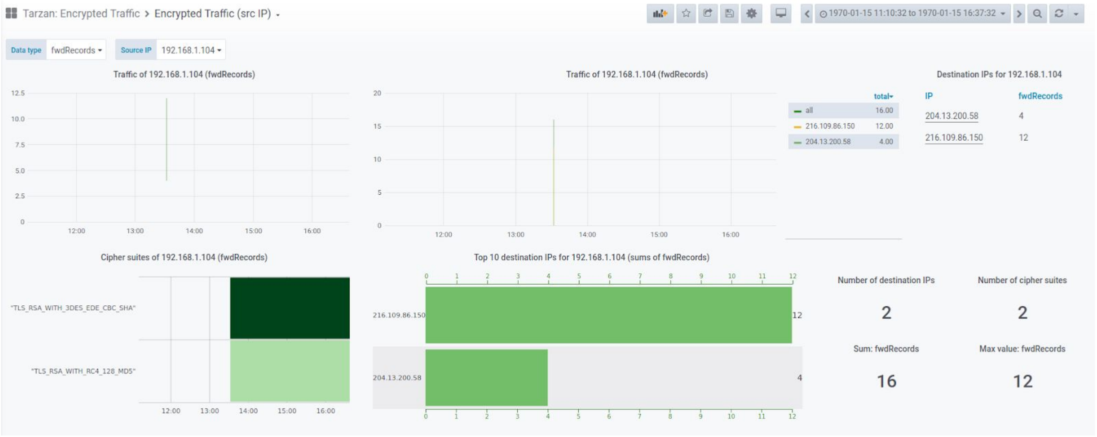
Number of source IPs
9

Number of destination IPs
27

Sum: fwdRecords
371

Max value: fwdRecords
82

Figure: 4.10: And example of the overall dashboard displaying data of encrypted traffic. Users can see the traffic in time and analyze relationships between IP addresses. They can use hyperlinks (the tables on the right) to open detailed data perspectives.



Tarzan: Encrypted Traffic > Encrypted Traffic (src IP)
1970-01-15 11:10:32 to 1970-01-15 16:37:32
Traffic of 192.168.1.104 (fwdRecords)
12.5
10.0
7.5
5.0
2.5
0
12:00
13:00
14:00
15:00
16:00
Traffic of 192.168.1.104 (fwdRecords)
20
15
10
5
0
Cipher suites of 192.168.1.104 (fwdRecords)
"TLS_RSA_WITH_3DES_EDE_CBC_SHA"
"TLS_RSA_WITH_RC4_128_MD5"
Top 10 destination IPs for 192.168.1.104 (sums of fwdRecords)
0
1
2
3
4
5
6
7
8
9
10
11
12
216.109.86.150
204.13.200.58

