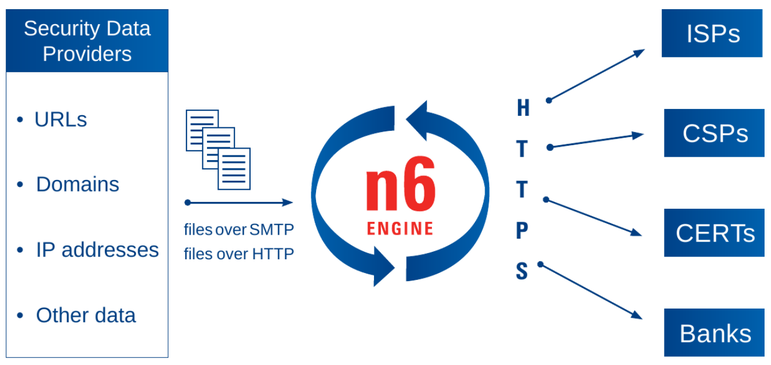
WORKPACKAGE 1: Threat Data & WORKPACKAGE 3: Cyberdefense

The n6 platform

n6 was designed and developed entirely at CERT Polska as a platform for acquisition, processing and exchange of information regarding Internet threats. Currently, millions of security events are processed daily in an automated manner. The goal is efficient, reliable and fast delivery of large volumes of network incident data to interested parties: network owners, administrators and Internet Service Providers. The project disseminates information gathered from various security systems operated by security organizations, software vendors, independent researchers, etc. The platform exchanges information on the source attacks in the form of URL addresses, domains, IP addresses or names of malicious software, and also information on special data (ZeuS config etc.).



The n6 platform

Data sources

The n6 platform handles many different types of data feeds, including malicious URL addresses, infected hosts (bots), C&C servers, phishing, spam, scanning, DDoS, brute force attacks, open-proxies and open-resolvers. Many different sources are providing data for the platform, including some integrated as part of NECOMA.

n6 REST API & n6 SDK

One of the main design goals of the n6 API is to make integration as easy as possible from the client point of view. This property, combined with extendibility provided by JSON, ability to handle large amounts of data and filtering capability, makes the API a good candidate for exchange of heterogeneous datasets and it was chosen as the basis of the common data exchange mechanism in the NECOMA project. In order to accomplish that goal, the n6 API has been extended with additional attributes required to represent information present in all datasets used by the consortium.

To make it as easy as possible to make other datasets accessible using the same API without requiring direct integration with the n6 platform operated by NASK, the n6 SDK was developed and released as open source. A detailed tutorial on enabling n6 API in a dataset is provided in deliverable D3.2 of NECOMA.

n6 stream interface

The REST interface described above is suitable for delivery of any amount of data and provides filtering capabilities. Nevertheless, its architecture is inherently pull-based, meaning that the client (recipient of data) must explicitly request new data from the server, which limits its timeliness and scalability. This fact led to the design and implementation of n6 stream interface, an alternative API for n6 that has an asynchronous push-based architecture. At this point, the API does not provide a way for the client to send data, although such addition without breaking backwards compatibility is feasible in the selected approach. The new stream interface is built on top of STOMP, an open, text-based, message-passing protocol.

By design, the new interface is not meant to replace the existing REST interface, but rather to complement it. The stream interface solves a different use case: getting new data as it is added to the n6 platform. It does not provide a way to serve old data, already present in the system before the client subscribed to the feed, the REST API should be used for that purpose.

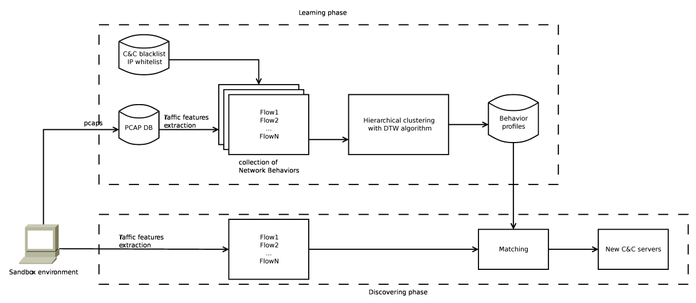
WORKPACKAGE 2: Threat analysis

SVM and fp-growth

This approach has already been presented last year; in the meantime it was developed further, obtaining new results and integrating other data sources. Results of new research are currently being published.

