Marie Sklodowska Curie,

Research and Innovation Staff Exchange (RISE)

Ref. Ares(2016)7203956 - 31/12/2016

European Commission

Horizon 2020
European Union funding
for Research & Innovation

ENhancing seCurity and privAcy in the Social wEb: a user-centered approach for the protection of minors

ENCASE
ENhancing seCurity and privAcy in the Social wEb

WP2 – Requirements and System Architecture

Deliverable D2.2 “System Requirements and Software Architecture”

Editor(s):	Michael Sirivianos (CUT)
Author(s):	Michael Sirivianos, Kostantinos Papadamou, Antigoni Parmaxi, Panagiotis Zaphiris (CUT), Rig Das (ROMA3), Pantelis Nicolaou, George Sielis (CYRIC), Thanassis Lekkas, Demetris Soukaras (INNO), Antonia Gogoglou, Despoina Chatzakou (AUTH), Emiliano De Cristofaro, Gianluca Stringhini (UCL), Jeremy Blackburn (TID)
Dissemination Level:	Public
Nature:	Report
Version:	2.3

ENCASE Project Profile

Contract Number	691025
Acronym	ENCASE
Title	ENhancing seCurity and privacy in the Social wEb: a user-centered approach for the protection of minors
Start Date	Jan 1st, 2016
Duration	48 Months

Partners

	Cyprus University of Technology	Cyprus
	Telefonica Investigacion Y Desarrollo SA	Spain
	University College London	United Kingdom
	Cyprus Research and Innovation Center, Ltd	Cyprus
	SignalGenerix Ltd	Cyprus
	Aristotle University	Greece
	Innovators, AE	Greece
	Universita Degli Studi, Roma Tre	Italy

Document History

AUTHORS

(CUT)	Michael Sirivianos, Kostantinos Papadamou, Antigoni Parmaxi, Panagiotis Zaphiris
(ROMA3)	Rig Das
(CYRIC)	Pantelis Nikolaou, George Sielis
(INNO)	Thanassis Lekkas, Demetris Soukaras
(AUTH)	Antonia Gogoglou, Despoina Chatzakou
(UCL)	Emiliano De Cristofaro, Gianluca Stringhini
(TID)	Jeremy Blackburn

VERSIONS

Version	Date	Author	Remarks
0.1	15.11.2016	CUT	Initial Table of Contents
0.2	18.11.2016	CUT, INNO	Update of the related web-based tools
0.3	15.11.2016	CUT	Update of initial literature review
0.4	20.11.2016	CUT, CYRIC	Revised use cases
0.5	28.11.2016	CUT	Addition and update of user stories
0.6	30.11.2016	ROMA3	Added dataset description for sexually abusive behaviour detection
0.7	02.12.2016	AUTH, TID	Added dataset description for abusive behaviour and hate speech detection
0.8	02.12.2016	UCL	Added dataset description for cyber bullying and hate speech detection
0.9	04.12.2016	AUTH	Added dataset description for early malicious activity detection
1.0	06.12.2016	CUT, CYRIC, ROMA3	Added the reference architecture description
1.1	08.12.2016	CYRIC, ROMA3	Added established architectures section
1.2	12.12.2016	CUT, CYRIC, ROMA3, INNO	Added the description of the architectural components
1.3	13.12.2016	ROMA3, CYRIC	Added architectural components diagrams
1.4	14.12.2016	INNO	Added infrastructure design
1.5	15.12.2016	CUT	Updated architectural components diagrams
1.6	17.12.2016	CUT, CYRIC	Added Front-end technical requirements
1.7	19.12.2016	CUT	Added Web-proxy server

			technical requirements
1.8	21.12.2016	CUT	Added Middleware technical requirements
1.9	23.12.2016	CUT	Added Back-end technical requirements
2.0	24.12.2016	CUT, CYRIC	Revision of use cases and technical requirements
2.1	26.06.2016	CUT, CYRIC, ROMA3	Comments and corrections
2.2	29.06.2016	CUT	Document layout and format check
2.3	30.06.2016	CUT, CYRIC, ROMA3	Proof reading – Final version

Executive Summary

The overall aim of the ENCASE project is to leverage the latest advances in usable security and privacy of minors (age 10-18) in order to design and implement a user-centric architecture for the protection of minors from malicious actors in Online Social Networks (OSNs). This deliverable, “System Requirements and Software Architecture”, is a revision of D2.1, which was an initial specification of the system’s requirements and the software architecture.

In order to identity the magnitude of the problem, we have also surveyed the existing security and privacy enhancing web-based tools and performed a research state-of-art on cyber security risks and on security in OSNs. Moreover, we further investigated the problem by collecting data from various OSNs and through our own measurements we intended to verify the results of the literature review provided in D2.1.

According to the use cases requirements engineering methodology that ENCASE has opted for, the user stories are the next step. This deliverable provides more sophisticated use cases and the usage scenarios has been transformed into user stories - first person narrative stories - that essentially provide another level of detail, addressing each user role individually and from a first person perspective. The benefit using this approach is that enables you to produce accurate technical requirements without taking into account the limitations imposed by the technology at hand.

Based on the aforementioned we were in position to design the reference architecture of the ENCASE ecosystem and to define and describe the various components that consist this architecture and the interactions between them. This architecture comprises three browser add-ons, a web-proxy service that in collaboration with a middleware service will be responsible to detect malicious behaviour, fake identities and activity, and sensitive content in OSNs based on sophisticated machine learning detection rules generated by a data analytics software stack, which is the back-end of our architecture.

Lastly, this document relies on the above efforts and mainly in the user stories in order to produce a list of specific technical requirements. The requirements are classified according to the four components and services that will together form the ENCASE platform. These are:

1. Front-End
2. Web-proxy Server
3. Middleware Service
4. Data Analytics Software Stack (Back-End)

The requirements description for each component is then subdivided into Functional requirements, Security and privacy requirements, and Operational requirements.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1.Purpose of the document

The main purpose of this document is to define the system technical requirements and the reference Architecture of the ENCASE ecosystem to be implemented. First, an updated survey of all the existing security and privacy enhancing web-based tools along with a survey of the research state-of-the-art, defined in D2.1, are provided. Next, following the requirements engineering methodology that was adapted for ENCASE and was initially described in D2.1, this deliverable continuous with the process of requirements elicitation by gradually transforming the user scenarios into user stories.

1.2. Structure of the document

The document is structured as follows. Section 1 describes the purpose of the document and the user stories methodology that we followed in order to describe the functionality of the ENCASE platform. The templates for the user stories are also included in Section 1. Section 2 provides a survey of the existing related web-based tools and the research state-of-the-art. Section 3 provides a description of the measurements and the test data preparation that has been done in the context of ENCASE. Section 4 defines the usage scenarios that our system is going to handle while Section 5 provides the user stories based on the usage scenarios. Section 6 describes the architecture design of ENCASE along with a detailed description of its main components. Finally, Section 7 provides the identified system technical requirements. We conclude in Section 8.

1.3. User Stories Methodology

According to William C. Wake author of “Extreme Programming Explored[1,2]“ good user stories should follow **INVEST** acronym:

- **I**ndependent: User stories need to overlap as little as possible in order t be able to be independently implementable.
- **N**egotiable: User stories need to be flexible in order to allow for small customisations that will be decided upon between the user and implementer.
- **V**aluable: Every user story needs to have a certain value to the end user and a lack thereof would essentially mean that some of the user needs are not met
- **E**stimable: User stories need to be understandable to both customers and implementers at a level such that a fairly good estimation of effort and scheduling can be agreed upon by both parties
- **S**mall: User stories need not be massively complex or large. Good user stories should be equivalent to a few person-weeks of work.
- **T**estable: The user stories need to be easily testable in order to provide a validation mechanism for requirements.

In ENCASE the user stories are written based on those ground rules. A set of technical requirements are then provided based on the user scenarios and user stories and they are categorised according to the four main components of the ENCASE architecture. Those requirements are specific so that

[1] William C. Wake, “Extreme Programming Explored”, 804-934-8194, 2000

[2] http://xp123.com/articles/invest-in-good-stories-and-smart-tasks/

they leave no room for ambiguity or misinterpretation. The specification documents in ENCASE are D2.1 and D2.2 which have to be maintained over the life of the project.

1.4.User Story template

To provide a structured User Story description, a template is defined as follows:

Code number	Coded identification of every user story
Title	User story title
Description	User story text in the following format: As a ... I want to … so that ...
Acceptance criteria	Criteria based upon which the successful implementation of the user story will be established. • Criterion 1 • Criterion 2 . . • Criterion n

Table 1. User Story template

2. State-of-the-art

During the last decade, children’s lives have been completely transformed due to the rapid growth of digital technology. Access to the Internet has been increased and year by year even younger people are getting access. According to a UNICEF study over 40 per cent of 10,000 young people, who participated in a poll started using the Internet before they were 13 years old [1]. As a consequence the number of young people accessing the Online Social Network’s (OSNs) has been also increased. OSNs provide a lot of opportunities to its users, especially young people. They have opportunities for communication with other people, entertainment, education, and even for innovation.

