SEVENTH FRAMEWORK PROGRAMME

Trustworthy ICT

Project Title:

Enhanced Network Security for Seamless Service Provisioning in the Smart Mobile Ecosystem

Grant Agreement No: 317888, Specific Targeted Research Project (STREP)

DELIVERABLE

D3.1: Network information sources

Deliverable No.		**D3.1**	
Workpackage No.	**WP3**	Workpackage Title	**Network data collection infrastructure**
Task No.	**T3.1**	Task Title	**T3.1: Survey of information sources**
Lead Beneficiary		**CERTH/ITI**	
Dissemination Level		**PU**	
Nature of Deliverable		**R**	
Delivery Date		**31 July 2013**	
Status		**D**	
File Name		**NEMESYS_Deliverable_D3.1.doc**	
Project Start Date		**01 November 2012**	
Project Duration		**36 Months**	

Authors List

Author’s Name	Partner	E-mail Address
Leading Author / Editor		
D. Tzovaras	CERTH/ITI	Dimitrios.Tzovaras@iti.gr
Co-Authors		
V. Mavroudis	CERTH/ITI	Vasilios.Mavroudis@iti.gr
G. Papadopoulos	CERTH/ITI	gpapadopoulos@iti.gr
G. Bekoulis	CERTH/ITI	bekou@iti.gr
M. Baltatu	TI	madalina.baltatu@it.telecomitalia.it
L. Delosičres	HIS	ldelosieres@hispasec.com

Contents

List of Figures

List of Tables

Abbreviations

AMGP	Android Malware Genome Project
ANUBIS	Analyzing Unknown Binaries
API	Application programming interface
APK	Android Application Package
ASN	Autonomous System Number
AV	Antivirus
BL	Blacklist/Blocklist
CDR	Call Detail Record
CSV	Comma-Separated values
CVSS	Common Vulnerability Scoring System
DB	Database
DNS	Domain Name System
DoW	Description of Work
EPS	Encapsulated PostScript
GB	Gigabyte
GPS	Global Positioning System
HTTP	Hypertext Transfer Protocol
ID	Identity Document
IIS	Internet Information Services
IP	Internet Protocol
IRC	Internet Relay Chat
ISC	Internet Systems Consortium
JSON	JavaScript Object Notation
LAN	Local Area Network
MAC	Media Access Control
NIST	National Institute of Standards and Technology
NVD	National Vulnerability Database
OMC	Operating and Maintenance Centers
OS	Operating System
OSVDB	Open Source Vulnerability Database
PC	Personal Computer

PTR	Pointer Record
REGEX	Regular Expression
RFC	Request for Comments
RSS	Rich Site Summary
SCAP	Security Content Automation Protocol
SMS	Short Message Service
SQL	Structured Query Language
SSH	Secure Shell
TCP	Transmission Control Protocol
TSV	Tab-separated Values
URL	Uniform Resource Locator
WLAN	Wireless LAN
XML	Extensible Markup Language

Abstract

This document provides an outline of all available information sources followed by evaluation and comparisons under the light of the NEMESYS objectives. The purpose of deliverable D3.1 is to provide the consortium with an overview of the available information sources related to mobile signaling, user behavior and attacks against end hosts and edge networks and identify those that are the most likely to give us interesting hints with respect to the correlation of attacks. The task T3.1 contributed to the production of this deliverable.

The main body of this report examines in detail the information sources, divided in four discrete categories. The first category introduces sources which collect data using monitoring software installed directly in mobile devices or monitor the signaling traces of devices in a confined geospatial area. The second category describes datasets composed using monitoring systems deployed in the Operating and Maintenance Centers (OMCs) of the network providers. The third category focuses on malware and vulnerability databases and repositories. Finally, the forth category includes DNS monitoring services, honeypot projects, malicious URL blocklists and IP and URL reputation services. An “Evaluation & Comparison” section is included for each of the above categories including comparison tables coupled with our remarks and suggestions. The final part of the document is dedicated to the conclusions drawn from the previous examination and puts the deliverable into the higher context of the project.

It is strongly emphasized that this is an ongoing document that is being evolved along with the project progress and will be regularly updated to reflect up-to-date information.

1 Introduction

1.1 Purpose of the Deliverable

The purpose of this deliverable is to provide the consortium with all relevant information sources relating to attacks targeting both the end hosts and edge networks and identify the ones that are most likely to give us interesting hints with respect to the correlation of core network related attacks with attacks against mobile devices.

This deliverable presents the results of Task T3.1 “Survey of Information Sources” and includes different types of information sources that are in accordance with the objectives of NEMESYS as it has been specified in “Annex I- Description of Work” (DoW) of the project. The information sources that are reported either pertain wireline and wireless networks infrastructure attacks observed in the Internet or relate to attacking processes that target mobile devices. The sources of information that this document reports will be later manipulated and correlated using the data collection infrastructure developed in the T3.3.

1.2 Organisation of the Deliverable

This deliverable is organised as follows:

Chapter 2 presents an outline of all information sources that compose their datasets using data collected directly from the user devices using either monitoring software installed in the device or device tracking techniques.

Chapter 3 introduces the datasets collected from a central node in the mobile network infrastructure and not directly from the devices.

Chapter 4 shows the databases and the repositories that include information and details on mobile malware and vulnerabilities found in mobile software and operating systems. Additionally, some malware analysis tools are discussed.

Chapter 5 describes the information sources that are maintained using mainly internet monitoring techniques. The information sources vary from DNS monitoring services to honeypots and IP address & domain blocklists.

Finally, chapter 6 concludes the deliverable and enumerates the outcomes of the information sources research. Furthermore, it underlines important findings and sources that are most likely to give interesting hints.

2 Mobile devices direct activity monitoring

2.1 Introduction

In this first chapter, we introduce different information sources that release datasets beneficial for NEMESYS. Those datasets will be the input of the anomaly detection mechanisms that are going to be developed in NEMESYS. Additionally, within the NEMESYS project, we aim to analyze and correlate available signaling and user behavior data. The data sets of the chapter are divided in data sets which contain user data collected using monitoring software installed in the mobile device and data sets containing Bluetooth and WiFi traces. We expect the information sources in this category to be very useful in tasks T4.1 and T4.2 as well as in visualization tasks from workpackage 5.

2.2 Monitoring Software

This section introduces information sources that distribute data sets composed using activity monitoring software installed in the mobile device of the end user. Each source uses a different setup and the collected data vary. The combination of the collected data provides a base for developing user profiles to be processed by anomaly detection algorithms.

2.2.1 NODOBO

NODOBO [1] is a software suite, developed at the University of Strathclyde, which allows precise capture and replay of smartphone user interactions sessions.

- **Content**

  The Nodobo-2011-01-v1 is a dataset generated using the NODOBO software suite. It contains data gathered during a study of the mobile phone usage of 27 high-school students, from September 2010 to February 2011. This dataset includes 13035 call records, 83542 message records, 5292103 presence records, and other related data. NODOBO monitors the user's calls and messages, the connected celltowers, device details, nearby presences, the users and any Wifi activity.

- **Format**

  The dataset is available in CSV format.

- **Tools**

  Three sample programs written in ruby programming language are also provided.

- **Access**

  The dataset is public. In case the dataset is used for research, it is required that the most relevant NODOBO publication is cited.

2.2.2 Lausanne Data Collection – NOKIA Mobile Data Challenge

The *Lausanne data collection* campaign [2] was a people centric sensing campaign that was operated in Lausanne, Switzerland, and ran by a corporate research laboratory from the mobile industry. During the campaign, Nokia N95 phones were allocated to a heterogeneous sample of nearly 170 participants from Lausanne to be used over a period of one year.

- **Content**

  The data collection software ran on the background of the phones in a non-intrusive manner, yielding data on modalities such as social interaction and spatial behavior. The dataset aims to illuminate multiple aspects of human behavior.

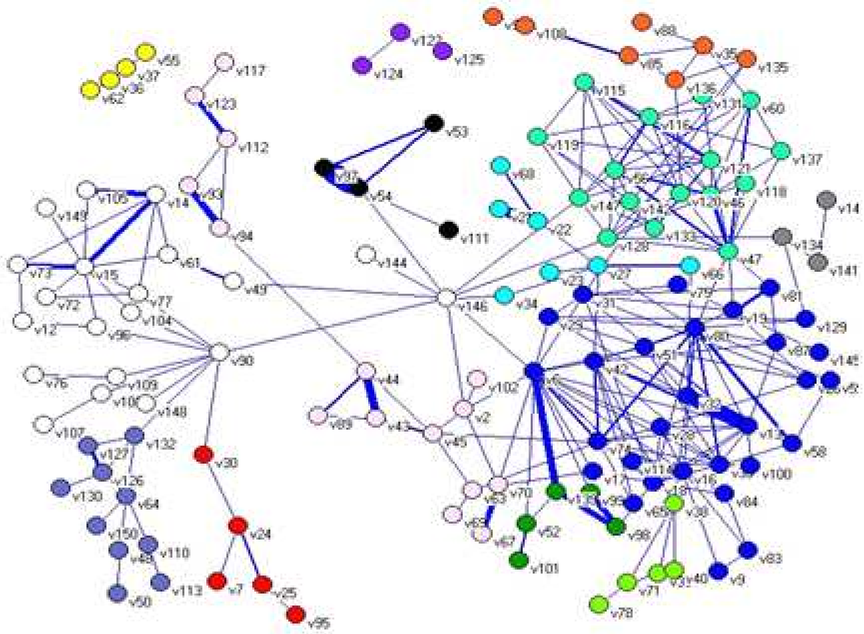


Figure 1: The graph was generated using a technique called modularity optimization. As can be seen, the sample consists of several smaller communities, several of them connected to each other by bridging individuals.

It contains data that feature GPS points, user calls, SMS, pictures taken and videos shots. More specifically, 132.000 calls, 88.225 SMS, 28.000 pictures and 2.100 videos shots were collected. Furthermore, the Bluetooth and WiFi activity is monitored with bluetooth and WLAN scans. Besides, additional information on user behavior, activity accelerometer samples and audio samples are available, though they have been edited properly to ensure the users' privacy. Last but not least, application, calendar and phone book entries were also recorded.

- **Format**

  The format was not known at the time the document was composed.

- **Tools**

  Few tools were available only to the campaign participants but no tools are provided for analysis of the datasets.

- **Access**

  The access to the database is restricted and permission for the NEMESYS project is pending.

2.2.3 Reality Commons - Reality Mining Dataset

The *Reality Mining experiment* [3] studied community dynamics by tracking a sufficient amount people with their personal mobile phones and resulted in the first mobile data set with rich personal behavior and interpersonal interactions.

- **Content**

  The produced dataset includes traces from 100 human subjects (75 students or faculty in the MIT Media Laboratory, and 25 students at the MIT Sloan business school) over the course of nine months and approximately 500,000 hours of data on users' location, communication and device usage behavior. Of the 75 Media Lab participants, 20 were incoming masters students and 5 were incoming MIT freshmen, and the rest had remained in the Media Lab for at least a year. The mobile devices used were Nokia 6600 smart phones pre-installed with monitoring software and the *context* application [4] from the University of Helsinki. The information collected includes: Call logs, Bluetooth devices in proximity, Cell tower IDs, Application usage and the Phone status. The sensor data provide information on:

- Proximity
- Location, location labels, latitudes and longitudes
- Call log, SMS: time with hourly resolution (or date + early Morning/morning/afternoon/evening/midnight), duration, unique callee identifier (natural number)
- Running Nokia applications

Moreover, the survey data provide information on perceived friendships, personal attributes of the user, the research group, the user's academic position (e.g., graduate student, undergraduate student, staff, professor), living area and lifestyle details (e.g., when in the office, how often to travel, predictability of life, where to hangout, how often get sick).

- **Format**

  Dataset is provided in MAT (matlab) format.
- **Tools**

  No specific tools are provided.
- **Access**

  The dataset is released to the public and citation is requested.

2.2.4 SMS Corpus

The *SMS corpus* [5] is composed by researchers from the School of Computing of the National University of Singapore, who aim to expand an existing dataset of SMS messages, to study its language, network and other characteristics.

- **Content**

  As of June 2013, 42140 English SMS messages and 31205 Chinese SMS messages have been collected. Additionally, statistics are updated every week, while detailed, individual monthly statistics are also provided. It is important to note that English messages are largely from Singaporean university students. For each SMS a message ID is generated and a new entry is created containing the:

  - Message Text
  - Source (srcID, country)
  - Destination (destID, country)

    - Phone Model (manufacturer, modelNumber)
    - User Profile (userID, Experience Level, InputMethod)
    - Message Profile (date, language, forwarded, mass, reply)
- **Format**

  The corpus is available in XML and SQL formats.
- **Tools**

  No specific tools provided.
- **Access**

  The dataset is public for non-commercial use. A free registration is required.

2.2.5 Michal Ficek Dataset

The *Michal Ficek dataset* [6] contains mobile phone records (Call Data Records) and cell transitions of Czech Ph.D. student Michal Ficek, stored by his own mobile terminal in 2010-2011.

- **Content**

  The dataset covers more than 99.99% of 142 days of mobile phone usage in mobile networks of 8 different providers in 5 countries: Czech Republic, Slovak Republic, Germany, Austria and the USA. The source of the data is Michal Ficek's mobile phone (Nokia E52). The coordinates of positions within the cells were obtained by translating the Cell-IDs to their geographical coordinates by querying the Google Location API. For each activity event a timestamped record is created defining its type, direction (Incoming/Outgoing) and duration. The "LogExport" application was used to record time and type of communication events (voice, SMS, data). For cell-transition recording, the free CellTrack91 application was utilized.
- **Format**

  Datasets are available in CSV format.
- **Tools**

  No tools are provided.
- **Access**

  Access to the database is public and a free registration is required.

2.3 Bluetooth Traces

The information sources enumerated below provide datasets that contain Bluetooth interactions between populations of mobile devices. The analysis of the Bluetooth traces using the algorithms developed within the NEMESYS project will give a detailed overview of the user mobility and usage patterns and based on them malicious activity and anomalies will be detected and classified in further steps.

2.3.1 Nottingham Mall Research

The *Nottingham Mall Research* dataset [7] is composed of real-world Bluetooth contact data collected from shop employees of a shopping mall.

- **Content**

  The contact data from shop employees of the mall were collected over six days. The devices used to collect data in this experiment are smart phones running symbianOS and using Bluetooth technology. The mobile devices were strictly used within the shopping mall. For six days the devices were all collecting data from around 9:15 a.m. to around 8:45 p.m..

- **Format**

  The dataset is provided in EPS format.

- **Tools**

  The shopping mall layout is provided for reference and correlation based on location.

- **Access**

  Access to the database is public.

2.3.2 Android Bluetooth Tracing Experiment

This is a dataset originating from the *Android Bluetooth tracing experiment* [8]. The experiment was performed for a period of 35 days in an academic environment (University Politehnica of Bucharest) between November 18 and December 22 2011.

- **Content**

  The data was collected only inside the grounds of the faculty between 8 AM and 8 PM during week-days. There were a total of 22 participants, chosen to be as varied as possible in terms of year, in order to obtain a better approximation of mobility in a real academic environment. The participating members were asked to start the application whenever they arrived at the faculty and to turn it off when they left, because we were only interested in the mobility patterns and social interaction in the academic environment. Data was collected using a Social-Tracer android application that registers contacts between mobile devices with Bluetooth.