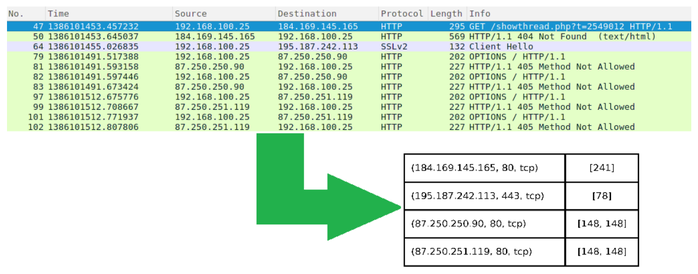
C&C identification in sandbox results

We propose a new mechanism to identify previously unknown C&C servers based on behavior of malware samples in a sandbox environment. We assume that similar malware samples not only use similar C&C protocols but also show similarities in other network activities due to shared code. Our system processes traffic traces collected during execution of malware samples, building clusters of samples showing the same sequence of connections both to legitimate and C&C servers.



Overview of the C&C identification module

In the **learning phase**, for each sample a network behavior profile is created in the form of a sequence of flows identified by protocol, destination address and destination port and associated with a sequence of packet payload lengths. Then, using C&C blacklist and legitimate servers whitelist, each flow is classified as either "benign" or "C&C" based on destination IP address.



No.	Time	Source	Destination	Protocol	Length	Info
47	1386101453.457237	192.168.100.25	184.169.145.165	HTTP	295	GET /showthread.php?t=2549012 HTTP/1.1
50	1386101453.645037	184.169.145.165	192.168.100.25	HTTP	569	HTTP/1.1 404 Not Found (text/html)
64	1386101455.026835	192.168.100.25	195.187.242.113	SSLv2	132	Client Hello
79	1386101491.517388	192.168.100.25	87.250.250.90	HTTP	202	OPTIONS / HTTP/1.1
81	1386101491.593158	87.250.250.90	192.168.100.25	HTTP	227	HTTP/1.1 405 Method Not Allowed
82	1386101491.597446	192.168.100.25	87.250.250.90	HTTP	202	OPTIONS / HTTP/1.1
83	1386101491.673424	87.250.250.90	192.168.100.25	HTTP	227	HTTP/1.1 405 Method Not Allowed
97	1386101512.675776	192.168.100.25	87.250.251.119	HTTP	202	OPTIONS / HTTP/1.1
99	1386101512.708667	87.250.251.119	192.168.100.25	HTTP	227	HTTP/1.1 405 Method Not Allowed
101	1386101512.771937	192.168.100.25	87.250.251.119	HTTP	202	OPTIONS / HTTP/1.1
102	1386101512.807806	87.250.251.119	192.168.100.25	HTTP	227	HTTP/1.1 405 Method Not Allowed

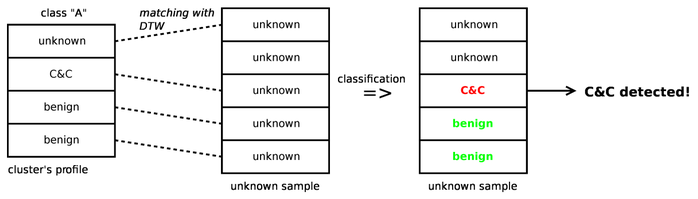
(184.169.145.165, 80, tcp)	[241]
(195.187.242.113, 443, tcp)	[78]
(87.250.250.90, 80, tcp)	[148, 148]
(87.250.251.119, 80, tcp)	[148, 148]

Creation of a sample's network behavior description from traffic trace

In order to derive network behavior profiles, we cluster the initially classified samples with similar network behavior using agglomerative hierarchical clustering with Dynamic Time Warping (DTW) as the algorithm to compute the distance function, enabling matching of sequences with different lengths. For each cluster a network profile is built, which is essentially a sequence of flows similar to mentioned above but extended with weight for each payload length of flow.