However, OSNs have a lot of risks besides opportunities. Unsupervised access to social networks exposes children to a lot of threats such as cyber bullying, sexual grooming, sexual abuse and any other kind of malicious or abusive behaviour. Also, OSNs give to malicious actors the ability to create a fake identity pretending someone else even younger in age. This enables them to easily deceive a child. Based on UNICEF’s study, the majority minors participated in the study recognized that the treats exists and think that their friends participate in risky behaviours online. Additionally, another problem identified by this study is that more of the participants will turn to a friend for help instead of parents in the case they will face such threats online.

Much effort has been made to tackle such threats in OSNs and the Internet in general, but none of them has successfully solved the problem. This section summarizes all the research state-of-the-art and existing web-based tools for mitigating threats that renders minors and other vulnerable population groups susceptible to cyberbullying, malicious or abusive behaviour and fake activity in OSNs.

2.1.Benefits and risks of Web 2.0 tools

The advancement of Web 2.0 tools offers a rewarding source of knowledge sharing, interaction and socialization. Amongst the benefits reported in the use of these tools include the development of 21st century skills such as creativity, innovation, team building, critical thinking, confidence, information sharing, higher academic achievement and improvement of ICT skills and competences [2], [3]. Yet, being present in OSNs such as Facebook, Twitter, Snapchat and MySpace presents particular risks such as exposure to cyberbullying, child abuse, inappropriate material and contact with dangerous strangers.

Social Web can facilitate abuse of children by adults - being in place to assume fake identities online, a possible “danger” can intrude a child’s private zone leading to violence or even sex crimes [4]. The risks and threats that minors encounter on the web can be classified under the following five categories [5]: a) content risks: instances or events in which children are exposed to illegal content, harmful content or age inappropriate content and harmful advice; b) contact risks: instances or events in which children have direct interaction with other children or adults. Frequent threats under this category are cyber grooming and cyberbullying; c) Children targeted as consumers: instances or events in which children face the risk of being treated as consumers of products and/or services designed only for adults; d) Economic risks: instances or events in which children spent money in gambling and other online games; e) Online privacy risks: instances or events in which children share personal data with inappropriate audience.

2.2.Security and privacy enhancing web-based tools review

This subsection provides a review of the existing web-based tools and mobile applications that are trying to address the security and privacy issues in the OSNs. We list below the most relevant ones to the concepts of ENCASE.

Qustodio is parental control software available in most of the platforms [6]. It enables parents to monitor and manage their kids’ web and offline activity on its devices. It also allows them to track with whom their children are communicating with in OSNs and manage their whole OSN activity. In addition, Qustodio can be used as a sensitive content detection and protection tool.

SocialShield is a Social Network Protection application developed by Avira [7]. It is a monitoring tool that informs parents of their children’s online activities. It monitors and checks their child’s social network accounts for any comments, photos etc. that may influence the child’s reputation in a negative way or may indicate that the child is in danger. Furthermore, SocialShield is able to protect the children from cyberbullying, to prevent them from participating in online discussions with inappropriate content and it is also able to verify the identities of the child’s online friends.

Kidlogger is a free parental control software that helps parents get to know their children are doing on their computer or smartphone [27]. It is able to track what children types, which websites they visit, what applications they use, with whom they are communicating with in Facebook. Besides these, Kidlogger offers a voice-activated sound recorder for parents who are concerned about what their children are talking with other people online.

Spyrix Personal Monitor is a complete and detailed remote monitoring software of user’s activity [28]. It can be used as a parental control tool and is suitable for children’s activity monitoring. It is able to monitor the children’s activity in Social Networks (e.g., Facebook), Skype, websites they visit, and all running and active applications. It offers a lot of monitoring features to the parents and it renders standalone reports coupled with screenshots.

Zoodles Kid Mode is able to turn a device in to a safe learning and playing environment for kids [29]. It combines filtered browsing and a dedicated web browser where everything in it is safe for kids and there is no risk of anything awful popping up. It also allows parents to customize how they prefer Kid Mode to work for their children like forcing them to play more mathematic games. Kid Mode is available in Windows, Mac, Android and iOS.

Web of Trust (WoT) is a safe browser add-on for website reputation rating that helps users to make informed decisions about whether to trust a website or not when browsing online [8]. In order to provide its users with an extra layer of security against malicious links posted by malicious users, Facebook uses WOT’s reputation data to inform users about low reputation links.

WebWatcher is a parental control, cross-platform compatible, monitoring software [9]. It is able to capture the content of emails and instant messages in OSNs, as well as actual keystrokes and screenshots. It assists parents in keeping their children safe online by viewing what is captured in their child’s screen from everywhere.

Cloudalc WebFilter Pro is cloud-based content filtering application [10]. Cloudalc monitors billion of web pages to protect families and especially kids from malicious attacks and threads and to have a safer Internet surfing experience. It blocks web pages, spam servers and adult material.

Abuse User Analytics (AuA) is an analytical framework aiming to provide information about the behavior of OSN users [11]. This framework processes data from users’ activities in the online social networks with the goal to identify deviant or abusive activities through visualization.

FoxFilter - THE Parental control for Firefox is a free browser add-on produced by Mozilla and is known as the parental control for Firefox browser [12]. It is a personal content filter that helps blocking pornographic and other inappropriate content. A user can block content for an entire site or enter custom keywords that will be used to block content for any site that contains those keywords.

Parental Control and Web Filter from MetaCert is a parental control browser add-on that blocks pornography, malware and spyware [13]. It protects kids and adults across multiple categories. It allows you to choose among two main categories (extra strong for kids and Strong for adults) while also allows you to define the specific categories that you prefer to be protected (such as Bullying, Drugs, Aggressive behaviour, Gambling, Sex etc.).

MetaCert Security is a Security REST API [14]. It provides a layer of security on top of web applications so that the application can protect users from Phishing attacks, Malware and Pornography.

eSafely is a parental control browser add-on that provides kid-safe access to popular web resources, free of adult content [15]. Generally, it offers the following: a) Kid Safe Facebook that protects children against cyberbullying by replacing harassing messages with friendly icons in Facebook chat; b) Kid Safe Images that when a site is identified as hosting adult content it replaces the images with images more suitable for children; c) Kid Safe YouTube; and d) Kid Safe Search.

Nude.js: Nudity Detection with JavaScript and HTMLCanvas is a JavaScript implementation intended for client side nudity detection based on approaches from research papers [16]. It analyses image and video data and returns whether it contains nude content or not.

ReThink is a non-intrusive, patented software product that stops cyberbullying before the damage is done [17]. When an adolescent tries to post an offensive message on social media, ReTHink uses patented context sensitive filtering to determine whether or not it is offensive and gives the adolescent a second chance to reconsider their decision.

PureSight Multi is monitoring and filtering cross-platform software that allows children to use the internet without fearing bullies or harassment and keeps parents in the know [18]. It features cyberbullying protection on Facebook, Web filtering, Reports and alerts, file sharing control and parent portal.

MM Guardian Parental Control is a mobile application that allows you to block incoming calls and texts, monitor alarming texts and control which apps on the device can be used and when on a children's' smartphone [19]. It also allows the parent to locate and lock his children’s mobiles with a text message, as well as to set time restrictions to limit their use.

Funamo Parental Control is a mobile application that allows parents to monitor their children's mobile devices [20]. Contacts, calls, SMS, browser history, applications and locations are automatically logged and history data is uploaded to the Funamo server each day. It also allows parents to enable safe search engines in the web.

Screen Time Parental Control is a mobile application that empowers parents with the ability to monitor and manage the time that their children spent on their devices and set time usage limits on selected apps, as well as a bedtime curfew, lights out and school time curfews [23]. The app runs in the background of the mobile device and it can be controlled via any web browser.
The following table summarizes the web-based tools analysis. Such analysis is aiming at overviewing the features provided by existing web-tools for user protection. Moreover, such capabilities can be considered as benchmark capabilities which could be supported by ENCASE solution at web add-on level and/or be carried out in an integrated and more efficient way. Of course it is the ENCASE actual scope and priorities to determine whether these features should be incorporated in the ENCASE add-ons.