- **Format**

  There are two files in this trace. The first one (interactions.dat) contains all interactions between devices participating in the experiment. Each entry contains the following features:

  - Observer ID
  - Observed ID
  - Encounter start
  - Encounter end
  - Number of encounters between two devices
  - Duration until the next encounter between these two devices

  The IDs range from 1 to 22 for internal devices, and higher than 22 for the external ones. The encounter start and end times are in seconds since January 1 1970, while the duration until the next encounter between the same two devices is in milliseconds. The second file (social.dat) contains the social network between the participating nodes. It is a comma-separated matrix with each line corresponding to a device. Each column represents a node in the experiment and it can be either 1 if the nodes have a social relationship and 0 otherwise.

- **Tools**

  No specific tools were provided.

- **Access**

  Access to the database is public and a registration is required.

2.3.3 SIGCOMM 2009

During the SIGCOMM 2009 conference in Barcelona, Spain 76 persons used the opportunistic mobile social application (MobiClique) and the collected data were combined to form a single dataset.

- **Contents**

  The *SIGCOMM 2009* dataset [9] contains traces of Bluetooth device proximity, opportunistic message creation and dissemination, and the social profiles of the participants. During the experiment, each device performed a periodic Bluetooth device discovery to find out about nearby devices. Upon discovering new contacts, the device formed a RFCOMM link on a preconfigured channel for data communications. Each device recorded the results of the periodic device discovery and all data communications (RFCOMM link setup and bytes send/received). In addition, the devices recorded details of the user's social profile and its evolution and application level messaging. All collected traces are timestamped based on the device clock and reported as a relative time in seconds since the start of the experiment, 17/08/2009 08:00.

- **Format**

  The dataset is in CSV format.

- **Tools**

  No tools are provided for analysis or parsing.

- **Access**

  The access is free to the public upon registration.

2.3.4 Toronto Bluetooth Worms Investigation

Researchers from the University of Toronto based on several reports for Bluetooth malware composed a dataset containing Bluetooth activity traces. The *Bluetooth Worms Investigation* dataset [10] serves as a base for further analysis in order to examine the potential outbreak of mobile malware using the Bluetooth interface of mobile devices.

- **Content**

Three different traces of Bluetooth activity are available depending on the geographical location of the captures. The duration of captures includes data gathered in a single day. The trace records include:

1. 32-bit timestamp: the encounter start time.

2. Same timestamp as per #1, but in a human readable format
3. 32-bit timestamp: the encounter end time
4. Same timestamp as per #3, but in a human readable format
5. Location
6. Scanner ID
7. Anonymized MAC address of foreign Bluetooth device encountered.
8. Type of Bluetooth device
9. Manufacturer of Bluetooth device

- **Format**
  The dataset is in TSV format.
- **Tools**
  No tools are provided for analysis or parsing.
- **Access**
  The dataset is publicly released.

2.4 Evaluation & Comparison

In order to efficiently evaluate and compare the information sources outlined above, we designed Table 1. This table highlights the features and differences between the information sources and it is advisable to consult it when choosing datasets to be included to the data collection infrastructure (T3.3).

The size of the samples in this category is relatively small compared to datasets collected using other monitoring techniques. However, in most cases a large amount of data is collected from each participant and thus, the user behavior analysis and profiling become more accurate.

Upon examination, the Reality Mining and the NODOBO datasets are very useful for NEMESYS as they provide detailed information on the user activity and behavior. More specifically, the Reality Mining dataset contains many hours of user activity recordings and has a quite large number of participants compared to other sources. However, it is quite old (2005) and this may affect the accuracy of the reported data when it comes to user behavior patterns and activities. On the other hand, the NODOBO dataset is more recent and has similar features with Reality Mining (e.g., calls, location) but the size of the sample is much smaller (27). Furthermore, the Lausanne Data Collection appears to be one of the most suitable datasets for NEMESYS, since it provides excellent details on the user activity (e.g., SMS traces), has a large number of participants and took place in a recent year (2011). It is worth noting that the “Lausanne Data Collection” in comparison to the Reality Mining data, is expected to provide richer information to study mobile traffic and user behavior. This is because, in the Lausanne Data Collection the mobile phones of the participants were more powerful and equipped with more sensors. The SMS Corpus can be very useful when creating synthetic SMS traffic using simulation software or by merging datasets.

From the second category, which includes information sources with Bluetooth and WiFi activity traces, the “SIGCOMM 2009” and the “Bluetooth worms Investigation” datasets, appear to be more useful for our objectives. The first one provides traces from the activity of a medium number of participants in a confined space and also records their social interactions. The role of mobile devices in social interactions is necessary to be analyzed and understood (e.g., visualization) as attackers often base their malicious attempts in social properties. Moreover, the “Bluetooth worms experiment” dataset can be useful in simulating the user behavior and possible infections from nearby devices. The “Michal Ficek” is a dataset worth noticing, since it includes data from only one user for a number of months and can be used to identify normal changes in the user behavior over time (e.g., public holidays, mobile usage evolution).

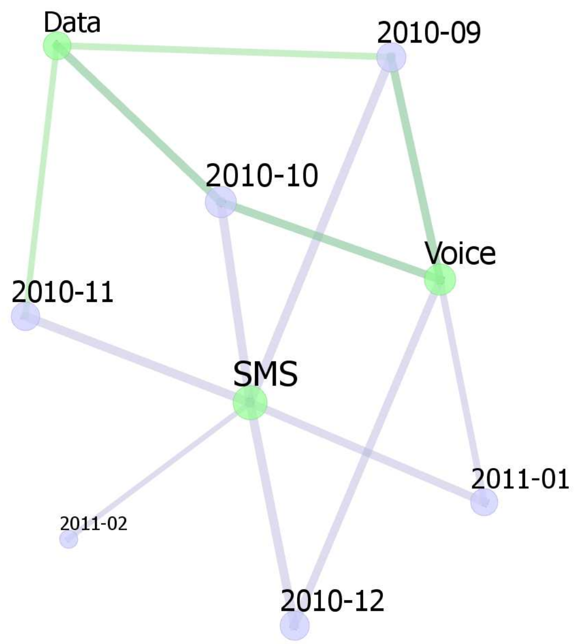


Figure 2: Michael Ficek dataset visualization that indicates the change in mobile service usage habits between different months.

It is important to note that all datasets contain timestamped records but only the SMS Corpus is an ongoing effort with regular updates.

Table 1: Mobile devices direct monitoring information sources

ID / Name	Timestamp	Calls	Messages	WiFi/Bluetooth	Tower	Device	Updates	Size	R/S	Year	Access
NODOBO	✓	✓	✓	✓	✓	✓	-	27	R	2011	Public
Reality Commons - Reality Mining Dataset	✓	✓	✓	✓	✓	✓	-	94	R	2005	Public
SMS Corpus	✓	-	✓	-	-	✓	✓	>116	R	2013	Public
Michal Ficek Dataset	✓	✓	✓	-	✓	✓	-	1	R	2011	Public
Nottingham Mall Research	✓	-	-	✓	-	✓	-	n/a	R	2007	Public
Android Bluetooth Tracing Experiment	✓	-	-	✓	-	-	-	22	R	2011	Public
SIGCOMM 2009	✓	-	✓	✓	-	-	-	76	R	2009	Public
Bluetooth Worms Investigation	✓	-	-	✓	-	✓	-	267	R	2006	Public
Lausanne Data Collection – NOKIA	✓	✓	✓	✓	✓	✓	-	170	R	2011	Pending

– ID/Name: Name of the information source or dataset.
– Timestamp: If the data records are time stamped.
– Calls: If the released datasets include information on the user calls.
– Messages: If the released datasets include information on the user's messaging activity.
– WiFi/Bluetooth: If the released datasets include information on the user's WiFi or Bluetooth activity.
– Tower: If the released datasets include information on the used cell tower.
– Device: If the released datasets include information on the user's device.
– Updates: If the datasets are periodically updated.
– Size: Sample size is refers to the number of user/devices monitored.
– R/S: If the information source compiles the datasets based on real or synthetic data.
– Year: Year of release or last update.
– Access: Access policy as defined by the source and the NEMESYS project's permissions.

3 Operating and Maintenance Centers

3.1 Introduction

In this chapter, we outline information sources whose datasets involve mobile services signaling, mobile services billing and call detail records. These datasets differ from the ones presented in the previous chapter over the data collection methodologies. More specifically, in contrast to the information sources of the previous chapter the data collection is made in the centralized infrastructure of the Operating and Maintenance Centers operated by the mobile services carriers. These data will assist the NEMESYS consortium in developing and testing anomaly detection algorithms (i.e., T4.1, T4.2) and visualization techniques (i.e., T5.2, T5.3) which are efficient and practical to be used by carriers to protect their infrastructure and their clients. Finally, it is important to note that all personal information included in the data sets is removed or anonymized.

3.2 Signaling

In mobile networks, the signaling method is responsible for addressing, call information and supervisory functions. Additionally, it also determines the status of the network and controls the amount of traffic. In this section we identify information sources which provide data sets that contain signaling traces of mobile protocols. These data sets are of great importance for NEMESYS, since they are essential in malware detection and identification. In this scope, malfunctioning mobile devices will be traced by using the signaling data from control planes and applying anomaly detection algorithms developed for mobile networks and devices.

3.3 Accounting & Billing

Accounting and billing information sources are important in order to build a complete user profile. Semantically enriched data extracted from accounting and billing information (e.g. statistical models) can be correlated with other types of information (e.g., signaling) to detect deviations from normal users' behavior. However, due to privacy concerns and the limited access to billing information sources, alternative methods to compile synthetic billing datasets need to be researched. The conclusion of our research and the outcome of the consortium discussion is that the billing data and expenses profiles can be accurately inferred by combining the user call detail records with the service costs and expenses per region.

3.4 Call Detail Records

In this section, we present Call detail record (CDR) information sources. CDR datasets are very important for NEMESYS as their records document the details of all telecommunication transactions that passed through the central node.

3.4.1 Orange "Data for Development" Challenge

Orange is a French multinational telecommunications corporation and a global provider for mobile phone, landline, Internet, mobile internet, and IP television services. The datasets origin from the *Data for Development (D4D) Challenge* [11], which is an open challenge, set by Orange, who has provided anonymized records of their customers in the Ivory Coast.

- **Content**

  The data were collected for 150 days, from December 1, 2011 until April 28, 2012. The original set of Call Detail Records (CDRs) contains 2.5 billion calls and SMS exchanges between around five million users. CDRs provide timestampted information on the caller id, the callee id, the call duration and the antenna code. For the challenge four mobile phone datasets were released. These datasets are based on the anonymized Call Detail Records (CDR). The datasets are:

  1. *Antenna-to-antenna traffic* on an hourly basis
  2. *Individual trajectories* for 50,000 customers for two week time windows with antenna location information
  3. *Individual trajectories* for 500,000 customers over the entire observation period with sub-prefecture location information
  4. A sample of *communication graphs* for 5,000 customers.

  The provided datasets contain also the geographical positions of cell phone antennas. Orange considers the exact antenna location as sensitive information and therefore the locations have been slightly blurred so as to protect Orange's commercial interests. The 255 sub-prefectures of Ivory Coast along with Orange's cell phone towers are shown in Figure 3.

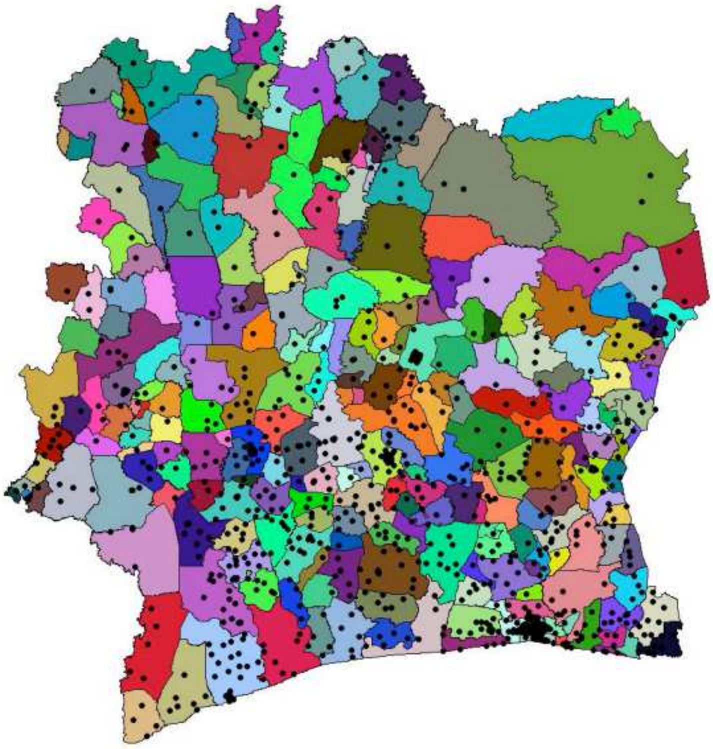


Figure 3: Orange's cell phone towers in Ivory Coast and sub-prefectures administrative regions

- **Format**

  All datasets are available in Tabulation Separated Values (TSV) format.

- **Tools**

  The first dataset can be visualized with *Geofast* [12]. Geofast is a web-based tool for the interactive exploration of mobile phone data. The data is aggregated on different administrative levels and users are able to select administrative regions and visualize the amount of communication traffic on selected days.

- **Access**

  Restricted only to the challenge participants. The orange's committee was contacted to grant access to the partners of the NEMESYS project. Permission is pending.

3.4.2 IEEE VAST 2008 - Challenge Datasets

The IEEE Symposium on Visual Analytics Science and Technology (VAST) annually holds a challenge which focuses on data analytics using visualization techniques. For these challenges certain datasets are released each year.

- **Contents**

  For the *IEEE VAST 2008 challenge* [13] four heterogeneous synthetic datasets were generated. The datasets are a blend of computer and hand-generated data and thus are considered synthetic. For our purpose the Cell Phone Calls dataset offers interesting data and in certain cases can be correlated with the other datasets (e.g., geospatial).

The Cell Phone Calls dataset contains:

- From: Identifier for the calling cell
- To: Identifier for the receiving cell
- Datetime: a yyyymmdd hhmm format date and time
- Duration: duration of the call in seconds
- Cell Tower: Location of the call origination cell tower

- **Format**

  The dataset is in CSV format.

- **Tools**

  No tools are provided.

- **Access**

  The dataset is publicly released and requires registration.

3.5 Evaluation & Comparison

In order to efficiently compare the information sources outlined above we designed Table 2. This table highlights specific features that we took under consideration for the evaluation of the datasets. A common characteristic of the datasets in this category is the fact that no updates are provided.

In our evaluation, we found the orange dataset to be an excellent source of information. The orange dataset is one of the most comprehensive datasets we were able to find and propose it to be one of the main information sources for NEMESYS. The data collection involved a wide range and number of users. This resulted in a dataset suitable for extracting diverse types of information and a vast amount of users for a lengthy amount of time. However, orange is keeping the dataset private and explicit permission has to be given to the NEMESYS project.