Data type: fwdRecords

Source IP: 192.168.1.104

	total
all	16.00
216.109.86.150	12.00
204.13.200.58	4.00

Destination IPs for 192.168.1.104

IP	fwdRecords
204.13.200.58	4
216.109.86.150	12

Number of destination IPs: 2

Number of cipher suites: 2

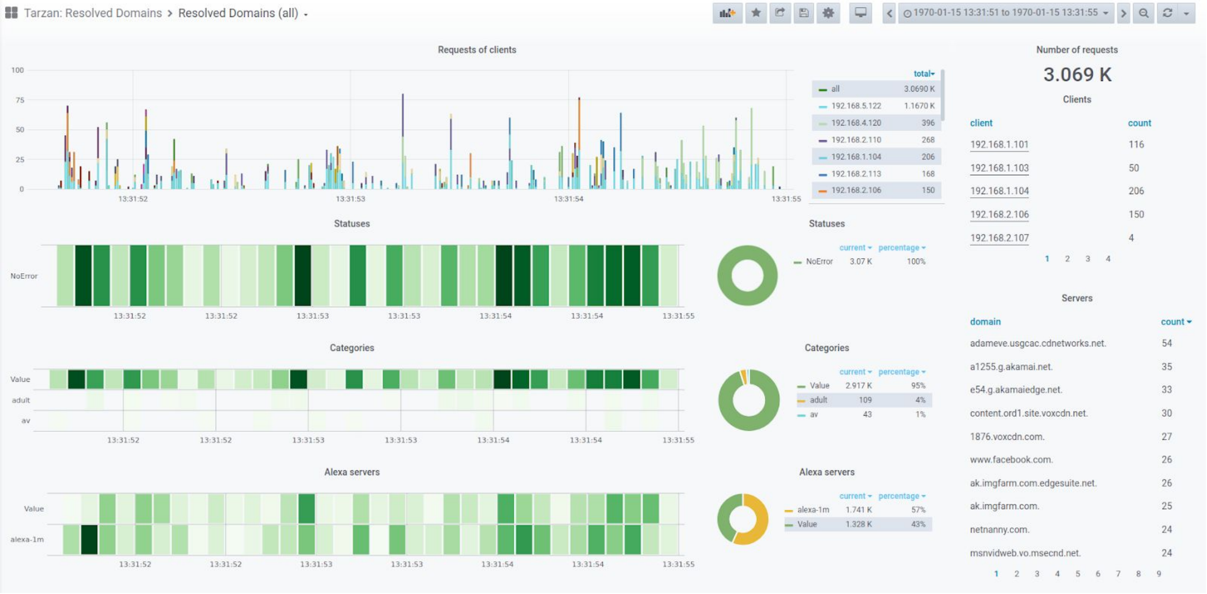
Sum: fwdRecords: 16

Max value: fwdRecords: 12

Figure 4.11: And example of the dashboard presenting data regarding encrypted traffic for the specific source IP address (192.168.1.104).

4.5 Resolved Domains

The last measurement represents the information about DNS requests, which is a good way to quickly understand behaviour of clients. The information is highly comprehensive since we work with the domain names rather than IP addresses. The dashboard contains the table of translated domain names which can be sorted by the numbers of DNS requests. Also the dashboard contains statistics about the status of requests, usage of whitelist/blacklist methods (category), or alexa top lists, which are provided in the form of heat maps and pie charts. The dashboard presents information about traffic and client IP addresses as well. An example of the dashboard is presented in Figure 4.14. Then, users can analyze behavior of particular client using the dashboard displaying filtered data for particular client IP address (Figure 4.15).



	total
all	3.0690 K
192.168.5.122	1.1670 K
192.168.4.120	396
192.168.2.110	268
192.168.1.104	206
192.168.2.113	168
192.168.2.106	150

Statuses

	current	percentage
NoError	3.07 K	100%

Categories

	current	percentage
Value	2.917 K	95%
adult	109	4%
av	43	1%

Alexa servers

	current	percentage
alexa-1m	1.741 K	57%
Value	1.328 K	43%

Number of requests

3.069 K

Clients

client	count
192.168.1.101	116
192.168.1.103	50
192.168.1.104	206
192.168.2.106	150
192.168.2.107	4