Key Capabilities	Tools	SW	App	Browser add-on	API
Monitoring (Monitor web online/offline activity, track OSN communication activity, etc.)	Qustodio	🟦			
	Avira SocialShield		🟦		
	PureSight Multi	🟦			
	Screen Time		🟦		
	WebWatcher	🟦			
	Spyrix Personal Monitor	🟦			
	Kidlogger	🟦			
Website reputation rating	Web of Trust			🟦	
Content Detection	Qustodio	🟦			
	WebWatcher	🟦			
Cyberbullying protection	Avira SocialShield		🟦		
Content Filtering (OSN content filtering and flagging, offensive message determination, identify deviant or abusive activities etc.)	CloudAlc WebFilter Pro		🟦		
	FoxFilter			🟦	
	Meta Cert			🟦	
	eSafety			🟦	
	ReThink	🟦			
	PureSight Multi	🟦			
Content Blocking (Content replacement, webpage blocking, etc.)	CloudAlc WebFilter Pro		🟦		
	FoxFilter			🟦	
	MetaCert			🟦	
User protection (e.g., OSN identity verification, etc.)	MetaCert API				🟦
	Abuse User Analytics	🟦			

Nudity Detection	Nude.js	■			
User Device App control (e.g., app locking, time usage limitation, etc.)	MM Guardian		■		
	Funamo		■		
	KidsPlace [21]		■		
	AppLock [22]		■		
	Zoodles Kid Mode		■		

Table 2. Categorization of web-based tools based on key capabilities

Despite the aforementioned web-based tools, there are other tools like NetNanny, Safe Eyes, Elite Keylogger etc. that are trying to solve some of the problems that ENCASE does [24], [25], [26]. Most of them are parental control tools that monitor children’s online activity using the following methods:

1. **Keystroke logging:** It is the action of recording the keys struck on a keyboard, typically covertly, so that the person using the keyboard is unaware that their actions are being monitored. In ENCASE we will
2. **Screen capturing and monitoring via screenshots:** The action of monitoring a user’s activity on his device by capturing screenshots periodically. ENCASE will not adopt such type of monitoring.
3. **Android monitoring:** It is the action of monitoring, recording and reporting the activity of a user in his android mobile device. In ENCASE we will adopt this method but only for monitoring the user’s OSN activity.
4. **SMS and instant messaging in OSNs capturing:** It is the action of monitoring a user’s OSN activity and reporting, via SMS or instant messaging, any abnormal incident based on keywords. In ENCASE we will use more intelligent techniques, like machine learning techniques, to capture and analyse a user’s OSN activity.
5. **Content blocking and filtering:** It is the action where the administration declares a set of websites or keywords (e.g., pornography) that he wants to be blocked and cannot be accessed from a user’s device.
6. **Time usage limits:** A parent is allowed to set limits regarding the time that his child is able to use his device. We will not adopt this method in ENCASE.

Overall, most of the existing tools rely on monitoring and parent review to detect any abnormal activity. Some of them search for keywords to create alerts, while some others block the usual list of websites.

Cyber-bullying, cyber-grooming, fake identity, false information dissemination, and exchange of sensitive content is not *intelligently* detected by existing web-based tools and this has a negative social effect on the children i.e. they are monitored to an excessive degree and this will probably lead them to find alternative ways to go online. ENCASE platform will be a good compensator for protecting the children from communicating with a person that is willing to bully or exploit them.

2.3. Research state-of-the-art on cyber security risks for minors

This section provides a review concerning the Internet activity and motivation of use by minors and presents in a coherent manner the identified risks and threats that children using the web and OSNs are exposed to.

2.3.1. Minors’ access to the Internet and use of OSN

Nowadays, children are very familiar with technology. Research has shown that they have the ability to familiarize themselves with any electronic gadget very fast and they are able to do sophisticated tasks using these devices. Research has also proven that as soon as children come in touch with electronic device such as PC’s, tablets, smartphones and so on, they can use them instantly, in contrast with adults who may need to study the instructor manual of the gadget.

More and more children, these days, have access to the internet through handheld electronic devices such as smartphones, tablets and portable game consoles [46]. According to Ofcom reports on Internet safety measures and strategies of parental protection for children online in the UK tablets became the favourite device for online access for children aged 8-11 who mostly used them for playing games [47], [48]. Smartphones are the most popular device for social networking and, according to the same reports, the children aged 12-15 have their own smartphone. Most parents believe that children are more at risk when they are online at home than outdoors. However, statistics have proven the opposite. This is because smartphones, tablets and other handheld devices, offer instant access to the Internet everywhere and children prefer that as they are not supervised by their parents. According to the 2016 ITU report on child online protection in USA the number of children who have access to the internet is constantly increasing since 2011 [46]. Children below five years of age use the internet on a weekly basis and as age increases the frequency of access to the internet also increases. The 40% of children aged 8-11 years old make use of the internet daily while the 36% of them use it multiple times per day. The same report reveals that 70% of teenagers are online daily while 25% of them reported that they are permanently connected online. A survey conducted by South Korean government has shown that one out of ten children aged 10-19 years are addicted to the Internet [49]. According to that study, when children are connected online they enjoy using a variety of activities whose number increase by age. For instance, children under 9 years old search for information about school, play games or watch videos [50]. Children aged 10 to 19 also listen to music as well as the above mentioned activities, however their basic everyday use of the internet is for social networking reasons.

The intrusion of online social networks in people’s everyday life the last decade, has met with huge success. There are many social networks services available, so as to meet different needs according to age, language, profession and culture. According to the ‘Net Children Go Mobile’ network report [51], approximately 70% of children in Europe have at least one social network profile while most of them have a profile in media sharing services such as YouTube or Instagram. In UK one out of four children use Twitter to share photos and other content [52] rather than tweeting. A study conducted by Pew Research Center for USA [50] concluded at similar findings. Facebook is the most popular social media site among American teenagers aged 13 to 17 since 71% of them are using the corresponding website. Half of teens use Instagram, while the popularity of Snapchat increases rapidly reaching a 41% of teen’s population. Snapchat allows people to send and receive pictures and videos directly to their phone and created new security concerns for parents [53]. The study of Pew Research Center showed also that about 71% of teens are using more than one online social network site [50].

2.3.2. A taxonomy of online risks for minors

It has being shown in the previous section that the popularity of Internet in general and OSN in particular is high and with increasing tendency among children and teenagers. Thus, the online risks for these sensitive age categories received increased awareness. Several different international organizations and research groups have been trying to study and categorized the dangers which have emerged in the past years including EU Kids Online, ITUs-Child Online Protection (COP), Youth Protection Roundtable (YPRT), Net Children Go Mobile and many others. These organizations conduct surveys in regular time intervals and, based on the findings, recommend safety measures for every identified potential danger that the Internet might pose to children. However, the security and privacy risks themselves are rarely mentioned making it difficult to define energetic actions and to design tools that proactively try to minimize the aforementioned risks and dangers. For instance, in contrary to a few studies such as those of Australian Communications and Media Authority where dangers, of Internet and OSN use, such as electronic fraud, malware and other e-safety threats, are explicitly mentioned, research in Europe usually describes generic categories of risks such as sexual and commercial [54], [55].

Categorization of online risk for children is not easy. In most cases risks are caused or affected by a variety of reasons emanating not only from children’s online lives but their real lives as well. In addition many risks and threats are crossing several categories. In the corresponding literature the following distinctive situations have been defined [46], [52], [56]:

- Online risks which are the expansion of problems in real life, for example pornography.
- Risks which arise from the interaction of two under-agers such as cyberbullying.
- Risks which arise from the interaction between a child and an adult, such as cyber grooming.
- Risks which arise by the collection of data, against the protection of privacy, such as viruses and other malware.

In addition of potential dangers, children on the internet might be exposed to, can be assessed based on the legal importance and by discriminating the cases where the child is the victim or the predator.

Another popular, in the related bibliography, categorization of online risks is based on the way the Internet is ‘used’ and/or perceived by the children. The first clearly concerns the risks of the Internet as product of technology or simply stated the risks that arise due to minors’ access to Internet content. The second category, concerns incidences where the Internet provides the means through which the children are exposed to dangers, i.e. contact risks, and finally, the third category refers to cases where children are aimed at as online consumers [57], [58].

2.3.2.1. Content risks

As already stated, children are able to familiarized themselves with Internet and generally with technology as they grow up parallel to it. This fact combined with the fact that in 2015 there were more than one trillion websites, turns children into a vulnerable group or exposed to many dangers related to the content of the Web. Content risks are divided, according to bibliography, in three categories: (a) illegal content; (b) harmful content or age inappropriate content and; (c) harmful advice.