On the other hand, the IEEE VAST 2008 dataset is publicly released and it is one of the few information sources that provide synthetically generated datasets. This can be beneficial in the NEMESYS case, since the lack of noise can be valuable when testing and training anomaly detection algorithms. Additionally, it has very similar features with the orange dataset but for a much smaller number of users. The lack of the “Messages” feature is not of great importance and can be made up by incorporating the SMS messages collected by the SMS Corpus.

Table 2: Operating and Maintenance Centers Monitoring

ID / Name	Timestamp	Calls	Messages	Tower	Device	Updates	Tools	Size	Dataset #	R/S	Year	Access
Orange "Data for Development" Challenge	✓	✓	✓	✓	✓	-	✓	5 Million	4	R	2012	Pending
IEEE VAST 2008 - Challenge Datasets	✓	✓	-	✓	✓	-	-	~400	4	S	2008	Public

- ID/Name: Name of the information source or dataset.
- Timestamp: If the data records are time stamped.
- Calls: If the released datasets include information on the user calls.
- Messages: If the released datasets include information on the user's messaging activity.
- WiFi/Bluetooth: If the released datasets include information on the user's Bluetooth/WiFi activity.
- Tower: If the released datasets include information on the used cell tower.
- Device: If the released datasets include information on the user's device.
- Updates: If the datasets are periodically updated.
- Size: Sample size is refers to the number of user/devices monitored.
- Dataset #: Number of released datasets.
- R/S: If the information source compiles the datasets based on real or synthetic data.
- Year: Year of release or last update.
- Access: Access policy as defined by the source and the NEMESYS project's permissions.

4 Databases, repositories and analysis tools

4.1 Introduction

In recent years, due to the wide spread use of mobile devices and the telecommunications evolution the attackers changed their modus operandi in order to exploit this new domain for their malicious purposes. To countermeasure this, numerous information sources provide data sets containing information that can be of great interest for the NEMESYS objectives. In this section, we present sources that maintain databases and repositories for mobile and non-mobile malware and vulnerabilities in mobile environments. These sources will help the consortium to comprehend the mobile threat landscape (T7.1) and build robust anomaly detection and correlation mechanisms for tasks T4.1 and T5.1.

4.2 Malware

There is a wide variety of malware developed for mobile devices and can be analyzed to identify patterns in the attacker modus operandi and accordingly develop the NEMESYS anomaly detection algorithms. The information sources of this section, maintain repositories of mobile and non-mobile malware and in some cases detailed analysis and classification of each malware is available.

4.2.1 Berkeley files - Mobile Malware Survey

The *Mobile Malware Survey* [14] is a survey of iOS, Android, and Symbian 9.x malware [15] that spread in the wild from January 2009 to June 2011 and the results of the survey are distributed online (the website is updated automatically every 5 minutes).

- **Content**

  The behavior of the pieces of malware is reported, their permissions (for Android), and their certificate types (for Symbian). The list contains online malware nor personal spyware or grayware. Grayware refers to applications that behave in a manner that is annoying or undesirable, and yet less serious or troublesome than malware (e.g., adware). Among others, the list provides details on: the malware name, the affected platform, its type, its current spread, the location, date of detection, its modus operandi and links for additional technical details.

- **Format**

  The dataset is provided online as an online Spreadsheet which is periodically updated.

- **Tools**

  No tools are provided for analysis or parsing.

- **Access**

  The dataset is publicly accessible.

4.2.2 F-Secure – Mobile Security List

F-Secure Corporation [16] is an anti-virus and computer security company, which tracks global internet and mobile security threats.

- **Content**
  In the *Mobile Security List* [17], the latest mobile threats (e.g., malware, spyware) are given, along with brief description of the threat, screenshots, disinfection instructions and additional details. Although, the list is relatively small, it is constantly updated and constitutes a quality source of the latest mobile malware definitions.
- **Format**
  The dataset is provided in XML/RSS format.
- **Tools**
  No processing tools are provided.
- **Access**
  The dataset is public.

4.2.3 Android Malware Genome Project

The *Android Malware Genome Project* [18] attempts to systematize or characterize existing Android malware.

- **Content**
  During the project more than 1,200 malware samples that cover the majority of existing Android malware families, ranging from their debut in August 2010 to October 2011. What differentiates the project from malware repositories is that malware is analyzed and characterized from various aspects, including installation methods, activation mechanisms and nature of carried malicious payloads.
- **Format**
  A zip archive containing malicious APK (Android application package) files categorized in families.
- **Tools**
  No processing tools are provided.
- **Access**
  The dataset is private. Permission from the authors is required in order to gain access [19]. The NEMESYS project has access to this dataset.

4.2.4 Panda Security – List of Viruses & Panda Mobile

Panda provides a *list of all viruses* [20] catalogued in Panda Security's Collective Intelligence servers. Additionally, Panda Security maintains a *list of mobile threats* [21].

- **Content**

The panda security list of viruses contains 173.068 lemmas, which mostly affecting personal computer. Furthermore the mobile threats historical record contains 152 threats designed for mobile devices. For each malware detailed information are provided, including its common & technical name, its type and effects. Besides these, the time and date of the first detection and anti-virus definitions updates are shown. Furthermore, panda attempts to provide a classification scale based on the threat level, the potential damage and the distribution.

- **Format**
  The data are available in text/html format.
- **Tools**
  No processing tools are provided.
- **Access**
  The dataset is publicly available on the CloudAntivirus website and the pandasecurity.com.

4.2.5 Kaspersky Lab – SecureList

Kaspersky Lab is a developer of secure content and threat management systems and the world's largest privately held vendor of software security products. Kaspersky maintains Securelist.com, a computer security portal that is devoted to educating the general public about different aspects of internet security.

- **Content**

  The *securelist.com* [22] contains 10284775 signatures for mobile and non-mobile malware. The list provides detailed information on the viruses and includes technical information for their operation. For each threat a detailed description is provided and various time-related information regarding detection, release and publication. Moreover, supplementary technical details are available such as infection installation, the malicious payload, and removal instructions.
- **Format**
  The list is available in a text/html format.
- **Tools**
  No processing tools are provided.
- **Access**
  The dataset is publicly available on the securelist.com.

4.2.6 SOPHOS

Sophos is a developer and vendor of security software and hardware, providing endpoint, encryption, email, web, mobile and network security as well as Unified Threat Management products. They developed the Threat Center [23] where a wide variety of threats is analyzed.

- **Content**

  The Threat Center contains information on viruses, spyware, suspicious behavior and files, adware, PUAs, and controlled applications and devices for a variety of platforms. For each entry the type of threat, its prevalence, different aliases, affected operating systems and a runtime analysis are given.

  The prevalence serves as a malware ranking meter for the Threat Center. Moreover, SOPHOS provides signatures (e.g., MD5, SHA-1, CRC-32) of the different releases of the same malware along with their file size and date of capture.

- **Format**
  The list is in a Text/html format. The latest threats list is in RSS/XML.
- **Tools**
  No processing tools are provided.
- **Access**
  The dataset is publicly available.

4.2.7 Anubis - Analyzing Unknown Binaries

Anubis [24] is a tool for analyzing the behavior of Windows executables and Android APK files (Andrubis) with special focus on the analysis of malware. The project is developed by *isecLAB* [25].

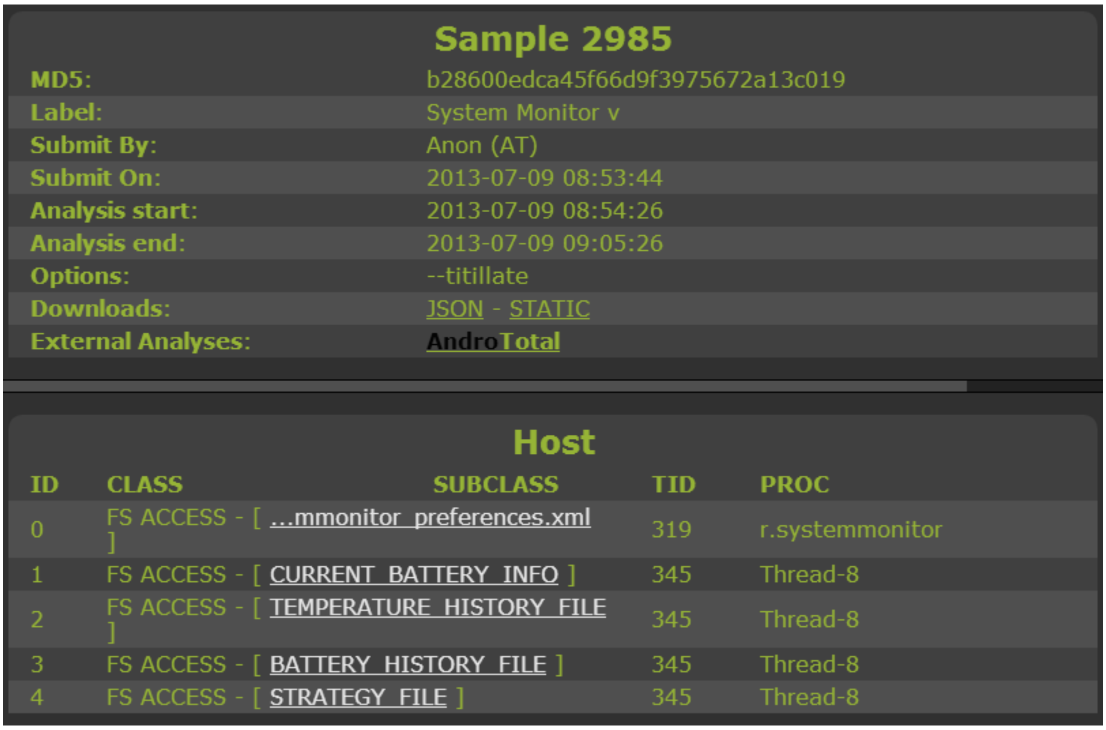
- **Contents**
  After the analysis the tool generates a report file that contains detailed data about modifications made to the Windows registry or the file system, interactions with the Windows Service Manager or other processes logs all generated network traffic [26]. Each report contains some general information, a static analysis report, a dynamic analysis report and screenshots. The static analysis focuses on: Activities, Services, Broadcast Receivers, Required Permissions, Used Permissions, Features and Urls, while the dynamic analysis on: File Operations, Network Operations, Started Broadcast Receivers, Started Services and Native Libraries Loaded. A very interesting feature is the malware network activity monitoring that the service supports.
- **Format**
  The report is available in text/html and XML format. Additionally the traffic files are in pcap format.
- **Tools**
  The service can be accessed via a web page. Additionally, automated submission of sample can be done with a python script or a submission URL.
- **Access**
  Samples of the dataset are public. There are no restrictions on the use of the service. The full dataset of analyses is private.

4.2.8 CooperDroid

CopperDroid [27] is a research effort to automatically perform out-of-the-box dynamic behavioral analysis of Android malware. The CopperDroid project is designed and developed by the Information Security Group (ISG) of Royal Holloway, University of London and the Network and Security Lab of Università degli Studi di Milano.

- **Content**

  CopperDroid generates a unified analysis to characterize low-level OS-specific and high-level Android-specific behaviors. Furthermore, CopperDroid's VMI-based dynamic system call-centric analysis is able to faithfully describe the behavior of Android malware whether it is initiated from Java, JNI or native code execution. CopperDroid features a stimulation technique to improve code coverage, aimed at triggering additional behaviors of interest.



Sample 2985

MD5:	b28600edca45f66d9f3975672a13c019
Label:	System Monitor v
Submit By:	Anon (AT)
Submit On:	2013-07-09 08:53:44
Analysis start:	2013-07-09 08:54:26
Analysis end:	2013-07-09 09:05:26
Options:	--titillate
Downloads:	JSON - STATIC
External Analyses:	AndroTotal

Host

ID	CLASS	SUBCLASS	TID	PROC
0	FS ACCESS - [...mmonitor_preferences.xml]		319	r.systemmonitor
1	FS ACCESS - [CURRENT_BATTERY_INFO]		345	Thread-8
2	FS ACCESS - [TEMPERATURE_HISTORY_FILE]		345	Thread-8
3	FS ACCESS - [BATTERY_HISTORY_FILE]		345	Thread-8
4	FS ACCESS - [STRATEGY_FILE]		345	Thread-8

Figure 4: CooperDroid Report

- **Format**

  The reports are available in JSON and in a static format, as well. Traffic records are kept and released as pcap.

- **Tools**

  The web interface is available to all users. However, it is easier to automate the task of submission with an automation script.

- **Access**

  The NEMESYS project has been granted permission to use the service and a user account has been issued. This is important, since the automatic submission can be easily implemented and the captcha protection mechanism is not deployed for registered users. Additionally, an API is being designed and will be available in the future.

4.2.9 Georgia Institute of Technology – Open Malware

Open Malware [28] was formed as a resource for the computer security community by Danny Quist. Its primary emphasis is on malware collections and analysis for the purpose of improving people's abilities to defend their networks.

- **Content**

The database, which serves also as a repository, contains malware samples affecting all widely used platforms including those used in mobile devices. The database is not available as a whole, but the user can get related listings by using the search functionality. For each malware the database provides:

- Various Signatures
- The original filename
- Filename according to antivirus suites
- Date of detection
- Other Information Sources

Additionally, the database provides resources such as live copies of malicious software, md5sums to search on and analysis of the malware to the general public. According to its claims it has the largest publicly available malware collection on the Internet.

- **Format**
  The list is in plaintext format.
- **Tools**
  No tools are provided.
- **Access**
  The dataset is public and available online.

4.2.10 WildList - Virus Bulletin

The *WildList Organization* [29] aims to provide accurate, timely and comprehensive information about computer viruses to both users and product developers. For this purpose, the organization compiles *The WildList* [30], a list of the viruses currently spreading throughout a diverse user population.

- **Content**
  The list is plain and contains only the signatures of malware detected, the proposed naming and the reporter that submitted the entry. Please note that, the list should not be considered a list of "the most common viruses", since no specific provision is made for a commonness factor. The list is released monthly and features an extended version, as well.
- **Format**
  The list is in plaintext format.
- **Tools**
  No tools provided.
- **Access**
  Access to the list is public.

4.2.11 Microsoft Malware Encyclopedia

Microsoft maintains the *Microsoft Malware Encyclopedia* [31] which contains descriptions for malware detected by Microsoft security products in various platforms.

- **Content**

The encyclopedia includes malware, which affects Android, Symbian and other mobile operating systems. The Microsoft malware definitions are generated using telemetry from millions of computers, and by operating a global network of research and response labs. For each threat there is a separate page which includes a) summary, b) symptoms, c) technical information, d) prevention methods, and e) recovery information. Microsoft maintains its own classification system for the threats based on the severity of the threat, as defined by it is security products.

- **Format**
  The database is available in text/html format.
- **Tools**
  A collection of tools by various vendors is provided for protection against threats [32]. However, these tools are not directly related with the NEMESYS objectives.
- **Access**
  The access to the encyclopedia and the tools is public.

4.2.12 FortiNet – Fortiguard Encyclopedia

The F*ortiguard Encyclopedia* [33] is developed by Fortinet a company which specializes in network security appliances. Part of the Fortiguard Encyclopedia refers only to threats against mobile devices [34] and provides various details.