Matching network behavior descriptions using DTW

In the **discovering phase**, the distance between a sample containing flows directed to unknown servers (potentially C&C) and each cluster represented by its network behavior profile is computed using again DTW algorithm. This time, however, weights for each packet payload length within a profile's flows are taken into account. Finally, each unknown endpoint in the sample being tested is classified using flows from profile (cluster) closest to sample, matched with DTW algorithm.



Matching a sample to a cluster (C&C discovery phase)

The algorithm was tested on data from the n6 database. The following table shows the results for 248 IP addresses classified using knowledge gathered from a dataset containing 433 known C&C IP addresses and 740 benign addresses (resulting in 414 clusters, 50 best ones were used).

Classification	Verified as	Count	[%]
C&C	true positive	66	26.61
C&C	false positive	7	2.82
C&C	unknown	0.745	58.47
benign	true negative	30	12.10

Graph-based cross-layer analysis

The graph-based analysis is a simple tool enabling exploration of the entire collection of datasets based on seed tokens: interesting values either identified by other analyses, or obtained in a different way by the user of the system. The goal of the approach is to find as much information related to a given token as possible, while limiting the cost of such operation. Related information within different datasets can be modeled as a graph, with information from a single dataset represented as a node and shared tokens represented as edges.

The algorithm proposed in deliverable D2.1 builds a graph of related information starting with a single token. Our solution creates a graph of tokens and assigns a cost to each query, building the graph by selecting cheapest queries until a preset budget is consumed.

The proposed analysis has two separate use cases. In interactive mode it is a tool used to explore the datasets, starting with a manually chosen token. In automatic mode the analysis processes tokens provided by other analyses or taken from some interesting datasets.

Automated rating and classification

NASK has developed several metrics of dataset usefulness, both manual (configurable or reputation-based) and automatically identifiable. The ratings are heavily used in the graph-based analysis to select the most useful data sources in the process of graph creation.

Another type of rating mechanism is proposed for the process of collecting external knowledge. In this case free APIs of search engines are used to assess the quality of search results provided for the tokens identified by the NECOMA platform. These results, coupled with the cost model of each engine, allow us to select the most cost-efficient choice of paid service capable of handling the large amount of queries generated for the tokens identified by the NECOMA platform with a good chance of collecting valuable information.

WORKPACKAGE 4: Case Studies

Malware campaign mitigation

NASK will demonstrate how the solutions developed in previous workpackages deal with a simulated new malware campaign, verifying each of the necessary steps: detection of individual incidents (attack attempts against servers, webpages turning malicious, etc.), correlation of the collected data showing a repeating pattern, identification of a global campaign, data enrichment, collection of external information and response to the threat.

The demonstration will be based on a snapshot of real threat datasets, showing all stages of analysis leading to the identification of malicious campaing and enabling proper reaction. We are currently finalizing the development of the metrics measuring the effectiveness of our approach.

About NASK

In 1991, NASK connected Poland to the Internet. Since December 1993, NASK has been a research & development organization and a leading Polish data networks operator. We offer state-of-the-art telecommunications and data solutions to business, administration and academic customers. NASK is also the Polish national registry of Internet names in the .pl domain.

Scientific Activity As a research institute, NASK carries out numerous scientific and research & development activities. Projects centre on telecommunications & data quality (QoS – Quality of Service), security of IT systems and biometric identification methods. NASK is an active member of many international organizations and associations (FIRST, CENTR, TERENA, RIPE) and participates in national and European Union projects.

NASK's participation in the NECOMA project is a joint activity of two groups within NASK (joined during the project by the Software Development Department):

CERT Polska A part of the NASK organization, CERT Polska is a Computer Emergency Response Team operated by NASK that handles incidents related to the .pl namespace. A part of the team's work is focused on researching new detection and analysis methods and developing tools to aid this process.

Network and Information Security Methods Team A part of the NASK Research Division dealing with security problems, the NISM team cooperates often with CERT Polska in security-related research projects. The team's more theoretical and exploratory approach complements the operational experience and focus of CERT Polska.