Illegal content refers to content which is illegal to be published online. For example, it might be content about sexual exploitation of children which is illegal in most countries. Inappropriate content usually depends on the age of children that have access to and may contain, for instance, adult pornography. Hatred or violence related content, although not illegal, may harm children in case they gain access to it. Age inappropriate content may be mentioned, as term in national or local cultures and social values, however, in literature and official documents this term focuses more widely on pornography and other sexual content [58]. The meaning of pornography may vary between countries and between groups within a country. Pornographic content is fairly easy to be found by anyone online, however, younger children are more exposed to offline pornography than online ones [59]. Nevertheless, a lot of studies agree that exposure of children to online pornography content increases by age. In addition it was found that random exposure of children to pornographic content, on the Internet, is more common than intentional access and it increases when the names of the websites or URLs are misleading for children. According to ITU the rates at which children of young ages are exposed to websites of pornographic content appears an increasing tendency [46]. This happens even to children whose parents have locked access to sites of inappropriate content. The high percentage of children that randomly access to pornographic content continues with intentional access. According to Dooley et al. only children of very early age reported being upset by being exposed to pornographic content [60]. As for the exposure of children to violate content researchers did not arrive yet at concrete findings and it seems that additional research is required.

Harmful advice refers to content which may lead a child to consume alcohol and drugs or to commit suicide or different psychological and nutritional disorders. In combination with the fact that anyone can provide such advice online through social networks and other platforms, it is very easy to children to have access to and be influenced by it. Researchers state that many of these advices maybe well intended; thus, it is difficult to be categorized to harmful or useful [58].

2.3.2.2. Contact risks

Contact risks refer to instances or events that children have direct interaction online, either with other children or with adults. This can be achieved through child's participation in online chat or social networks chats. A frequent phenomenon is when adults try to develop relationships of trust with children with the aim of having sexual intercourse with them. This constitutes a criminal act in almost all countries and is known as cyber grooming [58], [46]. Cyber grooming is often when an adult sexual predator seeks a communication with its victims in a direct online conversation with the aim of coming in offline sexual relation with them without mentioning his/her real age and identity to the children taking advantage of their naivety [61].

Cyberbullying is another contact risks for the children. The term cyberbullying refers to bullying that children undergo through the Internet. Bullying may come in different types such as threats, humiliation or harassment. Cyberbullying differs from cyber stalking and cyber harassment. While in cyberbullying there is participation of peers of both sides, in the event of an adult participant it constitutes cyber harassment [62]. Experiencing tense emotions such as anger, desperation or vengeance are frequent reasons causing children to be exposed to cyberbullying. Emotions which stem from problematic situation in the family background and problematic relationships in general are also common reasons. Researchers indicated that cyberbullying constitutes in many cases some form of entertainment, satisfying in this way power struggle needs.

In comparison with traditional bullying, cyberbullying offers some advantages to predators. The most important of which is the ability to remain anonymous which they achieve by using aliases, fake profiles, fake accounts, face social media profiles, text messages, instant messaging and other services that internet provides so they do not reveal their identity. Cyberbullying is one of the biggest threats that social networks pose. In recent years more than 3 million children have undergone cyberbullying in any form whether this constitutes harassment or threats. A high percentage (95%) of them reported that they have been victims of cyberbullying on Facebook.

Eight out of 10 adolescents who use social networks share personal information about themselves such as photos or videos, location information and contact information to a much greater extent compared to previous years. According to several studies sharing personal information such as age, phone number, school and location are the main reasons for young people to undergo cyberbullying through social networks [59], [63]. In recent years electronic games have shown an enormous increase. These games either through PCs or game consoles support features for online games and games with multiple players. Most of these games have special chat rooms so that communication among players may be easily achieved. Robinson’s research indicates that approximately 20% of the children who reported having undergone some kind of cyberbullying where cited cyberbullying to have being taken place during in an online game [49]. The most usual way of cyberbullying in an online game refers to schools, online game communities and direct communication between online players. OECD reports that the risks that minors run for sexual harassment by adults is limited; 25% young children share information and interact with strangers on the Internet, however, only 5% of them had spoken to a stranger discussing sexual matters [58]. In addition it is mentioned by OECD that most children tend to ignore the conversation and take proper steps. It is noteworthy that potential sexual predators are adolescences and adults younger than 21 years old. In general, the possibility of physical sexual contact with an adult through an online approach is very rare. Ybarra reports that only eight out of a sample of 1500 hundreds reported physical sexual contact, all of whom where aged 17 and above [64]. Furthermore, it was found cyber-grooming for children aged of 12 or less is extremely rare [63]. These results indicate that cyber grooming contains minimum danger; however it is difficult to measure precisely. Research agrees that online harassment constitutes the most widespread Risks that children face. Various individuals use the means of technology offers (social media, chartrooms etc.), with a view to harming others through bullying, humiliation and embarrassment and treats. Those who cause cyberbullying are underagers as are their victims. Despite this there have been instances where cyberbullying is caused by adults. Cyber talking refers to the event where an individual is exposed to an online extreme behaviour of another individual whose purpose is malevolent treats and/or psychological or physical predicament of the victim. Overall, cyberbullying and cyber harassment constitute an ever increasing field the prevalence of which is extremely worrying [62].

2.3.2.3. Children targeted as consumers

Children on the internet face the risks of consumers, mostly for products and services especially designed only for adults. Such cases relate mostly to products such as alcohol, tobacco and prescription medicines. Children may come in contact with advertisements about these products. Furthermore, children may come in contact with the promotional illegal products such as drugs or doping substances. A study in US showed that 75% of teenagers that tried to buy cigarettes online managed to do so, while in 2002 only a percentage smaller that 3% had succeed in doing so [64].

Minors and more specifically young children are not able to realize that content on the internet is produced a financed and that is why they have difficulty critically assessing advertisements and advert messages. There have also been instances where online marketing exclusively targets websites for children for example online games. This fact has caused many countries to question integrated ads on websites aimed to children. Online marketing and advertisements may harm children. This happens mainly with products or services aimed for adult such as gambling, pornographic content and dating services. A study by Netchildren show that about 10% of the ads were about games and 5% about dating services [48]. Advertisement of pornographic content from banners and popups constitute the main reason while children accidentally came in contact with improper content.

2.3.2.4. Economic risks

It is a frequent phenomenon for children to spend exorbitantly if they have access to payment methods either through a mobile phone or other online services, thus creating huge cost for parents [67]. The most usual instances are by registering and transferring money in gambling and other online games. Many games require some form of subscription for some particular reason or to support multiplayers. Players may spend a lot to buy virtual characters or other features. There are, however, cases where children may spend huge amount of money through fraudulent transactions [59]. This occurs when services do not clarify that after the purchase of a product or service there would be extra charges. A common example of this is ringtone download services for mobile phones who charge extra for registration. According to OECD in 2008 24% of Belgian adolescents reported having paid more for ringtone downloaded and 9% registered in such kind of service without realizing it [58]. All the above risks are exacerbated with children of younger ages because of their inexperience. Nevertheless, minors who do not own a bank account or have access to their payment methods are less likely to suffer economic fraud.

2.3.2.5. Online privacy risks

Safety risks for private life information relate to all users. Children however, constitute an especially vulnerable group as they do not possess the necessary critical thinking to understand and predict the consequences. Personal information privacy in the case of children is at risk where the personal data is collected on the internet automatically following their request to search engines or other services. This may happen in various ways; the most usual is which collecting cookies, electronic registration in surveys and filling information data in electronic forms. In addition children as well as most adults, skip user term in order to have access to services they are interested in. According to OECD, 40 websites especially offered for children will analysed and almost 75 of them ask for personal data [58]. In most websites it was not compulsory however they did ask for personal data such as email age, birthday etc. so that they could gain access to subpages of the site [66]. There are also different websites who target children and the collection of their personal information offering quiz, competitions, research, using marketing techniques, such as a discount or free service or an award managed to gain the personal data as well as their families or friends. The research shows that minors give out personal information easier than adults in order to receive an award [65]. Children may share and reveal personal data because they cannot realize how widespread online viewers are, neither all the possible consequences. Underagers have also addicted social networks and other apps to great extent, publishing information photos videos, thus revealing important information about their life family, friends and of course themselves [61].

2.3.3. Summary

In conclusion, the huge spread of the World Wide Web and the opportunities that it offers, besides the enormous advantages, poses many risks especially for children. Research shows vast adoption of the internet by children. However, the rates where children are exposed to risks vary by country, age and gender. Pornography and cyberbullying constitute perhaps the greatest risks which children are exposed to, as is an extension of the problem of real life. Online social networks and other Web 2.0 applications are at the greatest risk because they constitute the ‘vehicle through which children may be exposed to many dangers and threats. Summarizing, the Internet contains many risks for children as they are a vulnerable group. It is an issue that needs further study and protective measures to be taken to reduce the risks that threatened children physically and psychologically.

2.4.Research state-of-the-art on security/e-safety in online environments

This section explores the research development pertaining to safety and security in online environments.