- **Content**
  The whole encyclopedia contains 4.226.131 entries. However, for the NEMESYS' purposes we focus mostly on the mobile threats part of the database. The malware included in this database affects iOS, Android and Symbian mobile devices and is updated in daily basis. More details available for the majority of virus variants:
    - Analysis
    - Technical Details
    - Aliases
    - Symptoms
    - Recommended Action

  The database is updated at least once every two days.

- **Format**
  The database is in html format.
- **Tools**
  A collection of tools by various vendors is provided for protection against threats.
- **Access**
  The database is public and available online.

4.2.13 VirusShare

VirusShare [35] is a malware repository aiming to provide security researchers, incident responders and forensic analysts access to samples of malicious code and details regarding each sample.

- **Content**
  The repository contains 9.602.732 samples and their hashes. The malware samples affect personal computer operating systems, as well as mobile platforms. For each sample there is a report page which includes: a number of signatures (MD5, SHA-1, SHA256, SSDeep), the size of the malware, the file type, detections by popular solutions, ExIF Data and the submission date. The repository is published as lists of MD5 hashes for all of the malware samples contained in each of the zip files shared via the torrents. Each list is published after each torrent is uploaded. Each list is a 4.3MB plain text file with one hash per line [36]. The database is updated regularly and usually new samples are added daily.
- **Format**
  The MD5 hash lists are in plaintext format. The samples are in a compressed archive.
- **Tools**
  No tools are provided.
- **Access**
  Only MD5 Hashes are public. Search and download functionalities are restricted. Access is possible only after approval. The NEMESYS project was granted access permissions.

4.2.14 Malware.lu

The *malware.lu* [37] is a malware repository and a source for malicious software signatures.

- **Content**
  Contains 5,572,622 samples of malicious software, including software for mobile devices. Additionally, a list of signatures of the above malware is release. The hash formats are: md5, sha1 and sha256.
- **Format**
  Hashes list is provided in plaintext format.
- **Tools**
  An API is available to users after approval.
- **Access**
  Access to hashes is public. The malware repository requires user account approved by the owners. Access to the API for the NEMESYS project was granted.

4.2.15 Contagio Mobile – Mobile Malware Mini Dump

Contagio mobile mini-dump [38] is a part of *Contagio dump* [39]. Both databases are collections of the malware samples, threats, observations and analyses.

- **Content**

From the two lists the Mobile Malware Mini Dump has the greatest interest for NEMESYS. The list contains malware targeting mobile devices and occasionally brief analysis of the modus operandi is provided. The database does not have a stable release template and thus certain details are omitted per case. The only details that are consistently available with each sample are: filename, file Size and its MD5 Signature. Other information may be offered for certain samples.

- **Format**

  Samples can be downloaded individually or in one zip archive.

- **Tools**

  A collection of tools by various vendors is provided for protection against threats.

- **Access**

  The dataset is publicly available. However, newest samples (after 2012) require a password to be accessed. The password pattern has been given to the NEMESYS project by the owner of the dataset.

4.2.16 VirusTotal

VirusTotal [40] is a free service for analyzing samples and suspicious URLs. The samples are analyzed by 44 different antivirus engines. At the end of each analysis, a report is generated containing the malware name, the hash ID of the sample (MD5, SHA1, and SHA-256), the initial filename, the type of file (e.g., an image), etc. For some samples, the sample behavior is also added to the report. By behavior, we intend the actions of the malware in the system such as the files that are read, written, the communications, etc. Besides the analyses, the service also enables to retrieve the uploaded malware by their name and their hash, and download them.

- **Content**

  VirusTotal contains approximately 200 million of binaries. Each binary is identified by its hash (MD5, SHA-1, or SHA-256). This database is in constant evolution with 600,000 new samples received every day.

- **Format**

  Reports can be accessed either by the browser in HTML format or directly by the VirusTotal API which provides a JSON report.

- **Tools**

  The VirusTotal API enables automated analyses of samples and URLs, while sample downloading is also supported.

- **Access**

  For accessing to the database, the ideal is to use the VirusTotal API. Nevertheless for using it, users need to previously create an account. There are two types of API: public and private API. Private API provides more advantages than public API. For instance, it enables users to get more information on the JSON report but also increases the daily quota on the number of analyses, the number of reports accessed, etc.

4.3 Vulnerability

It is of special interest for the NEMESYS project to develop anomaly detection mechanisms that correlate network traffic and activity with known vulnerabilities of mobile platforms. Vulnerability databases list a variety of security flaws that can be exploited by malicious attackers in order to infect mobile devices and spread the infection to all vulnerable systems. Attacks of this kind are not based on social-engineering but exploit the deficiencies of the system and thus often no action from the end user is required. The most commonly exploited security vulnerabilities are either 0-day flaws or flaws that are not yet officially patched by the software developer.

NEMESYS anomaly detection algorithms (e.g., T4.1, T4.2) can take advantage of this fact and detect, classify and monitor malware threats based on the traffic of end user devices. More specifically, vulnerability information sources can provide a valuable knowledge base against which can be verified and potentially harmful request.

4.3.1 Symantec – Security Response Vulnerabilities

Symantec is one of the world's largest software companies and provides security, storage and systems management solutions. Symantec owns the Security Response organization which is an antivirus and computer security research group and maintains the *Security Response Vulnerabilities* [41] database.

- **Content**
  The database contains numerous vulnerabilities affecting various applications and operating systems. This vulnerability collection contains threats from 1997 up to 2013. Even though, its size is moderate (in comparison with other sources outlined in this document) the severity of included vulnerabilities is considered high. A severity rank of each entry is provided along with:
    - Risk
    - Date Discovered
    - Description
    - Technologies Affected
    - Countermeasures Recommendations

  The database is updated constantly and thus it is important to use a recent instance or access a feed.
- **Format**
  The database is available in html and RSS/XML format.
- **Tools**
  No tools are provided.
- **Access**
  The database is public and available online.

4.3.2 NIST - National Vulnerability Database

National Vulnerability Database (NVD) [42] is a product of the *NIST Computer Security Division* [43] and is sponsored by the Department of Homeland

Security's National Cyber Security Division. NVD is the U.S. government repository of standards based vulnerability management data represented using the Security Content Automation Protocol (SCAP).

- **Content**
  The database contains 56536 vulnerabilities and aims to enable automation of vulnerability management, security measurement and compliance. It includes databases of security checklists, security related software flaws, misconfigurations, product names and impact metrics. Each vulnerability summary contains:
  - Original release date
  - Last revised date
  - Source
  - Overview
  - Impact
  - References to Advisories, Solutions, and Tools
  - Technical Details

  The vulnerabilities affect all kinds of platforms including mobile operating systems (e.g., iOS, Android, Symbian) [44]. One can browse for vendors, products and versions and view CVE entries and vulnerabilities related to them. New vulnerabilities are added in the database in daily basis.

- **Format**
  The database is available in html, xml & SCAP [45] format and is updated daily.
- **Tools**
  SCAP validated product list.
- **Access**
  Public. The dataset can be accessed without restrictions online.

4.3.3 OSVDB - Open Sourced Vulnerability Database

OSVDB [46] is an independent and open sourced web-based vulnerability database created for the security community. Its goal is to provide accurate, detailed, current, and unbiased technical information on security vulnerabilities and maintain a truly comprehensive vulnerability database with extended features.

- **Content**
  The project currently covers 92,615 vulnerabilities, spanning 77,788 products from 4,735 researchers. OSVDB collects information on vulnerabilities on all types of products including many entries concerning software in mobile platforms. The difference from other databases is that it is not a plain list but provides quality information on each entry. More specifically, each record contains a timeline, detailed description and classification, proposed solution

(if available), products affected, related references, the CVSSv2 Score (if existent) and additional expert comments.

- **Format**
  Results from API calls are returned in either XML or CSV.
- **Tools**
  OSVDB Tools are available in the wild.
- **Access**
  Public and Open Source. An API and other export methods are provided by riskbasedsecurity.com

4.3.4 IBM Internet Security Systems – X-Force Vulnerability Search

The *IBM X-Force Vulnerability database* [47] is a very comprehensive threats and vulnerabilities database. It provides the latest information about cyber threats and mobile threats, in particular.

- **Content**

  The database is composed of more than 70,000 computer security vulnerabilities with specific details such as: risk level, description, consequences, remedy, platforms affected and date of report.
- **Format**
  The database is in text/html format.
- **Tools**
  No tools are provided.
- **Access**
  The access is public.

4.4 Evaluation & Comparison

In order to efficiently compare and evaluate the information sources outlined above we designed Table 3. The table presents important features and details for the specific category of sources and can be consulted when choosing datasets to be included to the data collection infrastructure.

We consider the existence of mobile malware samples to be of great importance. Based on this, we propose the following up-to-date mobile malware repositories: VirusShare, Android Malware Genome Project (AMGP), Contagio Mini Dump and Malware.lu. All these sources provide mobile malware samples of high quality and especially AMGP and VirusShare offer very rich and actively updated collections of mobile malware. Samples from these malware repositories can be analyzed using ANUBIS and the network traffic file can be utilized in order to perform automated analysis of malware activity and signaling patterns. Additionally, package and traffic signatures are provided by many sources (e.g., VirusTotal, OpenMalware) and will be very useful during the anomaly detection phase. For the NEMESYS objectives it will be beneficial to merge or combine the already large repositories and databases to achieve even better results. This is rather interesting, since we can further enrich our data sets both in terms of quality and variety of entries.

Besides malware, vulnerability information can be used to create signatures and correlate them with malware activity recordings from ANUBIS to develop robust anomaly detection mechanisms and comprehensive visualization techniques. The NIST Vulnerability database and the Open Source Vulnerability Database are great sources of information for mobile security flaws that are often exploited by the malware creators. NIST is supported by the US government and thus we consider it the most reliable of all vulnerability sources listed in this document. The X-force Vulnerability Search is another well-known information source for vulnerabilities, which is maintained by IBM. However, even though it lists a large number of vulnerabilities we recommend using it mostly as a supplementary source because of difficulties in parsing its data.

Finally, we also outlined various malware encyclopedias which analyze different kinds of malware. These sources are regularly updated and either they contain mobile and non-mobile malware analyzes (e.g., Microsoft Encyclopedia) or they specialize exclusively in mobile threats (e.g., Berkeley files). These sources are useful for NEMESYS, since they give a deep insight of each threat. On the other hand, we found that a special parsing application is needed for each one of them. Thus, we propose to use only selected malware analysis sources. These are: FortiNet because of its size and wealth of information and the Sophos Threat Center which provides signatures of the examined malware. Finally, according to the specific needs of the task, other sources of these proposed ones can also be used if necessary.

Table 3: Databases, repositories and detection tools

ID / Name	*Malware*	*Vulnerabilities*	*Mobile*	*Analysis*	*Samples*	*Signatures*	*Updates*	*Tools*	*EPF*	*Size*	*Year*	*Access*
Berkeley files - Mobile Malware Survey	✓	-	✓	✓	-	-	-	-	-	46	2011	Public
F-Secure - Mobile Security List	✓	-	✓	✓	-	-	✓	-	-	N/A	2013	Public
Android Malware Genome Project	✓	-	✓	✓	✓	-	-	-	-	1200	2011	Granted
Panda Security - List of Viruses & Panda Mobile	✓	-	✓	✓	-	-	✓	-	-	N/A	2013	Public
Kaspersky Lab – SecureList	✓	-	✓	✓	-	-	✓	-	-	N/A	2013	Public
SOPHOS - Threat Center	✓	-	✓	✓	-	✓	✓	-	-	N/A	2013	Public
Anubis - Analyzing Unknown Binaries	✓	-	✓	✓	-	✓	✓	✓	✓	-	2013	Public
Georgia Institute of Tech - Open Malware	✓	-	✓	-	✓	✓	✓	-	-	N/A	2013	Public
WildList - Virus Bulletin	✓	-	-	-	-	✓	✓	-	-	>2500	2013	Public
Microsoft Malware Encyclopedia	✓	-	✓	✓	-	-	✓	-	-	N/A	2013	Public
FortiNet - Fortiguard Encyclopedia	✓	-	✓	✓	-	-	✓	-	-	>4 Million	2013	Public
VirusShare	✓	-	✓	-	✓	✓	✓	-	-	11080	2013	Granted
Malware.lu	✓	-	N/A	✓	✓ (PR)	✓	✓	-	-	>5 Million	2013	Granted
Contagio Mini Dump	✓	-	✓	✓	✓	✓	✓	-	-	111	2013	Public
VirusTotal	✓	-	✓	✓	-	✓	✓	✓	✓	-	2013	Public
Symantec – Security Response Vulnerabilities	-	✓	✓	✓	-	-	✓	-	✓	~1600	2013	Public
NIST - National Vulnerability Database	-	✓	✓	✓	-	-	✓	✓	✓	56690	2013	Public
Open Source Vulnerability Database	-	✓	✓	✓	-	-	✓	✓	✓	93,018	2013	Public
X-Force Vulnerability Search	-	✓	✓	✓	-	-	✓	-	-	N/A	2013	Public

- ID/Name: Name of the information source or dataset.
- Malware: The datasets include information on malware.
- Vulnerabilities: The datasets include information on vulnerabilities.
- Mobile: A proportion or all malware/vulnerabilities affect mobile devices.
- Analysis: Detailed analysis is provided for the malware/vulnerabilities.
- Samples: Malware samples are provided (not necessarily for mobile environments).
- Signatures: Malware/Vulnerabilities signatures are supported.
- Updates: If the datasets are periodically updated.
- Tools: Utilities for data manipulation are offered
- EPF: The datasets are in an Easy to Parse Format
- Size: Sample size is refers to the number of records
- Year: Year of release or last update.
- Access: Access policy as defined by the source and the NEMESYS project's permissions.

5 Internet Activity Monitoring

5.1 Introduction

The information sources of this chapter provide data essential to develop and train algorithms capable of correlating reported suspicious IP addresses and domains with possible traces of attacks targeting the mobile infrastructure and/or the end user devices. Initially, we will focus on sources that apply DNS monitoring techniques that are efficient in identifying large scale attack campaigns. In the second section, we present honeypot projects, while in the third various malware block lists are introduced. The "DNS monitoring" sources will provide valuable information for the scenarios development in task T7.2, while the honeypots will provide a solid foundation for building a honeyclient (task T3.2) and detecting abnormal events using correlation analysis (T5.1). The malware block lists are proposed as additional sources to be utilized in tasks T4.1, T5.3 and T7.1.

5.2 DNS Monitoring & Network Monitoring

DNS is the way that most modern malware is connecting back to their operators' command-and-control infrastructure in order to obfuscate their operation overall architecture from security researchers. Having rich DNS traffic information is very important for identifying malicious behavior and NEMESYS should consider information sources that capture DNS traffic. Such a technique will enable reliable anomaly detection. As a result, it is of particular interest for the NEMESYS project to search and review the most complete DNS monitoring projects and services that are publicly available. Later on, we could correlate this information with other possible traces of attacks and achieve very high detection rates with minimum false positives.