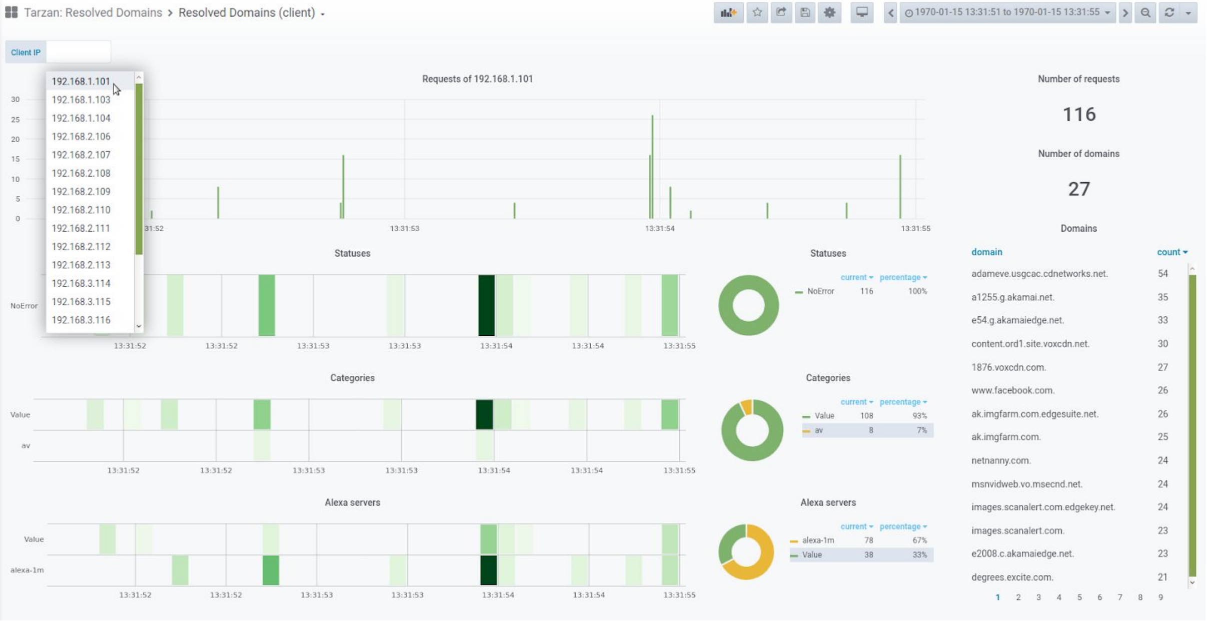
1 2 3 4

Servers

domain	count
adameve.usgcac.cdnetworks.net.	54
a1255.g.akamai.net.	35
e54.g.akamaiedge.net.	33
content.ord1.site.voxcdn.net.	30
1876.voxcdn.com.	27
www.facebook.com.	26
ak.imgfarm.com.edgesuite.net.	26
ak.imgfarm.com.	25
netnanny.com.	24
msnvidweb.vo.msecnd.net.	24

1 2 3 4 5 6 7 8 9

Figure 4.14: And example of the overall dashboard displaying data of DNS requests.



Tarzan: Resolved Domains > Resolved Domains (client)

1970-01-15 13:31:51 to 1970-01-15 13:31:55

Client IP

192.168.1.101
192.168.1.103
192.168.1.104
192.168.2.106
192.168.2.107
192.168.2.108
192.168.2.109
192.168.2.110
192.168.2.111
192.168.2.112
192.168.2.113
192.168.3.114
192.168.3.115
192.168.3.116

Requests of 192.168.1.101

Statuses

Categories

Alexa servers

Statuses

	current	percentage
NoError	116	100%

Categories

	current	percentage
Value	108	93%
av	8	7%

Alexa servers

	current	percentage
alexa-1m	78	67%
Value	38	33%

Number of requests

116

Number of domains

27

Domains

domain	count
adameve.usgcac.cdnetworks.net.	54
a1255.g.akamai.net.	35
e54.g.akamaiedge.net.	33
content.ord1.site.voxcdn.net.	30
1876.voxcdn.com.	27
www.facebook.com.	26
ak.imgfarm.com.edgesuite.net.	26
ak.imgfarm.com.	25
netnanny.com.	24
msnvidweb.vo.msecnd.net.	24
images.scanalert.com.edgekey.net.	24
images.scanalert.com.	23
e2008.c.akamaiedge.net.	23
degrees.excite.com.	21

Figure 4:15: An example of the dashboard presenting the data about DNS requests for the specific client’s IP address (192.168.1.101).

5 Summary

Network forensics data visualization provides one of the possible approaches for coping the big data problem in digital forensics. Visual analytics enables to provide different complementary views on the data that enables the analysts to quickly focus on the location and timeframes with possible incidents and suspicious activities.

In this report, we described our experiment in the is of network data visualization for the purpose of digital forensics. Based on the extracted data from the Traffix tool we designed several dashboards for visualization of different types of data. Because the source data are organized are events there is a natural timeline representation. In addition to temporal analysis, the visual representation also supports to apply aggregation operations and drilling to move the focus vertically at different levels of abstraction. The experimental environment was created consisting of the Traffix engine, InfluxDB, and Grafana. The Traffix engine is a distributed cluster-based tool for processing large amounts of capture files. InfluxDB is database with temporal support enabling to store millions of events. Finally, Grafana is an open-source visual analytics framework. Several dashboards were designed and realized in this environment to demonstrate the proposed approach.