2.4.1. Introduction

The rapid evolution of social technologies, or the so-called Web 2.0 technologies, has occurred in many aspects of business, communication and education. Crook and Harrison define Web 2.0 as “a catch-all term to describe a variety of developments on the web and a perceived shift in the way the web is used [30]. This has been characterized as the evolution of web use from passive consumption of content to more active participation, creation and sharing – to what is sometimes called the ‘read/write’ web”. This term encompasses technologies that emphasize social networking, collaboration and media sharing such as Facebook, Twitter, YouTube and Flickr.

A fundamental dilemma parents, educators and everyone need to address when considering the use of social technologies with children relates to e-safety. Increase use of social technologies, and their ubiquity in children’s lives, demand that actions are taken for ensuring children's safety and security. The question of how to allow such tools in children's life (e.g., their education to allow productivity, engagement and learning) without violating their safety and personal rights has been a key issue in a number of research papers in journals and conferences (cf. Special Issues in Journal of Computer Assisted Learning: Social Software, Web 2.0 and Learning). Some studies have been guided by the wish to understand students’ and teachers’ concerns in incorporating social technologies in the classroom and some by the wish to identify methods for handling e-safety in a cost-effective way [32], [34]. Despite the popularity of social technologies in our daily lives, they are surrounded by concerns (from students, educators, parents, social workers, researchers and other stakeholders) with regard to their vulnerability linked to safety and security issues. For example, concerns about the use of Web 2.0 technologies in the school environment relate to exposure to online bullying, inappropriate material and risk of contact with dangerous strangers. This section provides the research state-of-the-art on e-safety in online environments.

2.4.2. Methodology

In order to synthesize the findings of research regarding e-safety in online collaborative environments, we followed a three-step approach (see Figure 1), which included: (a) compilation of the e-safety corpus which included research manuscripts related to e-safety from manually search in scientific databases; (b) refinement of the e-safety corpus and (c) synthesis of the research manuscripts.

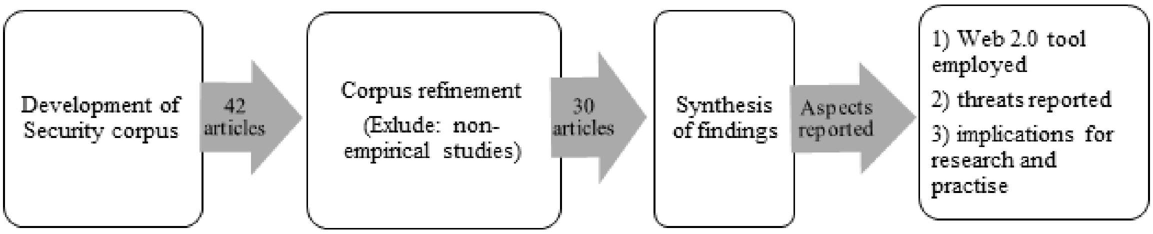


Figure 1. Flow diagram of the methodology adopted for exploring scholarly activity in e-safety in online collaborative environments.

2.4.3. Development of Security corpus

In order to understand scholarly activity on children's e-safety and security in online environments, we started by selecting appropriate resources which compiled the e-safety corpus. Appropriate articles for inclusion were selected via manual keyword (e.g., “security”, "safety", "social media", "e-safety", “threat”, Web 2.0 etc.) search in manuscripts’ title, abstract and given keywords. in the following databases: ERIC, Education Research Complete, Academic Search Complete, Computers & Applied Sciences Complete, Springer Link, Research Starters, Psychology and Behavioural Sciences Collection, Food Science Source, Taylor & Francis Group. The keyword search returned 45 manuscripts which comprised the preliminary e-safety corpus of this review.

2.4.4. Corpus refinement

Each manuscript from the corpus was screened in order to elucidate the aim of each study which was given in the form of a quote in the authors’ own words. This stage facilitated the screening of articles to be included in the e-safety corpus, excluding seven articles as reporting on non-empirical studies. The final corpus included 30 manuscripts.

2.4.5. Synthesis

Each paper in the e-safety corpus was then examined in detail to extract information related to the following pre-defined dimensions: (a) tools and threats dominant in online environments; (b) methods and tools for handling threats in online environments and; (c) implications for stakeholders including researchers, parents and educators.

2.4.6. Findings

Recent debates about students’ activities with social technologies strive between the perceived benefits and the potential threats. The social web is seen to have the capacity to foster the 21st century skills, yet students, teachers, IT administrators and parents demonstrate increased concern about the online risks and threats, often related to child sex abusers, pornographic content, and bullying. Concerns about online safety fit within a broader agenda related to students’ e-safety, recognizing the need to develop the skills and competences needed for taking advantage of the benefits that ICTs can provide. In the following sections, some themes are presented from the safety corpus of manuscripts.

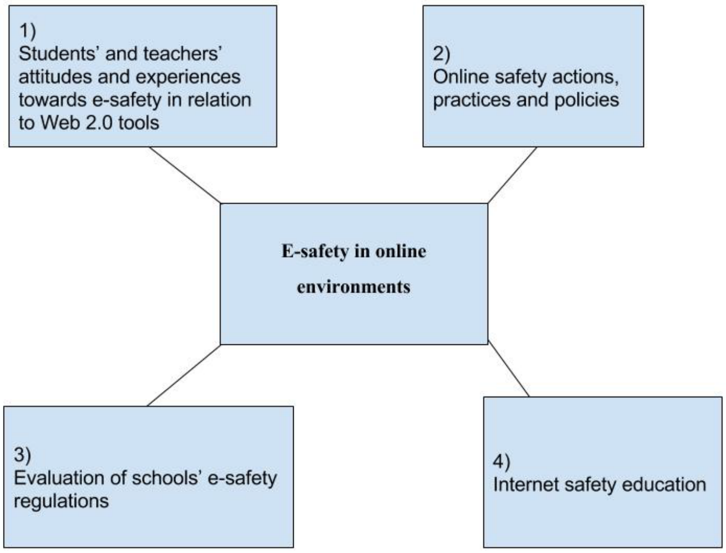

1)
Students’ and teachers’ attitudes and experiences towards e-safety in relation to Web 2.0 tools
2)
Online safety actions, practices and policies
E-safety in online environments
3)
Evaluation of schools’ e-safety regulations
4)
Internet safety education

Figure 2. Overview of e-safety in online environments

2.4.6.1. Students’ and teachers’ attributes and experiences towards e-safety in relation to Web 2.0 tools

In this line, Sharples et al. report results of a survey of children, teachers and parents of teenage children across England [32]. The survey data were complemented with focus group interviews with students and individual interviews with teachers, managers and technical staff (IT administrators) to gain a thorough understanding of Web 2.0 activities and concerns. Findings demonstrated that a high percentage of the children surveyed (74%) have used social networking sites (SNS), whilst a substantial minority interacted regularly online with people they have not met face-to-face. Although teachers demonstrated the desire to take advantage of the benefits of Web 2.0 for creative and social learning, they reported being limited by a need to show a duty of care that prevents worst-case risk to children, to restrict access to SN sites. The respondents also report concerns about Internet bullying and exam cheating. Finally, a Policy Delphi process voiced the need for schools to allow access to Web 2.0 sites, with children being educated in responsible and creative learning.

2.4.6.2. Online safety actions, practices and policies

Within this theme researchers engage in online safety actions, practices and policies. For example, Searson et al. describe the need for developing informed policies and practices that would involve a wide range of sectors of the society [33]. Such practices would inform technology integration in educational settings addressing the following factors: national and local policies, bandwidth and technology infrastructure, educational contexts, cyber-safety and cyber-wellness practices and privacy accountability. Two organizations offer examples and set guidelines for digital citizenship in educational settings that is ISTE and iKeepSafe. On the same line, Waters highlights the multifarious security challenges that school districts encounter, using as a stepping stone the example of a high school’s page that has been hijacked by a former student [34]. The manuscript concludes by suggesting two web browser add-ons -Firesheep and BlackSheep- for users on unsecured Wi-Fi networks to identify the social networking sessions of others on that Network. Similarly, the Parent Teacher Association demonstrates its action in educating children and parents about Internet Safety (A SAFER DIGITAL WORLD). On the same line, Ramnath discuss how school administrators can protect students’ safety while integrating technological advancements in teaching and learning [36]. The study engages in topics such as cyberbullying and cyber-stalking, the use of social networking sites for collaboration and the use of Mobile Device Management for the safety of mobile devices within and outside the school network. Similarly, Campbell-Wright examined e-safety in e-learning, the benefits and dangers of online interaction and guidelines for preparing organizations to handle e-safety [37]. Similarly, Wespieser, upon a survey distributed in 14,309 young people in London, demonstrated the high percentage of internet usage and social network sites, as well as issues of bullying and exposure to inappropriate material [39]. The British Educational Communications and Technology Agency (BECTA) investigated the use and impact of Web 2.0 technologies in and out of school [45]. Findings demonstrated that at Key Stages 3 and 4, learners’ use of Web 2.0 is extensive and is currently done outside school, and for social purposes. The major challenge for schools in considering the adoption of Web 2.0 technologies is how to support children to engage in productive and creative social learning while protecting them from potential risk. Most learners demonstrated awareness of internet dangers, though many performed poorly in e-safety (e.g., in practice around password security). Whilst parents are generally positive in the use of technology for learning, yet concerns about e-safety exist. It is schools’ responsibility in raising children's awareness on safe engagement with Web 2.0 and internet safety in general.