5.2.1 ISC - DNSDB

The internet systems consortium (ISC) aims to provide the industry a greater insight into how the cyber-criminals are using DNS to violate the Internet. *DNSDB* [48] is a database, maintained by ISC, which stores and indexes both the passive DNS data (available via ISC's Security Information Exchange) as well as the authoritative DNS data that various zone operators make available.

- **Content**

  DNSDB stores and indexes both the passive DNS data available via ISC's Security Information Exchange as well as the authoritative DNS data that various zone operators make available. It supports search for individual DNS RRsets and provides additional metadata for search results such as first seen and last seen timestamps as well as the DNS bailiwick (zone) associated with an RRset. DNSDB also has the ability to perform inverse or rdata searches. DNSDB features two lookup modes (RRset and Rdata) depending on the user’s desired query. Individual DNS RRsets are possible and additional metadata for search results such as first seen and last seen timestamps as well as the DNS bailiwick associated with an RRset are provided. It also has the ability to perform inverse or rdata searches.

  Rrset query result fields

  - Rrname: The owner name of the RRset in DNS presentation format.
  - Rrtype: The resource record type of the RRset, either using the standard DNS type mnemonic, or an RFC 3597 generic type, i.e. the string TYPE immediately followed by the decimal RRtype number.
  - Rdata: An array of one or more Rdata values. The Rdata values are converted to the standard presentation format based on the rrtype value. If the encoder lacks a type-specific presentation format for the RRset's rrtype, then the RFC 3597 generic Rdata encoding will be used.
  - Bailiwick: The "bailiwick" metadata value described in the section above.
  - Count: The "count" metadata value described in the section above.
  - time_first, time_last: UNIX epoch timestamps with second granularity indicating the first and last times the RRset was observed via passive DNS replication. Will not be present if the RRset was only observed via zone file import.
  - zone_time_first, zone_time_last: UNIX epoch timestamps with second granularity indicating the first and last times the RRset was observed via zone file import. Will not be present if the RRset was only observed via passive DNS replication.

  Rdata query result fields

  - Rrname: The owner name of the resource record in DNS presentation format.
  - Rrtype: The resource record type of the resource record, either using the standard DNS type mnemonic, or an RFC 3597 generic type, i.e. the string TYPE immediately followed by the decimal RRtype number.
  - Rdata: The record data value. The Rdata value is converted to the standard presentation format based on the rrtype value. If the

encoder lacks a type-specific presentation format for the resource record's type, then the RFC 3597 generic Rdata encoding will be used.

- Count: The number of times the resource record was observed via passive DNS replication.
- time_first, time_last: UNIX epoch timestamps with second granularity indicating the first and last times the resource record was observed via passive DNS replication.
- zone_time_first, zone_time_last: UNIX epoch timestamps with second granularity indicating the first and last times the resource record was observed via zone file import.

Finally, the database supports IPv4 and IPv6 addresses, network prefixes and dash-delimited address ranges. Prefixes and address ranges on non-octet boundaries are also allowed.

ns07.ipv6testu-digkwx5w.info.	AAAA	::101.45.75.219
ns07.ipv6testu-y4gqrwyj.info.	AAAA	::101.45.75.219
ns1.7dns.info.	A	0.0.0.1
ns1.crudefuel.info.deleted.gandi.info.	A	0.0.0.1
ns2.crudefuel.info.deleted.gandi.info.	A	0.0.0.1

Figure 5: Part of the results in DB query

- **Format**
  The database is in text/html format. Additionally, an API is available for bulk queries. The DNSDB API supports two result formats: the ad hoc "text" format which is reminiscent of the DNS master file format, and the "json" format which returns one result per line encoded in JSON format. The two formats use the exact same URL scheme. A specific result format can be selected by specifying an Accept header in the HTTP request.
- **Tools**
  Various tools are provided. A set of tools is provided in order to setup a sensor and join the global passive DNS network. CERTH set up a passive DNS (pDNS) monitoring sensor in order to collect and generate datasets for analysis by the NEMESYS anomaly detection and visual correlation algorithms. Even though, our datasets are not expected to be very large the firsthand experience in DNS data collection is valuable [49]
- **Access**
  The access to the database is restricted. The NEMESYS project was granted access to the database after request. Access to API requires additional permission.

5.2.2 Iseclab – EXPOSURE BlackList

EXPOSURE [50] is a service that identifies domain names that are involved in malicious activity by performing large-scale passive DNS analysis. EXPOSURE has been developed as part of the *WOMBAT project* [51].

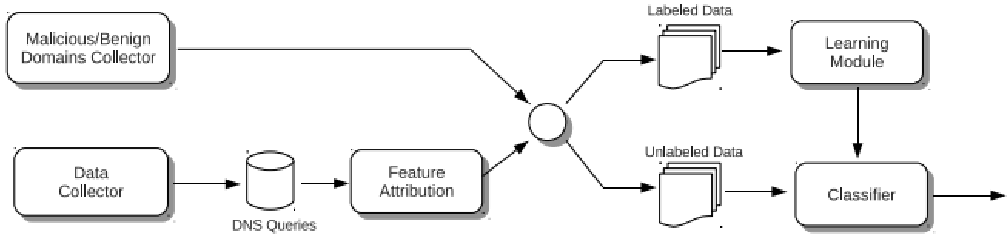


Figure 6 EXPOSURE Overview

The ability to identify malicious domains as soon as they appear would significantly help mitigate many Internet threats that stem from botnets, phishing sites, malware hosting services and others. The Exposure tool identifies domain names that are involved in malicious activity by performing large-scale passive DNS analysis (Figure 6). The tool is based on the fact that it is beneficial to monitor the use of the DNS system for signs which indicate that a certain name is used as part of a malicious operation. This is because malicious services are usually as dependent on DNS services as benign services.

- **Content**
  The data collected by the tool are compiled in different datasets. Experiments with real-world data sets show that the tool effectively identifies unknown malicious domains that are misused in a variety of malicious activity (e.g., botnet command and control, spamming, phishing). For each malicious domain the dataset includes:
    - *Geo-location Information*
    - *Malicious timeline* of requests
    - *Correlated Network Topology* where the domain is associated with other IP nodes. For each of these nodes a further analysis of related domains is available.

The database is updated on daily basis and can be found on the project's website.

- **Format**
  The dataset is provided in the form of a blacklist in plaintext format.
- **Tools**
  Advanced search functionality is available along with the blacklist. The search enables the user to scan for malicious DNS based on domain names, IP addresses or detection date.
- **Access**
  The list is publicly available on the project's website. However, temporarily the blacklist is not released due to some issues with commercial users.

5.2.3 CAIDA - UCSD Network Telescope

The Cooperative Association for Internet Data Analysis (CAIDA) is a collaborative undertaking among organizations in the commercial, government and research sectors. It aims to promote the greater cooperation in the engineering and the maintenance of a robust and scalable global Internet infrastructure. CAIDA collects several different types of data at geographically and topologically diverse locations, and makes this data available to the research community to the extent possible while preserving the privacy of individuals and organizations who donate data or network access. Among these collections, there are datasets for malicious network activity and malicious scanning activities (e.g., DoS attack dataset, worm dataset). The *UCSD Network Telescope* [52] is a passive traffic monitoring system applied on global traffic. CAIDA archives and analyzes the telescope data and release datasets for security researchers.

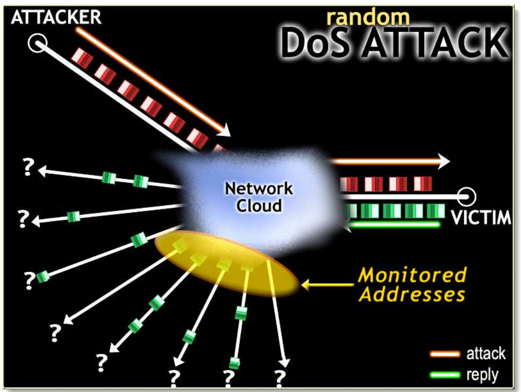


Figure 7: UCSD network telescope operation overview

- **Content**
  The UCSD Network Telescope datasets provide a unique view of anomalous traffic with no legitimate destination carried on the Internet. Among others, Network Telescope monitors malicious events including Denial-of-Service attack backscatter, Internet worms, and host scanning. The UCSD Network Telescope observes many types of scans continually, including ping based scans for the existence of a device at a given IP address, sequential scans of ports on a single IP address, methodical scans for a single or a small number of vulnerable ports sequentially through an IP address range, and even scans utilizing TCP resets (Figure 7). Common malicious events are the scanning of address space by attackers or malware looking for vulnerable targets, backscatter from randomly spoofed source denial-of-service attacks and the automated spread of malware.

  CAIDA releases a number of datasets for use by external researchers. These datasets represent the major sources of telescope traffic:
  1. Two Days in November 2008 Dataset (Raw Traffic)
  2. Three Days of Conficker Dataset (Raw Traffic)
  3. Backscatter-2008 Dataset (DoS Attacks)
  4. Backscatter-2007 Dataset (DoS Attacks)
  5. Backscatter-2006 Dataset (DoS Attacks)
  6. Backscatter-2004-2005 Dataset (DoS Attacks)
  7. Backscatter-TOCS Dataset (DoS Attacks)
  8. Code Red Worms Dataset (Internet Worms)

9. Witty Worm Dataset (Internet Worms)

Note that the datasets include traffic in a great detail and consequently the total size of the datasets is few terabytes. Each release is divided in smaller one-hour chunks and anonymization for the source and destination is applied. Unprocessed, anonymized pcap traces corresponding to the dataset are available upon request and are expected to consume about ten times the capacity of the initially released set. From the technical side, the UCSD network telescope consists of a globally routed /8 network that carries almost no legitimate traffic. More specifically, all legitimate traffic is being filtered out so the resulting data provides a snapshot of anomalous 'background' traffic to 1/256th of all public IPv4 destination addresses on the Internet.

- **Format**
  The files are in xml format and the unprocessed data are released in pcap (packet capture) files.
- **Tools**
  A wide variety of processing and parsing tools are listed and provided in [53]. Many of the tools are open-source and the source code is released freely for developers to tweak. The tools are divided in the following categories: a) Geographic, b) Library, c) performance, d) Plotting and Data Curation, e) Topology, f) Workload.
- **Access**
  The datasets are private. The access is possible only after request and written authorization from CAIDA management. Unprocessed raw data are available after special request.

5.3 Honeypots

In this section, honeypot projects will be introduced and examined under the light of the NEMESYS objectives. A honeypot is a trap set to detect and deflect unauthorized use of information systems. Honeypots gather information about the motives and tactics of the attackers and would help greatly in analyzing the modus operandi of the attackers in mobile platforms. We are interested in all types of research honeypots (i.e., pure, high-interaction, low-interaction) as long as they capture extensive information.

5.3.1 Project Honey Pot

Project Honey Pot [54] provides honey pot scripts to put on web sites in order to catch email harvesters, blog and forum comment spammers, spam servers and dictionary attackers. The idea is that each email address harvested from a honey pot is linked to a specific spamtrap so that all spam received can also be linked to the harvester.

- **Content**
  Using a network of honey pots the project maintains an IP black list of hosts that should be blocked by web servers to prevent them from harvesting email

addresses, posting spam comments and to block other malicious activity (e.g., linking or uploading malware). The datasets are updated daily and 112,139,468 IPs and 134,956,689 spam traps are monitored. The available IP Blacklists are:

1. Directory of IPs
2. Lookup IP
3. Harvesters
4. Spam Servers
5. Dictionary Attackers
6. Comment Spammers

Malicious IP	Event	Total	First	Last
216.99.146.218 | CR	Bad Event	20,123	2012-11-11	2013-06-06
112.101.64.11 | C	Bad Event	23,033	2012-10-16	2013-06-06
112.101.64.130 | C	Bad Event	16,504	2013-01-11	2013-06-06
96.47.225.74 | HC	Bad Event	1,403,888	2012-06-16	2013-06-06
112.123.168.61 | C	Bad Event	29,555	2013-01-10	2013-06-06
198.2.213.33 | C	Bad Event	525	2013-06-06	2013-06-06

Figure 8 Harvester IPs database

The project offers its spam feed in real-time for anti-spam filter developers and companies assessing the reputation of IP addresses. The spam feed is extremely high quality with a very low rate of false positives. All message data, including envelope data, is preserved.

- **Format**
  The datasets are available in text/http format and in RSS/XML feed. Different formats may be available upon request.
- **Tools**
  Http:BL is a system that allows website administrators to take advantage of the data generated by Project Honey Pot and restrict access to suspicious IPs. It has been implemented on a number of different web servers, content management systems, blogging platforms, and forums. These systems query the http:BL servers for the IP addresses of the website visitors and restrict their access if they are found to be malicious. Such tools are: AbyssGuard, Drupal http:BL Module, HoneyPotJect, http:BL WordPress Plugin, HttpBlacklist-Component, HTTPBL - Project Honeypot Blocklists Plugin and IIS httpBL Module.
- **Access**
  In order to access the http:BL a free registration is mandatory and a request must be filed and approved. The NEMESYS project was granted access and a key was issued.

5.3.2 Honeynet Project

The *Honeynet Project* [55] is an international security research organization, dedicated to investigating the latest attacks and developing open source security tools to improve Internet security. The organization has various chapters around the world and their members contribute to the fight against malware by discovering new attacks and creating security tools used by businesses and government agencies all over the world.

- **Content**

  For data collection purposes and malicious activity monitoring the following honeypots are operated by the HoneyNet project:

  - *Dionaea* is a low-interaction honeypot that captures attack payloads and malware. Dionaea is meant to be a nepenthes successor, embedding python as scripting language, using libemu to detect shellcodes, supporting ipv6 and tls. CERTH set up an experimental installation in order to examine the operation of the Dionaea honeypot.

  - *Glastopf* is a low-interaction honeypot that emulates a vulnerable web server hosting many web pages and web applications with thousands of vulnerabilities. In a normal Glastopf installation attacks will pour in by the thousands daily.

  - *Google Hack Honeypot* is the reaction to a new type of malicious web traffic: search engine hackers. It is designed to provide reconnaissance against attackers that use search engines as a hacking tool.

  - *High Interaction Honeypot Analysis Toolkit (HIHAT)* transforms arbitrary PHP applications into web-based high-interaction Honeypots. Apart from the creating high-interaction honeypots, HIHAT comprises a graphical user interface which supports the process of monitoring the honeypot, analyzing the acquired data. Last, it generates an IP-based geographical mapping of the attack sources and generates extensive statistics.

  - *Honeywall CDROM* is a high-interaction tool for capturing, controlling and analyzing attacks. It features an architecture that allows researchers to deploy both low-interaction and high-interaction honeypots, but is designed primarily for high-interaction.
  - *Kippo* is a medium interaction SSH honeypot designed to log brute force attacks and, most importantly, the entire shell interaction performed by the attacker.

  - *Tracker* facilitates the identification of abnormal DNS activity. It will find domains that are resolving to a large number of IP's in a short period of time then continue to track those hostname->IP mappings until either the hostname nolonger responds or the user decides to stop tracking that hostname.

- **Format**

  The format of the dataset depends on the honeypot.