2.4.6.3. Evaluation of schools’ regulations

Being in place to understand and evaluate schools’ e-safety regulations is an issue that attracts high interest from researchers. Lorenz et al. explored 201 e-safety related stories presented by students aged from 12 to 16, parents, teachers, school IT managers and police [40]. Through the stories, typical behavioral patterns were mapped, beliefs, regulations and limitations regarding the use of social networks in schools in Estonia. The results demonstrated that few schools hold explicit policies which target e-safety issues. Yet, even these few school-level policy documents fail to address the topics which were most frequently mentioned in the stories written by students. Safety incidents related to cyberbullying or exposure to illegal material remain unsolved or even undetected. Schools delegate any safety incidents to parents who in turn look to schools for assistance. As a principle, e-safety policies should focus on topics with which all stakeholder groups agree being important: gaming, fraud, password, harassment, pornography and meeting strangers. Emphasis should be placed in assessing e-safety risks and how they can influence online learning activities. Similarly, Cranmer reports on excluded young people’s experiences of e-safety and risk demonstrating that the strategies they employ to manage their online safety are primitive and insufficient, thus pointing the need for developing further their online strategies and ultimately their digital literacy [43].

Following a somewhat similar path, Lorenz et al. moved further in analysing the types and sources of safety incidents, the solutions offered, the students’ reactions from these incidents and the solutions suggested by students and whether these solutions actually apply in real-life situations [41]. Findings demonstrated that many students do not understand what e-safety is, assuming that they are not involved in any way in an e-safety episode, even if they have been bullied or “attacked” on the internet. The awareness training about “stop-block-tell” does not work as it is something radically different from how students are thinking and acting. Blocking unwanted material is the least successful solution for the students, even if current typical awareness training is focusing on it. As findings demonstrated, students seem to be passive reactors to any malicious behaviour, thus training focusing on “stop-block-tell” or “don’t click everywhere” seems unsuccessful. The solution provided by the authors “is to include more technical and other practical aspects in the awareness training and distribute step-by-step, common-language how-to-s like how to set one's privacy settings, how to report a page, picture, video or how to behave when someone is being bullied, or what to do when one becomes a victim of fraud or slander. Awareness in those areas is also important for adults who are setting the standard how their students or children behave and deal with the problems in the future” [41]. Ultimately, it is of major importance for schools to develop policies, strategies and solutions that address the core issues of children.

2.4.6.4. Internet safety education

Internet safety education is a topic that attracts researchers’ interest as advancement of technological systems calls for schools to teach children to protect themselves on the web. Whilst internet safety was introduced with some “special occasion” events or a dedicated “Internet Safety Day”, yet these actions seem to serve no purpose and have no real learning impact [35]. On this line, Naidoo et al. present a cyber –safety awareness framework that introduces cyber safety awareness education to primary school children in the South African community [31]. The cyber safety awareness framework offers multifarious benefits for bridging the lack of cyber safety awareness both in schools and in communities. The framework proposes that schools are grouped into clusters, with a cluster coordinator as its head. Cyber safety awareness information is expected to be disseminated through workshops attended by teacher representatives of these school clusters, and distributed back to parents, children, other teachers and ultimately to their communities. On the same line, Orech elaborates on the Digital Citizenship Project that aimed at integrating Internet Safety in the educational curriculum [35]. Through the program, students learned about cyberbullying and prevention as well as strategies for protecting themselves in case of a cyber-insult. The project had successfully employed social media for engaging middle school teachers and students to discuss about netiquette, digital citizenship, cyber-crime prevention and managing digital footprint. Ultimately, sophomore students and teachers become cybermentors engaging in conversations about cyberbullying prevention and protection. Following a somewhat similar path, Moreno et al. consider internet safety education of vital importance for youth in US, thus they surveyed at what age should such education begin and what group is held responsible for teaching it [42]. Having distributed their survey to 356 teachers, clinicians, parents and adolescents they demonstrated that the optimal age for internet safety education is 7.2 years (SD = 2.5), whilst parents were identified as the stakeholder with the primary responsibility in teaching this topic. Clinician's role was also recognized as vital in providing resources, guidance and support.

2.4.7. Implications for researchers and practitioners

As the usage of social technologies advances, the more children and adolescents engage with these technologies on a daily basis. Internet usage has changed the way literacy is perceived and taught, raising the crucial need not only for information literacy, but also for digital literacy and specifically e-safety education. In this endeavour, the question of how parents and educators can accommodate children's behaviour on the net still needs to be further investigated. As noted by Lorenz et al. there is a need for more technical training as well as, more automated solutions that would set one's privacy settings, instructing on how to report a page, picture, video or how to react when someone is being bullied [41]. Within this spirit, the overall aim of ENCASE is to leverage the latest advances in usable security and privacy to design and implement a browser-based user-centric architecture for the protection of minors from malicious actors in online social networks. With an intelligent, malicious behaviour detection browser add-on, ENCASE promises to provide a solution for protecting the children from online harassment, cyber bullying victimization and other malicious activity.

3. Measurements and Test Data Preparation

The goal of the data collection is to identify the magnitude of the problem and to extract requirements for the ENCASE platform. Through our own measurement and analysis of real OSN data we intent to:

a. Validate the results of the research state-of-the-art provided in D2.1, e.g., quantify the severity and occurrence frequency of the different security and privacy problems; and
b. Prepare a test input for the development of the security and privacy enhancing tools, for the testing and piloting activities of subsequent WPs.

3.1. Dataset for Sexually Abusive Behaviour Detection (YouTube)

This section provides a detailed report of the work performed at ROMA3 and INNO for the purposes of Task 2.2. This report contains the online sexual abuse related data measurement and collection along with detailed description of a custom Social Network created using BuddyPress.

3.1.1. Background of the work

At first we investigated the approach of OSN emulation. We created a custom Social Network using BuddyPress and enabled chat/comments section in order to get real time cyberbullying and sexually abusive comments. But as the characters for this chat room are mostly fictional and most of them are handled by the same person, questions were raised upon the legitimacy of this custom Social Network. Based on those questions, the involved partners decided that both ROMA3 and INNO had to perform online sexual abuse related data measurement.

3.1.2. Sexually Abusive Word Dictionary

While investigating how we should create a test dataset we came to the conclusion that we need to define two different types of databases. The first contains the sexually abusive words and the second all the sexually abusive comments. While performing automatic abusive comment detection, the generated tool or software should search for the words contained in the words dictionary and try to identify the sexually abusive comments when it finds any in the test dataset. Dirty Sex dictionary [68] is an example of a dictionary, which contains all the sexually abusive words needed for the identification of the sexually abusive comments.

3.1.3. Test Data Collection

For the data collection of online sexually abuse content we targeted different celebrity singer's song videos and trending cartoons from YouTube platform and looked into their comments' section. Until now we manually gathered a total of 1000 comments from different videos. More specifically, we gathered comments from "Peppa Pig" cartoon's comment section, where these cartoons are most popular and trending and are meant for kids' age group of 4-8. The collected data are stored into an online Excel document with six column of information such as serial no. of comments, collection date and link, short description of the link content, rating and comments. We left rating column blank because we were not sure on which basis/rule we should rate them. One of the possible solutions is to rate the comments on the basis of number of abusive words present in that comments; Rate one if it contains only one abusive word, rate 5 if it contains 5 or more than 5 abusive words and same way for the rating from 2-4.

The idea of storing the collected comments in an excel document came from [69] and [70] that we surveyed and the reference link that they provided for the Kaggle dataset is [71]. The Kaggle dataset contains training and testing data in a '.csv' file with mainly three column of information i.e., Insult Rating 0 or 1 whether it is insult or not; date of collection and the actual abusive comments. Our dataset has been prepared in replica of this Kaggle dataset.

The document that contains our test dataset can be found at:
https://docs.google.com/spreadsheets/d/1ypwhPYlCWnaiICjJ9jf_lHQQCeunTUG0UE8WwhYRjCk/edit?usp=sharing.