- **Tools**

  The Honeynet project releases numerous tools either for monitoring or for raw data processing. Some of these tools, which are relevant with the NEMESYS' objectives, are:

  - *APKinspector* is a tool to aid the reverse engineering of compiled Android packages and their DEX code.
  - *Capture-HPC* is a high-interaction client honeypot framework. It identifies malicious servers by interacting with potentially malicious servers using a dedicated virtual machine and observing its system for unauthorized state changes.
  - *Droidbox* is a dynamic analysis platform for android applications.
  - *HFlow2* is a data coalescing tool for honeynet/network analysis. It features coalescing data from snort, p0f, sebekd into a unified cross related data structure stored in a relational database.
  - *Honeyd* is a low-interaction honeypot used for capturing attacker activity.
  - *Honeymole* is used for honeypot farms. You deploy multiple sensors that redirect traffic to a centralized collection of honeypots.
  - *HoneySink* is a network sinkhole. Sinkholing is a technique that allows security researchers to monitor botnets and proactively deny access to the bots from the botnet herders.
  - *WebViz* is a GL visualization project implemented by Oguz as part of GSoc 2011. It allows to easily visualize attack data on a world globe.
  - *HoneyMap* provides a real-time visualization (Figure 9) of the cyber-attacks taking place across the globe. The map shows many attacks**.** Red dots represent an attack on a computer. Yellow dots represent honeypots, or systems set up to record incoming attacks. [56]

Figure 9: HoneyMap screenshot

- **Access**

Certain datasets have been released to be used in the HoneyNet challenges. The data collected by the project's chapters can be freely used upon approval of the chapter or the project management. Finally, it is possible for any volunteer to create and operate his own honeypot within the HoneyNet network.

5.3.3 University of Victoria - ISOT Lab Botnet Dataset

The *ISOT dataset* [57] is the combination of several existing publicly available malicious and non-malicious datasets.

- **Content**
  The recording of the network trace happened over a three month period, from October 2004 to January 2005 covering 22 subnets. The dataset contains trace data for a variety of network activities spanning from web and email to backup and streaming media.
  This experimental dataset contains both malicious and non-malicious traffic, and is the result of the merge of various datasets and traces. More specifically, two separate malicious traffic datasets were obtained from the French chapter of the HoneyNet project [57]. These datasets involved the Storm and Waledac botnets, respectively. Waledac is currently one of the most prevalent P2P botnets and is widely considered as the successor of the Storm botnet with a more decentralized communication protocol. Unlike Storm using overnet as a communication channel, Waledac utilizes HTTP communication and a fast-flux based DNS network exclusively. Additionally, to represent non-malicious, everyday usage traffic, two different datasets were also incorporated. One originating from the Traffic Lab at Ericsson Research in Hungary [58] and the other from the Lawrence Berkeley National Lab (LBNL) [59]. The Ericsson Lab dataset contains a large number of general traffic from a variety of applications, including HTTP web browsing behavior, World of Warcraft gaming packets, and packets from popular bittorrent clients such as Azureus. Finally, five trace datasets from the LBNL trace data were incorporated to provide additional non-malicious background traffic. The LBNL is a research institute with a medium-sized enterprise network.
  The combination of these traffic sets is used to simulate the behavior of a real world bot infected subnet while at the same time the existing well-labeled data can be used for training or evaluation purposes.
- **Format**
  The database archive includes a compressed pcap file with 10 GB original filesize.
- **Tools**
  No additional tools are provided with the dataset. However, the data are well-labeled and it is easy for the user to process the traces as needed.
- **Access**
  The dataset is public.

5.3.4 Nothink Honeypots & Malware Blacklist

The *Nothink* [60] datasets aim to provide information in the form of statistics and activity data of malware and spyware. The datasets are collected from various honeypot projects that Nothink operates.

- **Content**
  The datasets, lists and collections available offer a detailed insight in the correspondence between a malicious binary, its activities in the network (DNS, HTTP and IRC connections) and the sources of the attacks. Additionally, monitored activity from the honeypots is released as the following datasets: a) network traffic details, b) DNS network traffic, c) HTTP network traffic, d) IRC network traffic. Nothink provides three daily updated blacklists containing all IP addresses and domains related with Command & Control servers. These blacklists are:
    - Malware DNS network traffic (FQDN)
    - Malware HTTP network traffic (IP address)
    - Malware IRC network traffic (IP address)

  Finally, all malware is collected and analyzed using a sandbox. Apart from the analysis data and the malware signature lists a collection of malware binaries is available. This malware archive contains 7085 malicious binaries (ASCII, data, HTML, MS-DOS, PE32) gathered from Aug 2009 until Feb 2013.
- **Format**
  The datasets are available in xml format and the blacklists are in plaintext.
- **Tools**
  No tools are provided.
- **Access**
  All information can be used to perform analysis and filters in corporate and non-corporate environment.

5.4 Malicious Domains/URLs Block Lists

Blocklists (or Blacklists) are lists of domains or IP addresses that are reportedly known to have malicious behavior. This behavior varies from malware distribution, and website crawling to vulnerability scanning and command & control server [61] operation. Blocklists fall into two categories: global worst offender lists (GWOL) and local worst offender lists (LWOL). GWOL provide up-to-date information pulled from various sources that are known to be used to propagate malware and spyware. On the other hand, LWOL are built by an organization based on its firewalls and network activities.

5.4.1 ParetoLogic – Malware Blacklist

The *Malware Blacklist* [62] is a project that started back in 2007 by ParetoLogic. ParetoLogic creates security applications for home PC users. The Malware Blacklist houses one of the largest online repositories of malicious URLs and aims to help researchers in their understanding of the ever evolving threat landscape. The

Malware Blacklist database is consulted by Virus Total and URL Void online scans, and is a vital resource for security companies and professionals.

- **Contents**

  The database is updated daily and currently contains 203947 blacklisted URLs. For each URL the database provides the date and time of detection, the exact URL, its registrar (if not private), associated IP addresses, the ASN and hosting services provider. Additionally, the country of origin, a download link for the executable and a submission mechanism are available to the users. The blacklist generation procedure is shown in Figure 10.

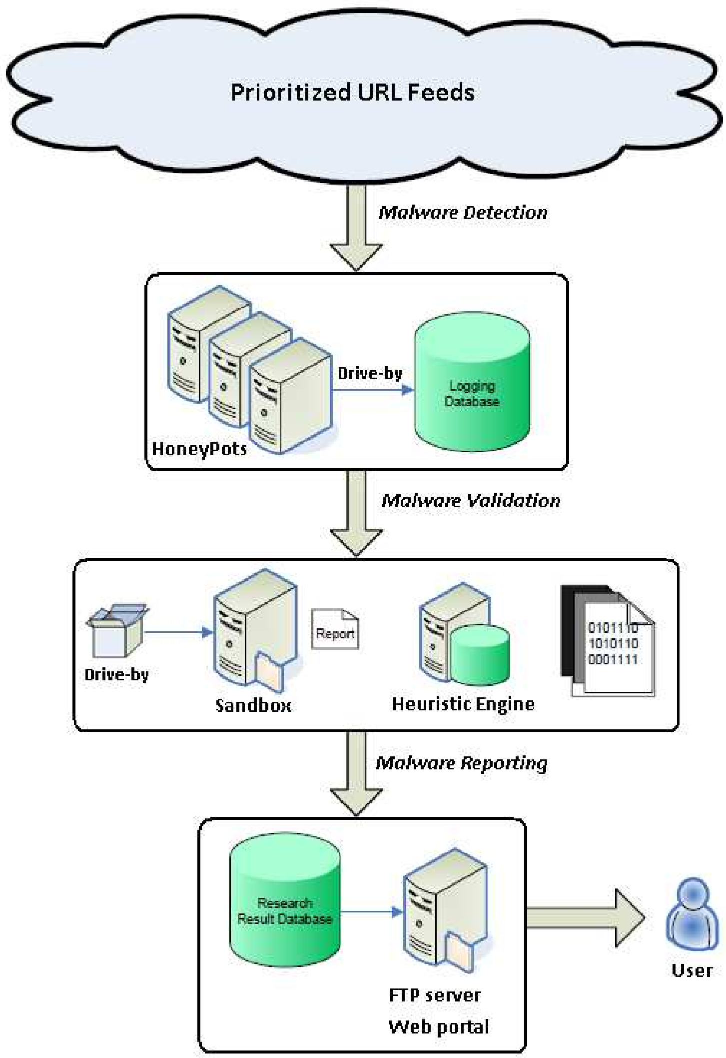


Figure 10: Malware Blacklist generation procedure

- **Format**

  The typical format is html. However, the list can also be obtained as RSS Feed.
- **Tools**

  A search engine is available for queries in the database. The search engine offers search based on a) URL b) IP c) Description d) Domain Extension e)

MD5 f) All. Additionally, a specific calendar period can be selected to investigate activity in a given period. Finally, Mwcrawler, a python script that parses malicious URL lists from well-known websites in order to automatically download the malicious code.

- **Access**
  The access to the database is public. However, access to the search functionality is available only to members approved by the project committee.

5.4.2 MalwareDomainList.com

Malware Domain List (MDL) [63] is a non-commercial community project.

- **Content**
  There are several versions of the list, which focus on different features (e.g., time, specific malware activity). The available lists are [64]:
    - The complete database
    - Recent database updates
    - URL updates
    - zbot/Zeus URL updates
    - All ZeuS URLs
    - Sites which are offline or have been cleaned
    - New database entries from previous day
    - New URLs from previous day
    - List of active IP addresses

- **Format**
  Different versions of the list are available in text/html, CSV and plaintext format. Special care has been taken so that all files are RFC complaint for easy parsing.
- **Tools**
  Mwcrawler, a python script that parses malicious URL lists from malware repositories in order to automatically download the malicious code.
- **Access**
  MalwareDomainList is a community project and the lists can be used for free by anyone.

5.4.3 Malware Patrol

Malware Patrol [65] is an automated and user contributed system that verifies URLs for the presence of Viruses, Trojans, Worms or any other software considered Malware. The system collects malicious URLs from multiple sources around the world, including security groups and companies, universities, spamptraps and personal contributors. Every URL is crawled and analyzed for the presence of malware.

- **Content**
  Currently, the database contains 8,510 blocked and 534,407 dangerous URLs. It is updated daily and URLs are monitored to ensure that block lists are fresh and trustworthy.

- **Format**
  Lists containing malicious addresses are available in several formats for many different applications and uses:
  1. BIND like DNS Servers
  2. ClamAV Virus DB (different versions)
  3. DansGuardian
  4. Firekeeper
  5. Hosts file (different versions)
  6. MailWasher block filters
  7. MaraDNS - CVS2
  8. MD5/SHA-1 hashes
  9. Microsoft DNS Server
  10. Mozilla cookie filtering
  11. Mozilla Firefox AdBlock
  12. Plain text
  13. Postfix MTA
  14. SmoothWall
  15. SpamAssassin
  16. Squid Web Proxy ACL
  17. SquidGuard
  18. Symantec Security for SMTP
  19. Symantec WebSecurity
  20. XML

- **Tools**
  No tools are provided.
- **Access**
  The use of the datasets is free and constitutes agreement to the Terms and Conditions. Use in commercial products or services require a special license.

5.4.4 Malc0de Blacklist and URLs

Malc0de [66] is a database of domains hosting malicious executables. Additionally, the *DNS Blackhole* [67] blacklist is released.

- **Content**
  The released datasets are automatically updated daily and populated with the last 30 days of malicious IP addresses. Each record to the database corresponds to one URL distributing malware. In order to enable linking between IPs and malware families, each entry includes: Date, Domain name, IP address, Country of Origin, ASN, Autonomous System Name, Md5 Signature and VirusTotal Report.
- **Format**
  The database is available in text/html format. Moreover, the dataset is available in the following formats:
  - BIND Format to be included in a DNS server for local networks.
  - Windows Format to be used in windows operating systems
  - Plaintext Format as simple IP blacklist for any other uses and easy parsing
- **Tools**

In the malc0de website there is a variety of perl scripts aiming to help researchers with parsing, lists resolution and conversion, malicious binaries submission and tracking. Some of these are:

- Search ThreatExpert From CLI
- Query Robtex.com to Reverse Lookup IP
- Submit to Anubis
- Convert Decimal IP to Dotted Quad
- Resolve List of Domain Names
- Create CSV out of Zeus Drop Server Log Files
- Wget through active Proxies
- Parse Zeus Tracker for Active Drop Sites
- Submit Binaries To VirusTotal (Python)
- Pastebin Scraper

Additionally, more online tools can be found in the website. These tools aim to cover simple encoding and validation needs: a) Format JavaScript C0de, Decode Base64, c) Decode Unescape. Another tool is Mwcrawler, a python script that parses malicious URL lists from certain websites in order to automatically download the malicious code. Among others, Mwcrawler supports downloading from Malc0de.

- **Access**

  The access to the database and any other tools and scripts is public and is not restricted by any means.

5.4.5 Malwr

Malwr [68] is a free malware analysis service and community launched in January 2011. Malwr is a free, independent and non-commercial service to the security community, independent or academic researchers with goal to facilitate everyone's daily work and give a contribution to the community. Members of the community can submit files to it and receive the results of a complete dynamic analysis back. It aims to serve as an alternative to other online analysis services which use closed and commercial technologies, often with intents to leverage people's data to own profit and with no real transparency on how the data is being used. The malware analysis is performed by the open source malware analysis tool Cuckoo Sandbox.

- **Content**

  The database contains 6344 analysis results and 23104 unique domains have been identified. The every new submission is automatically added to the project's website. Each submission is a separate entry containing the date and time of submission, an MD5 Hash, the file Name & Type and the antivirus results. The malware analysis results provide a quick overview, a static analysis report, a behavioral analysis report, a network analysis report and the dropped files.
- **Format**

  The database is in text/html format.
- **Tools**

Search functionality is available for all users. The users can query the database using one or more of the following criteria:

- Executable Signature (e.g., MD5, SHA1, SHA256, SHA512)
- File name pattern
- File type/format
- Open files matching the pattern
- Open registry keys matching the pattern
- Open mutexes matching the pattern
- Contact the specified domain
- Contact the specified IP address
- Search for Cuckoo Sandbox signatures
- Search using tags

Additionally, an API for queries to the database is supported.

- **Access**
  The access to the analyses is public. However, access to an API for queries to the database is permitted after approval.

5.4.6 Virus Tracker

Virus Tracker [69] is an automatically weekly updated banking trojan domain blocklist. Virus Tracker was created to track and monitor some of the largest botnets with a focus on financial malware botnets. The list is generated using domain sinkholing combined with other techniques. It is a very interesting blacklist as it correlates the criminals' modus operandi with the methods used to infect personal computers and mobile devices.

- **Content**
  The list contains domain names from more than 50 botnets and 60.000 Command & Control servers. The backend algorithms detect about 1.6 million unique infections daily and 10 million unique infections total, while the total number of infection records is 200 million. Furthermore, the project operates more than 300 sinkholes. The malware domain black list contains two columns one including the domain and a second for the botnet network name.
- **Format**
  The format of the list is plaintext.
- **Tools**
  No tools are provided.
- **Access**
  The list has switched from public to private and access for commercial uses is paid. The company needs to be contacted for licensing details.