3.1.4. Preliminary Testing of collected data

In order to verify the collected data we tested them using most of the web-based existing tools reported in D2.1 and also some other tools that we found later. Following is a detailed report of these tools along with the test results of each one:

a) First we tried **Qustodio**. Qustodio blocks most of the YouTube video links along with their comments section by turning on the "Restricted Mode" of the YouTube page and you cannot turn it off manually. This enables the tool to block most of the videos which are vulgar and contains abusive comments. On the other hand, since this tool works at real-time then we are not able to test our comments dataset's data by putting it as input to this tool.

b) We also investigated some other tools reported in D2.1 and we found that most of them can only block adult content and sexually abusive comments from being posted only at real-time. So we could not properly check our dataset using these tools as soon as they do not allow us to feed our collected comments as input.

c) In the end, we tested the **Aneesha** tool. Aneesha is a freely available Python based cyber bullying detection tool [72] and we investigated whether it can work with our dataset after some minor modifications. This tool has been produced in the context of a project called

“Developing an online cloud-based cyberbullying detection system to Bhuva Narayan at UTS” [73], which was funded with a small grand provided by AUDA Foundation.

This project consists of:

- A web-based application build in django (i.e., Python) that includes a dashboard for users to monitor cyberbullying (located in the cbd_project folder).
- A machine learning classification algorithm (e.g., Support Vector Machine) can be trained to identify cyberbullying messages and then the classified messages can be imported into a database and is summarised on a dashboard.
- The dashboard displays time series data, topic models (for cyberbullying messages and non cyberbullying messages) and a summary of the affective dimensions found in the test messages (for cyberbullying messages and non cyberbullying messages).
- Cron scripts (written in the cronscripts) folder that must be scheduled to perform topic modelling and affective sentiment analysis (i.e. topicmodelandaffectivelexicon.py).
- A moderation role that is able to mark classified messages as false positive.
- A sample python script to get data from Twitter (i.e. injest_twitter.py)

3.1.5. Customized Social Network

3.1.5.1. Data Repository

a. Local

To facilitate an easy way of storing and accessing data as well as running and elaborate complex queries we have setup a local machine running Ubuntu 14.04 LTS server and Mongo DB 3.2 with the necessary database users and accessibility.

b. Remote

Mongo DB instance on the mLab cloud service located in Northern Europe. First Install Mongo DB from the above mentioned installation guide. Then use following connection details from the “mongochef.exe” to remotely connect with the DB.

Connection details:

URL: ds040309.mlab.com

Port: 40309

Username: bio

Password: biometricsrm3

Authentication DB: encase

Authentication Mode: Standard (MongoDB-CR or SCRAM-SHA-1)

3.1.5.2.Data

As no data were available and according to our contingency plan we decided to generate our own scenario based dataset were we would have expected outcomes in a future algorithm based detection process. In the meanwhile, we were still looking for any other possible data sources.

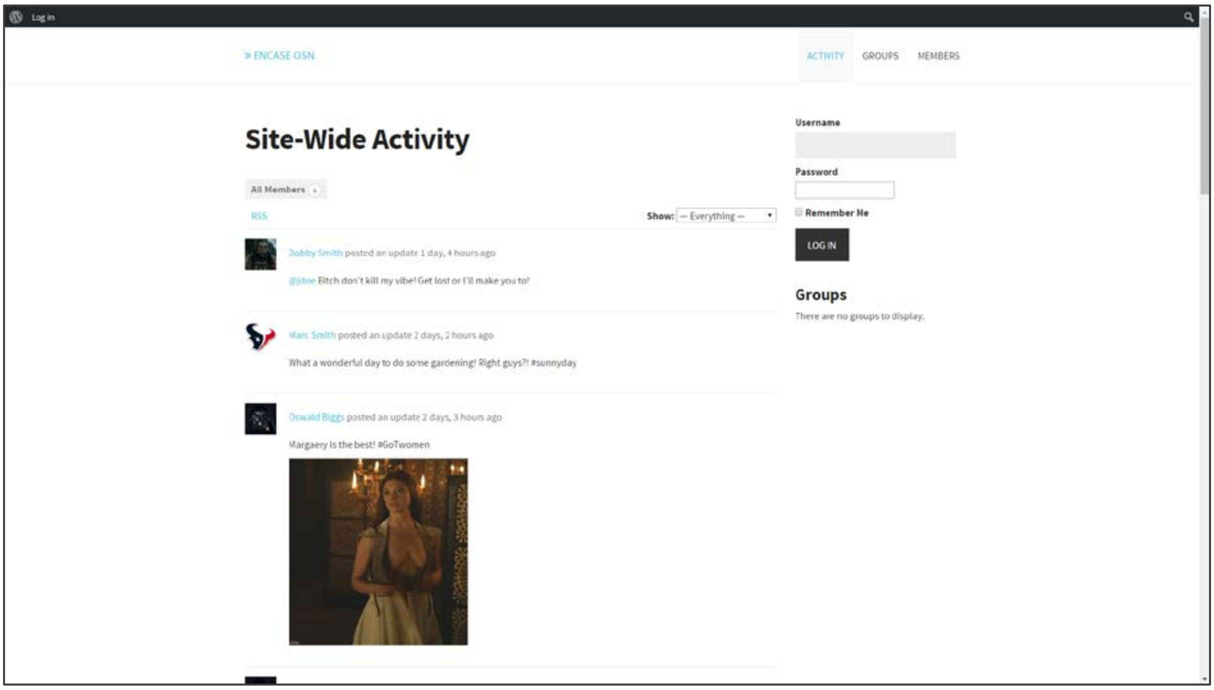
3.1.5.3. Controlled Scenarios

We created a frictional group of 5 actors to participate in our scenarios. Each scenario would be abide to a template where a type is assigned to denote if the scenario should raise a flag during an algorithm detection test or not.

3.1.5.4. Custom Social Network - ENCASE OSN

Since we discovered that Facebook, our first OSN choice, posed certain technical difficulties that would slow down our progress, we created an instance of BuddyPress on Azure to deploy our scenarios. BuddyPress is a WordPress plugin with a tremendous community to support it.

The OSN is located at https://encase.azurewebsites.net. For elevated access please contact Pantelis Nicolaou (CYRIC) at p.nicolaou@cyric.eu.



Site-Wide Activity

Bobby Smith posted an update 1 day, 4 hours ago

@jdoe Bitch don't kill my vibe! Get lost or I'll make you to!

Marc Smith posted an update 2 days, 2 hours ago

What a wonderful day to do some gardening! Right guys?! #sunnyday

Oswald Biggs posted an update 2 days, 3 hours ago

Margaery is the best! #GoTwomen

Groups

There are no groups to display.

3.1.5.5. API Access

To better simulate the access to a real Online Social Network such as Facebook or Twitter we enabled API access t BuddyPress. The documentation of all the available API calls is located at http://tweichart.github.io/JSON-API-for-BuddyPress/doc/index.html.

3.1.5.6. Produced Data

Once we finished implementing our scenarios, we pulled a data set sample from the OSN using the aforementioned API. In addition, scripts were created to ease the insertion of the produce data into the Data Repository.

3.1.6. Summary of the work

In total we have collected over 1000 sexually abusive comments from YouTube videos where most of them are in English and few of them are in Italian, Spanish and Portuguese. Any comment in any language other than English has been marked in the “Short Description” column of the dataset. Also, as soon as the dataset is in .xls format it can be changed into any other file type when this is desired.

3.2.Dataset for Early Malicious Activity Detection (Twitter)

This section provides a detailed report of the work performed at SignalGeneriX (SGX) during the placement from AUTH (July-January 2016) for the purposes of Task 3.1. This report contains the Twitter topology dataset and the identified spammers that have been utilized for the publication “Early Malicious Activity Discovery in Microblogs by Social Bridges Detection”.

3.2.1. Background – Related Work

There have been indications in existing literature that spammers are able to acquire large number of followers in OSNs, thus ending up as highly influential nodes in a social graph. The influence of spammers has been studied in various OSNs (Facebook, Twitter, Ask.fm, Flickr, LiveJournal, YouTube etc.) [90, 91] under different scopes like the topology of the network (connections) [92], textual analysis (posts and comments), link analysis (URLs) [93], etc.

Efforts to identify spammers in OSNs have also been made with popular approaches including automatic dissemination of spam like [94, 95], tools used by spammers to deceive search engines [96] or faking honest behaviours [97].

3.2.2. Selection of Data and Online Social Network

The majority of these networks are based on an explicit user graph to organize and share content as well as connections online. The links formed between users are often public and can thus be crawled automatically for a large number of users. Another advantage of utilizing the topology of connections is the accessibility of this information even for connections that are not part of a user’s existing network. If we consider the scenario where a new connection needs to be evaluated as a spammer or honest user, the information regarding the network of links around this connection are often public and attainable, whereas information about the posts this connection has shared are usually disclosed in its friends/followers network. As a result, we opted to utilize topological data about OSNs in order to analyse the spammers’ behaviour.