5.4.7 hpHosts

The *hpHosts database* [70] aims to provide an additional layer of protection against access to ad, tracking and malicious websites by releasing "hosts" files for windows and linux operating systems. Additionally, "hpHosts Online" is available for blacklisted IP addresses and domain names lookup.

- **Content**
  Currently, 196.977 hosts are listed of which 194.262 are classified and 2.715 are not. The classification refers to the reasons for inclusion into hpHosts and is done in the following categories:
  - ATS - Ad/tracking servers
  - EMD - Engaged in malware distribution
  - EXP - Engaged in or alleged to be engaged in the exploitation of browser and OS vulnerabilities
  - FSA - Engaged in the selling or distribution of bogus or fraudulent applications
  - GRM - Engaged in astroturfing (also known as grass roots marketing)
  - HJK - Engaged in browser hijacking or other forms of hijacking
  - MMT - Engaged in the use of misleading marketing tactics
  - PSH - Engaged in Phishing
  - WRZ - Engaged in the selling, distribution or provision of warez

  Apart from these, for each domain name the host, current IP address, IP PTR, ASN and submission date are provided. New entries are added daily to the database and once a month a new host file is released.
- **Format**
  The format of this database is quite unique, meaning that it is provided in Linux and windows host file format. Additionally, an RSS feed (XML) is available [71].
- **Tools**
  The web interface provides a simple search form for the end users to search for specific IP addresses. Furthermore the RSS feed supports classification filters in the form*: http://hosts-file.net/rss.asp?class=<classification_code>*
  In the "downloads" section a variety of tools and utilities can be found:
  - *DiamondCS:* MD5 utility for verifying MD5 hash.
  - *HostsMan: U*tility for managing the HOSTS file, with automatic updates and a built in server to enhance HOSTS file usage.
  - *HostXpert*: Utility for managing the HOSTS file, with automatic updates.
  - *Funkytoad:* Provide a server to be used in conjunction with HostsXpert.

- **Access**
  The service is free to use, however, any kind of automated use is strictly forbidden without express permission.

5.4.8 Cyber Crime Tracker

Cybercrime Tracker [72] is a database which stores the domain names and the IP addresses used by Control & Command servers of wide-spread botnets. This botnets use very often malware to infect mobile devices (e.g., banking botnets).

- **Content**

Currently, the database contains 1600 entries and is updated regularly. For each C&C server it provides the date of detection, URL, the IP address if detected and the type of the botnet.

- **Format**
  The database is available in RSS/XML format and in text/html, as well.
- **Tools**
  Some tools are provided but they are not particularly interesting for the NEMESYS purposes.
- **Access**
  Access is public.

5.4.9 ScumWare

ScumWare [73] is an online service that lists URLs and domain names that distribute malicious software. For the detection, a combination of user reports and automatic tools are used. All pages are scanned with multiple antivirus programs and analyzed by dedicated tools.

- **Content**
  The scumware database analyzed 1.587.796 threats in 2013 and it is growing fast. A useful feature of this database is that it supports categorization based either on Top Level Domains (TLD) and geographical location (GEO). For each malicious or infected IP address the database provides:
    - Date and time of detection
    - Full URL
    - MD5
    - GEO
    - Threat Details
    - Download Link (in some cases only)

- **Format**
  The public interface provides the data in text/html format.
- **Tools**
  The web interface provides search functionality, which enables database queries for specific MD5 hashes, IP addresses, hostnames or URLs.
- **Access**
  The access to the database is public and requires filling in a captcha. Use of automatic bots, crawlers, mirroring tools is prohibited.

5.4.10 VX Vault

VX Vault [74] maintains a malicious URL blacklist and a malware repository which includes about 24000 lemmas. Such repositories are ideal for correlating certain IP addresses with specific malware families.

- **Content**
  For each malicious URL that spreads malware the date, URL of the malware, MD5 hash, IP address, tools and the malware sample are provided. All malware samples collected are listed and the following information is made public:

- FileName
- Size
- MD5
- SHA-1
- Link
- Submission Date
- Detailed analysis of the executable

- **Format**
  The blacklist and the malware analysis details are in text/html format. However, the “latest 100” malicious URLs list [75] is in plaintext format.
- **Tools**
  The website provides an MD5 search engine for database queries. Additionally, Mwcrawler is compatible with VX Vault and enables the automatic downloading of malicious code.
- **Access**
  The list is open and public.

5.4.11 AlienVault Labs – IP Reputation Portal

The *IP Reputation Portal* [76] is an open source project which displays lists with the most wanted IPs with malicious activity worldwide.

- **Content**
  For all entries, the information associated with a specific IP address (i.e., country, city, risk, date) and type of malicious activity are provided. In case of malware distribution, the exact location of the malware, its MD5 signature and name are given. The entries of the list are updated hourly.
- **Format**
  The list can be obtained in the following formats: Generic, OSSIM, Snort Reputation, Iptables, Squid, Unix (hosts.deny).
- **Tools**
  A real-time attacks world-map is available in [77].
- **Access**
  The access to the reputation list is public.

5.4.12 Spam Domain Blacklist - Spam-IP

Spam-IP [78] maintains a spam blacklist that is powered by people from all over the world.

- **Content**
  The blacklist contains 714952 entries and is updated hourly. It contains the following details for each reported incident: unique ID, IP address, User, Email and Date & time.
- **Format**
  The spam IP blacklist is in CSV format.
- **Tools**
  No specific tools are given.

- **Access**
  The blacklist is released publicly and used in filtering application or hardware.

5.4.13 Wikimedia Spam Blacklist

Wikimedia Spam Blacklist [79] is a list of domain names which are blocked from the Wikimedia boards.

- **Content**
  The blacklist contains keywords and domain names that are reported to produce spam. The content of the list is structured as REGEXes and is of interest for us because it can be applied to identify spam SMS messages and filter URL requests.
- **Format**
  The list is available in plaintext format.
- **Tools**
  No tools are provided.
- **Access**
  Access to the spam blacklist is public.

5.4.14 Phishing Domain Blacklist - PhishTank

PhishTank [80] is a phishing URI list, which works by utilizing phishing reports from the community. Each time a URI is reported, it has to be verified before it is listed as pointing to a phishing web site.

- **Content**
  Apart from the URL, additional information are given such as the host network, the WHOIS information, whether the phishing has been verified or not and if so when, whether the site is still active or not, and the legitimate web site being targeted. In many cases, a screenshot of the phishing web page is available from the PhishTank web site only. The list is updated hourly.
- **Format**
  It can be queried or fetched via XML, CVS, PHP and JSON.
- **Tools**
  No additional tools are given.
- **Access**
  Access to the list is public. A registration is required to lift the limitation on the number of queries.

5.5 Evaluation & Comparison

Before starting with the evaluation we should note that most of the information sources do not aim to exclusively provide information solely on mobile traffic. Nevertheless, this is not a major issue, since most attackers coordinate and lunch their attacks from the internet. In order to efficiently compare and evaluate the information sources outlined above we designed Table 4. The table contains features that vary between the information sources and should be considered when choosing datasets for the data collection infrastructure (T3.3).

The DNS monitoring projects and all kinds of honeypots are considered valuable sources for NEMESYS. This is because, as seen in Figure 11, since 2011 fully DNS-based botnets started spreading and fast-flux techniques became popular. In our evaluation we found out that such sources offer high quality datasets on malicious and normal user traffic. From all collected sources, the DNSDB database, in comparison with EXPOSURE and CAIDA Network Telescope appears to be the easiest to use, mainly because it supports queries, automated access and is constantly updated with new entries. On the other hand, the CAIDA Network Telescope releases useful datasets CAIDA Network Telescope provides very large datasets, which are collections of traffic packages and are suitable for automated malware spread. It is quite hard to accurately characterize CAIDA Network Telescope as DNS monitoring system or honeypot, since the data collection is made using a non-standard technique. The drawback of CAIDA is its main benefit as well, the amount of data is overwhelming and thus, pre-processing is required prior to use.

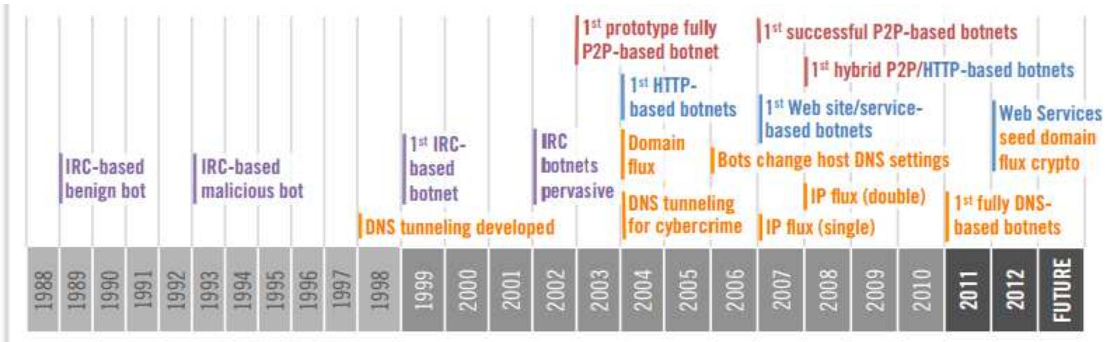


Figure 11: Botnet modus operandi evolution [81]

From the honeypot information sources we found the ISOT Lab Botnet dataset to be a valuable resource mainly because it is a combination of different datasets and contains a wide range of malicious and non-malicious incidents. That been said, the ISOT dataset is a careful merge of datasets originating DNS monitoring and honeypot sources but its pcap file type requires a special analysis in order to be useful for the NEMESYS purposes. Another source is NoThink, which we consider as the most complete collection of datasets in this category. It provides four types of datasets and various blacklists and repositories, with moderate size. These four datasets include network traffic monitoring data and they are easy to parse in order to extract useful features. In addition, NEMESYS can utilize three blacklists that are daily generated by the NoThink honeypots. The remaining honey pot projects are also interesting for NEMESYS and their use should be examined in a per case basis.

Finally, the malicious domain & IP address block lists are a different kind of sources that can be used in conjunction with other types of datasets. Sources of this kind release datasets that list IP addresses and/or domain names suspected or identified to be malicious. In most cases the malicious behavior is associated with malware but not always (e.g., phishing attempts). In this specific category of information sources the datasets can be easily merged to one. From the examined datasets we found the Malc0de and the VX Vault well-suited for NEMESYS, since both datasets are constantly updated, rich in content and all features can be easily extracted.

Furthermore, VirusTracker is a unique information source that NEMESYS was given permission to access and includes the IP addresses of numerous Command & Control (C&C) botnet servers. VirusTracker is not complete on its own but it can be used to further enrich other datasets. The remaining information sources are also reliable and beneficial for use in NEMESYS.

Table 4: Internet Activity Monitoring information sources

ID / Name	*DNS*	*Honeypot*	*Blacklist*	*Traffic*	*Malware*	*Samples*	*Signatures*	*Updates*	*EPF*	*Access*
ISC – DNSDB	✓	-	✓	✓	-	-	-	✓	✓	Granted
Iseclab – EXPOSURE BlackList	✓	-	✓	✓	-	-	-	✓	-	Public
CAIDA - UCSD Network Telescope	✓	✓*	-	✓	-	-	-	-	✓	Granted
Project Honey Pot	-	✓	✓	-	✓	-	-	✓	✓	Granted
The Honeynet Project	-	✓	N/A	✓	✓	N/A	N/A	✓	N/A	Pending
ISOT Lab Botnet Dataset	✓	✓	-	✓	-	-	-	-	✓	Public
Nothink Honeypots & Malware Blacklist	✓	✓	✓	✓	✓	✓	✓	✓	✓	Public
ParetoLogic - Malware Blacklist	-	✓	✓	-	✓	✓	-	✓	✓	Public
MalwareDomainList.com	-	-	✓	-	✓	-	✓	✓	✓	Public
Malware Patrol	-	-	✓	-	✓	-	✓	✓	✓	Public
Malc0de Blacklist and URLs	-	-	✓	-	✓	✓	✓	✓	✓	Public
Malwr	-	-	✓	-	✓	-	✓	✓	-	Public
Virus Tracker	-	-	✓	-	✓	-	-	✓	✓	Granted
hpHosts	-	-	✓	-	✓	-	-	✓	✓	Public
Cyber Crime Tracker	-	-	✓	-	✓	-	-	✓	✓	Public
ScumWare	-	-	✓	-	✓	-	✓	✓	-	Public
VX Vault	-	-	✓	-	✓	✓	✓	✓	✓	Public
AlienVault Labs – IP Reputation Portal	-	-	✓	-	✓	-	✓	✓	✓	Public
Spam Domain Blacklist - Spam-IP	-	-	✓	-	-	-	-	✓	✓	Public
Wikimedia Spam Blacklist	-	-	✓	-	-	-	-	✓	✓	Public
Phishing Domain Blacklist - PhishTank	-	-	✓	-	-	-	-	✓	✓	Public

- ID/Name: Name of the information source or dataset.
- DNS: The datasets were collected using DNS monitoring.
- Honeypot: The datasets were collected using honeypots.
- Blacklist: The source provides and IP/URL/Domain blacklist.
- Traffic: The information source provides network traffic data.
- Malware: The datasets contain information on malware traffic or activity.
- Samples: Malware samples are provided (not necessarily for mobile environments).
- Signatures: Malware/Vulnerabilities signatures are supported.
- Updates: If the datasets are periodically updated.
- EPF: The datasets are in an Easy to Parse Format.
- Access: Access policy as defined by the source and the NEMESYS project's permissions.

* The UCSD network telescope is a globally routed /8 network (approximately 1/256th of all IPv4 Internet addresses) that carries almost no legitimate traffic because there are few provider-allocated IP addresses in this prefix. Although it is marked as a honeypot

6 Analysis and Conclusion

Throughout this deliverable, we have listed and documented all relevant information sources that will be valuable inputs for the NEMESYS project.

As shown in Chapter 2, there are sufficient information sources that release datasets produced using mobile device monitoring. In particular, Lausanne Data Collection, Reality Mining and NODOBO are the most important information sources and they provide reliable and rich datasets. Moreover, a number of Bluetooth and Wifi datasets (e.g., SIGCOMM 2009) are released and it would be very interesting to explore malware spread techniques based on the traces they provide.

In chapter 3, we examined information sources that provide datasets collected using monitoring from a central network node. A shortage of information sources becomes easily apparent, mainly because of the monitoring complexity and the strict privacy policies the networks providers apply. Despite this, both datasets (i.e., IEEE VAST 2008, Orange D4D) were evaluated as very useful for NEMESYS and especially the private dataset from the D4D challenge which contains call detail records from 5 million users.

In Chapter 4, we examined various databases, repositories and analysis tools. More specifically, many malware repositories, databases, encyclopedias and analysis tools were presented and discussed. VirusShare, Android Malware Genome project, VirusTotal and Anubis are among others the most significant information sources from this category. On the other hand, vulnerabilities also provide a valuable source of information for the analysis of malware and attackers techniques. The most prevalent sources are NIST and OSVDB because they offer a wide variety of vulnerabilities in an easy to use format.

Chapter 5 presents DNS monitoring projects, honeypots and blocklists regarding malicious activity and malware. From those we found DNSDB from ISC, ISOT Botnet dataset and Nothink to be very close to what is needed for NEMESYS. Additionally, VX Vault and Malc0de should be considered the NEMESYS' primary sources for malware samples and signatures.

From all the above, we conclude that the information sources we introduced will effectively cover the needs of all NEMESYS' tasks that require such input. As planned within NEMESYS, correlating all those heterogeneous information sources should help us to efficiently identify, analyze and counteract against the attackers' modus operandi. Last but not least, it is important to note that even though we made every effort to examine and list all available information sources new projects are constantly initiated and thus it would be beneficial to stay alert for new sources.

7 References

[1] "NODOBO," [Online]. Available: http://nodobo.com/release.html.

[2] "Nokia Mobile Data Challenge," [Online]. Available: http://research.nokia.com/mdc.

[3] "Reality Commons Media," [Online]. Available: http://realitycommons.media.mit.edu/realitymining1.html.

[4] A. O. R. P. H. T. Mika Raento, "ContextPhone - A prototyping platform for context-aware mobile applications," *IEEE Pervasive Computing,* vol. 4, no. 2, pp. 51-59, 2005.

[5] "SMS Corpus," [Online]. Available: http://wing.comp.nus.edu.sg/SMSCorpus/.

[6] "Michal Ficek dataset," [Online]. Available: http://crawdad.cs.dartmouth.edu/meta.php?name=ctu/personal.

[7] "Mall Research dataset," [Online]. Available: http://www.mrl.nottingham.ac.uk/~azg/project/description.html.

[8] "Android Bluetooth tracing experiment," [Online]. Available: http://crawdad.cs.dartmouth.edu/meta.php?name=upb/mobility2011.

[9] "SIGCOMM 2009," [Online]. Available: http://uk.crawdad.org/meta.php?name=thlab/sigcomm2009.

[10] "Bluetooth Worms Investigation," [Online]. Available: http://www.cs.toronto.edu/~stefan/downloads/.

[11] "Data for Development (D4D) Challenge," [Online]. Available: http://www.d4d.orange.com/.

[12] "Geofast," [Online]. Available: www.geofast.net.

[13] "IEEE VAST 2008 challenge," [Online]. Available: http://www.cs.umd.edu/hcil/VASTchallenge08/.

[14] "Mobile Malware Survey," [Online]. Available: http://www.cs.berkeley.edu/~afelt/malware.html.

[15] A. F. e. al., "Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices," in *A survey of mobile malware in the wild*, New York, 2011.

[16] "F-Secure Corporation," [Online]. Available: www.f-secure.org.

[17] "Mobile Security List," [Online]. Available: http://www.f-secure.com/en/web/labs_global/mobile-security.

[18] "Android Malware Genome Project," [Online]. Available: http://www.malgenomeproject.org/.

[19] "Malgenome Project Policy," [Online]. Available: http://www.malgenomeproject.org/policy.html.

[20] "Panda List of viruses," [Online]. Available: http://www.cloudantivirus.com/en/listofviruses/.

[21] "Panda List of mobile threats," [Online]. Available: http://www.pandasecurity.com/homeusers/security-info/types-malware/mobile-threats/mobile-historical-record.htm.

[22] "securelist.com," [Online]. Available: http://www.securelist.com/en/descriptions.

[23] "Thread Center," [Online]. Available: http://www.sophos.com/en-us/threat-center/threat-analyses/viruses-and-spyware.aspx.

[24] "Anubis," [Online]. Available: http://anubis.iseclab.org/.

[25] "isecLAB," [Online]. Available: http://www.iseclab.org/.

[26] "Anubis About," [Online]. Available: http://anubis.iseclab.org/?action=about.

[27] "CooperDroid," [Online]. Available: http://copperdroid.isg.rhul.ac.uk/.

[28] "Open Malware," [Online]. Available: http://oc.gtisc.gatech.edu:8080/.

[29] "WildList Organization," [Online]. Available: http://www.wildlist.org/.

[30] "The WildList," [Online]. Available: http://www.wildlist.org/WildList/.

[31] "Microsoft Malware Encyclopedia," [Online]. Available: https://www.microsoft.com/security/portal/threat/Encyclopedia/Browse.aspx.

[32] "Tools Collection," [Online]. Available: https://www.microsoft.com/security/portal/shared/resources.aspx.

[33] "Fortiguard Encyclopedia," [Online]. Available: http://www.fortiguard.com/encyclopedia/.

[34] "Threats for mobile devices," [Online]. Available: http://www.fortiguard.com/antivirus/mobile_threats.html.

[35] "VirusShare," [Online]. Available: http://virusshare.com.

[36] "VirusShare Hashes," [Online]. Available: http://virusshare.com/hashes.4n6.

[37] "Malware.lu," [Online]. Available: malware.lu .

[38] "Contagio mobile mini-dump," [Online]. Available: http://contagiominidump.blogspot.gr/.

[39] "Contagio Dump," [Online]. Available: http://contagiodump.blogspot.gr/.

[40] VirusTotal. [Online]. Available: https://www.virustotal.com/en/.

[41] "Security Response Vulnerabilities," [Online]. Available: http://www.symantec.com/security_response/landing/azlisting.jsp.

[42] "NVD," [Online]. Available: http://nvd.nist.gov/.

[43] "NVD About," [Online]. Available: http://nvd.nist.gov/about.cfm.

[44] "NIST Vulnerabilities for the Android Platform," [Online]. Available: http://web.nvd.nist.gov/view/vuln/search-results?query=android&search_type=all&cves=on.

[45] "SCAP," [Online]. Available: http://scap.nist.gov/.

[46] "OSVDB," [Online]. Available: http://www.osvdb.org.

[47] "IBM X-Force Vulnerability database," [Online]. Available: http://webapp.iss.net/Search.do?searchType=vuln&keyword=.

[48] "ISC - DNSDB," [Online]. Available: https://dnsdb.isc.org/.

[49] "pDNS sensors," [Online]. Available: https://security.isc.org/Passive_DNS_Sensor_FAQ/.

[50] "EXPOSURE," [Online]. Available: http://exposure.iseclab.org.

[51] "WOMBAT," [Online]. Available: http://www.wombat-project.eu/.

[52] "UCSD Network Telescope," [Online]. Available: http://www.caida.org/projects/network_telescope/.

[53] "CAIDA Tools," [Online]. Available: http://www.caida.org/tools/.

[54] "Project Honey Pot," [Online]. Available: http://www.projecthoneypot.org/.

[55] "Honeynet Project," [Online]. Available: http://www.honeynet.org/.

[56] "HoneyNet Map," [Online]. Available: http://map.honeycloud.net.

[57] "ISOT dataset," [Online]. Available: http://www.uvic.ca/engineering/ece/isot/datasets/index.php.

[58] G. Szab, "Proceedings of the 9th international conference on Passive and active network measurement," in *On the validation of traffic classification algorithms*, Berlin, 2008.

[59] "Lawrence Berkeley National Lab (LBNL)," [Online]. Available: http://www.icir.org/enterprise-tracing/.

[60] "Nothink," [Online]. Available: http://www.nothink.org/honeypots.php.

[61] "VirusBTN," [Online]. Available:

http://www.virusbtn.com/resources/glossary/command_and_control.xml.

[62] "Malware Blacklist," [Online]. Available: http://www.malwareblacklist.com/showMDL.php.

[63] "Malware Domain List," [Online]. Available: http://www.malwaredomainlist.com/mdl.php.

[64] "Available Lists," [Online]. Available: http://www.malwaredomainlist.com/forums/index.php?topic=3270.0.

[65] "Malware Patrol," [Online]. Available: http://www.malware.com.br/lists.shtml.

[66] "Malc0de Database," [Online]. Available: http://www.malc0de.com/database/.

[67] "Malc0de Blacklist," [Online]. Available: http://www.malc0de.com/bl.

[68] "Malwr," [Online]. Available: https://malwr.com/.

[69] "VirusTracker," [Online]. Available: http://virustracker.info/.

[70] "hpHosts database," [Online]. Available: http://hosts-file.net/.

[71] "HP Hosts Feed," [Online]. Available: http://hosts-file.net/rss.asp.

[72] "Cyber Tracker," [Online]. Available: http://cybercrime-tracker.net/index.php.

[73] "Scumware," [Online]. Available: http://www.scumware.org/.

[74] "VX Vault," [Online]. Available: http://vxvault.siri-urz.net/ViriList.php.

[75] "VX Vault malicious URLs list," [Online]. Available: http://vxvault.siri-urz.net/URL_List.php.

[76] "The IP Reputation Portal," [Online]. Available: https://reputation.alienvault.com/reputation.data.

[77] "The IP Reputation Portal realtime attacks map," [Online]. Available: http://labs.alienvault.com/labs/index.php/projects/open-source-ip-reputation-portal/real-time-map/.

[78] "Spam-IP," [Online]. Available: http://spam-ip.com/spam-blacklist.php.

[79] "WikiMedia Spam Blacklist," [Online]. Available: http://meta.wikimedia.org/wiki/Spam_blacklist.

[80] "PhishTank," [Online]. Available: http://www.phishtank.com/.

[81] "OpenDNS - The Role of DNS in Botnet Command & Control," [Online]. Available: http://info.opendns.com/rs/opendns/images/OpenDNS_SecurityWhitepaper-DNSRoleInBotnets.pdf.

[82] "Android Versions Distribution," [Online]. Available:

http://developer.android.com/about/dashboards/index.html.

8 Appendix I

Table 5: Comparison Table for all Information Sources

ID / Name	Type	R/S	Updates	Format	Year	Access
NODOBO	Device	R	-	CSV	2011	Public
Reality Commons - Reality Mining Dataset	Device	R	-	MAT	2005	Public
SMS Corpus	Device	R	✓	XML,SQL	2013	Public
Michal Ficek Dataset	Device	R	-	CSV	2011	Public
Nottingham Mall Research	Device	R	-	EPS	2007	Public
Android Bluetooth Tracing Experiment	Bluetooth	R	-	Plaintext	2011	Public
SIGCOMM 2009	Device	R	-	CSV	2009	Public
Bluetooth Worms Investigation	Bluetooth	R	-	TSV	2006	Public
Lausanne Data Collection – NOKIA	Device	R	-	N/A	2013	Pending
Orange “Data for Development” Challenge	OMC	R	-	N/A	2012	Pending
IEEE VAST 2008 - Challenge Datasets	OMC	S	-	CSV	2008	Public
Berkeley files - Mobile Malware Survey	DB	R	-	Spreadsheet	2011	Public
F-Secure - Mobile Security List	DB	R	✓	XML	2013	Public
Android Malware Genome Project	Repo	R	-	ZIP	2011	Granted
Panda Security - List of Viruses & Panda Mobile	DB	R	✓	HTML	2013	Public
Kaspersky Lab – SecureList	DB	R	✓	HTML	2013	Public
SOPHOS - Threat Center	DB	R	✓	HTML	2013	Public
Anubis - Analyzing Unknown Binaries	Tools	R	✓	XML	2013	Public
Georgia Institute of Tech - Open Malware	DB/Tool	R	✓	Plaintext	2013	Public
WildList - Virus Bulletin	DB	R	✓	Plaintext	2013	Public
Microsoft Malware Encyclopedia	DB	R	✓	HTML	2013	Public
FortiNet - Fortiguard Encyclopedia	DB	R	✓	HTML	2013	Public
VirusShare	DB/Repo	R	✓	Plaintext/ZIP	2013	Granted
Malware.lu	DB/Repo	R	✓	Plaintext	2013	Granted
Contagio Mini Dump	DB/Repo	R	✓	ZIP	2013	Public
VirusTotal	Tool	R	✓	API	2013	Granted
Symantec – Security Response Vulnerabilities	DB	R	✓	XML	2013	Public
NIST - National Vulnerability Database	DB	R	✓	SCAP	2013	Public
Open Source Vulnerability Database	DB	R	✓	API	2013	Public

X-Force Vulnerability Search	DB	R	✓	HTML	2013	Public
ISC – DNSDB	DNS / BL	R	✓	HTML	2013	Public
Iseclab – EXPOSURE BlackList	DNS/BL	R	✓	Plaintext	2013	Public
CAIDA - UCSD Network Telescope	DNS	R	✓	XML	2013	Public
Project Honey Pot	BL/Honeypot	R	✓	XML	2013	Granted
The Honeynet Project	Honeypot	R	✓	N/A	2013	Public
ISOT Lab Botnet Dataset	DNS/Honeypot	R	-	PCAP	2013	Public
Nothink Honeypots & Blacklist	DNS/Honeypot	R	✓	XML	2013	Public
ParetoLogic - Malware Blacklist	Honeypot/BL	R	✓	RSS	2013	Public
MalwareDomainList.com	BL	R	✓	CSV etc	2013	Public
Malware Patrol	BL	R	✓	XML etc	2013	Public
Malc0de Blacklist and URLs	BL	R	✓	BIND etc	2013	Public
Malwr	BL/Tool	R	✓	HTML	2013	Public
Virus Tracker	Tool	R	✓	Plaintext	2013	Granted
hpHosts	BL	R	✓	HOSTS	2013	Public
Cyber Crime Tracker	BL	R	✓	XML	2013	Public
ScumWare	BL	R	✓	HTML	2013	Public
VX Vault	BL/Repo	R	✓	HTML	2013	Public
AlienVault Labs – IP Reputation Portal	BL	R	✓	HOSTS etc	2013	Public
Spam Domain Blacklist - Spam-IP	BL	R	✓	CSV	2013	Public
Wikimedia Spam Blacklist	BL	R	✓	regex	2013	Public
Phishing Domain BL – PhishTank	BL	R	✓	XML, CVS	2013	Public

- ID/Name: Name of the information source or dataset.
- Type: Monitoring type and the information source category.
- R/S: Real or synthetic data.
- Updates: If the datasets are periodically updated.
- Format: File type of the released datasets.
- Year: Year of release or last update.
- Access: Access policy as defined by the source and the NEMESYS project’s permissions.

Version	Codename	API	Distribution
1.6	Donut	4	0.1%
2.1	Eclair	7	1.5%
2.2	Froyo	8	3.2%
2.3 - 2.3.2	Gingerbread	9	0.1%
2.3.3 - 2.3.7		10	36.4%
3.2	Honeycomb	13	0.1%
4.0.3 - 4.0.4	Ice Cream Sandwich	15	25.6%
4.1.x	Jelly Bean	16	29.0%
4.2.x		17	4.0%

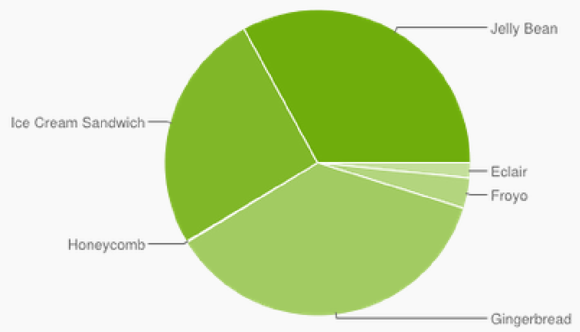


Figure 12: Android Version Usage. Data collected during a 14-day period ending on June 3, 2013. **[82]**