Our OSN of choice was Twitter, due to its popularity in the literature and its policy to suspend user accounts that share malicious URLs. Twitter constitutes also one of the most popular OSNs and 42% of teenagers in USA between the ages of 15-17 actively use Twitter, whilst the popularity in younger kids has risen also with 21% of 1314 year old kids using Twitter in 2015 []. What is more, the following relationship in Twitter allows for a directed unweighted graph to be formulated based on its topology, as compared to e.g. Facebook where the friend relationship is mutual (undirected). Since spamming is often targeted and aggressive, the directed follow relationship in Twitter allows for the dynamics of spam behaviour to be more clearly identified.

3.2.3. Original Dataset

For the purposes of our work, we utilized a massive and widely used in literature dataset from the Social Computing Research lab at Max Planck Institute (MPI) (http://socialnetworks.mpi-sws.org/datasets.html). This dataset (http://twitter.mpi-sws.org/) is comprised of two parts: the part regarding the social (follow) links between users and the part containing the tweets posted by the users included in the first part. As we focus on the topological study of Twitter, we have only utilized the first part of the dataset, which includes 54,981,153 user IDs and 1,963,263,821 links between them. The crawling process took place in August 2009 [92] and the queries to the Twitter API included all possible user IDs ranging from 0 to 80 million. The range was upper limited to 80 million, as no user in the collected data had a link to a user with an ID greater than 80 million. Also, not all possible IDs actually corresponded to user accounts; therefore only 54 million accounts were retrieved. Even though this dataset has been collected more than 7 years ago, it is still the largest openly available graph dataset for the Twitter network of connections.

In the above mentioned graph dataset from Twitter a set of 41,352 spammer accounts suspended by Twitter have been added and their connections to the rest of the users were retrieved. This dataset (http://twitter.mpi-sws.org/spam/) and the results of an analysis conducted on it to allow for combatting link-farming were presented. The spam accounts were collected based on Twitter’s official policy of suspending accounts, which were considered to have participated in malicious activity (http://tinyurl.com/22obg56). In particular, the way to identify whether a user account has been suspended is to crawl the Twitter API after a significant amount of time has gone by and attempt to access the same user accounts. If the user has been suspended, then the crawl would be redirected to the suspended page of Twitter. The approximately 55 million IDs, which were included in the Twitter dataset by MPI in 2009, were accessed again in February 2011 and over 379,340 of them were found to have been suspended. However, not all of these accounts were considered as spammers, since there are other reasons Twitter may suspend an account (e.g. being inactive for longer than 6 months). Therefore, two blacklisting services for URLs were utilized (bit.ly and tinyurl) to identify which of these suspended users have shared at least one blacklisted URL; this resulted into 41,352 user IDs identified as spammers.

3.2.4. Data Sampling

From this massive Twitter dataset, we sampled a graph starting from 500 spammer users and extracting their connections according to the relationships depicted in Figure 3.

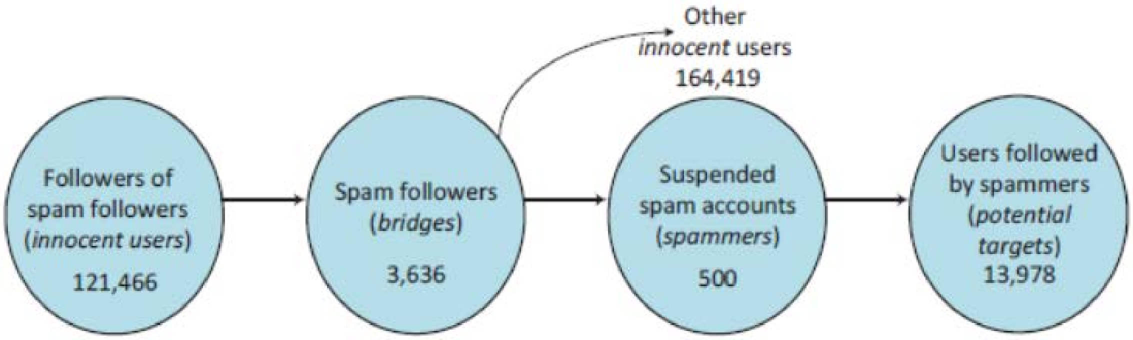


Figure 3. Groups of users and their respective sizes in the Twitter graph built around a group of spammers

Starting from a core of 500 spammer IDs we have retrieved all the users these spammers follow, who could be potential targets receiving spam URL postings, as well as the users that follow the spammers (spam followers). This way we have gathered the full network around the spammer seed set. Next, we aimed to retrieve the full network around the spam followers, which is the group of interest in our work, as they help connect the spammers to the rest of the network. To this end, we also retrieved for the spam followers all the other “innocent” users they follow excluding the spammers themselves. In addition, we included the entire followers’ network of spam followers. This process led to a graph containing 303,999 unique users and 1,002,316 links; the sizes of the respective groups are shown above. The graph constructed is directed, based on the “following” relationship of Twitter, and unweighted, as there is no natural measure of relationship strength in Twitter followers.

The sampling we performed on the original dataset to acquire our network of 303,999 users was based on the notion that in a closed interconnected network spammers and honest users follow distinguishably different patterns. Identifying these patterns would prove helpful in the construction of a prediction framework to distinguish potentially dangerous users (e.g. spammers) in an OSN based on the topological structural alone. Therefore, starting from the seed set of randomly chosen spammers (500 nodes) we acquired the entire network around them, meaning their direct links (followers and followed by) as well as the entire network of their followers. This way the behavioral patterns of spam followers may also be studied and the communities that help link spammers to the core of honest users can be identified. The following table contains the average numbers of links (incoming and outgoing) for the two basic groups of our sampled graph.

	Average number of Followers	**Average number of users followed by the group**
Spammers	163	886
Spam Followers	220	1112

Table 3. Average number of links (Incoming and Outgoing)

3.3.Dataset for Cyber bullying and Hate Speech detection (Twitter)

In order to create the dataset for early malicious activity detection we build upon two Twitter datasets collected using the Twitter Streaming API between June and August 2016.

The first dataset is a baseline of 1M random tweets. The second one is a hate-related set of 650k tweets collected based on 309 hashtags associated with bullying and hateful speech. To collect the 309 hashtags at first we obtained a 1% sample of public tweets from June to August 2016. Each tweet is stored in JSON format and contains information such as the text that was posted, profile information about the user who posted it, and the client application it was posted from. From this dataset, we extracted tweets that were likely to contain controversial content (e.g., politics, gender issues, racism), with the hope of observing bullying messages and hate speech in them. To get this set of tweets, we parsed the dataset to select all tweets containing #GamerGate, as it is one of the most well documented large-scale instances of bullying/aggressive behaviour. It stemmed from alleged improprieties in video game journalism which quickly grew into a larger campaign centered on sexism and social justice. Concerning the random set of tweets, it serves as a baseline as it is less prone to contain any kind of offensive behaviours.

To further filter the tweets and extract those that contain hate speech and other bullying content we will apply topic extraction with Latent Dirichlet Allocation (LDA), combined with the use of HateBase, an API for detecting hate speech words.

3.4. Dataset for Abusive Behaviour and Hate Speech Detection (4Chan.org)

Over the past decade, 4chan.org has emerged as one of the most impactful generators of online culture. 4chan is an imageboard site, built around a typical bulletin-board model. An “original poster" (OP) creates a new thread by making a post, with one single image attached, to a board with a particular interest focus. Other users can reply, with or without images, and possibly add references to previous posts, quote text, etc. Some of 4chan's most important aspects are anonymity (there is no identity associated with posts) and ephemerality (threads are periodically pruned).

4chan is generally considered a highly influential ecosystem, as not only has it given birth to significant chunks of Internet culture and memes, but also provided a highly visible platform to movements like Anonymous and the alt-right ideology. Although it has also enabled to positive actions, such as catching animal abusers, 4chan is one of the darkest corners of the Internet, featuring a high rate of hate speech, porn, trolling, and even murder confessions. Moreover, it often acts as a platform to coordinate distributed denial of service attacks as well as aggressive actions on other sites.

Despite its influence and frequent media attention, 4chan remains largely unstudied. To the best of our knowledge, there has been very little work providing a systematic analysis of its ecosystem. We thus set to understand the currently "hot" community in terms of hate speech and abusive behavior, namely, /pol/, i.e., 4chan's "Politically Incorrect" board. To some extent, /pol/ is considered a containment board, allowing users to discuss generally distasteful content – even by 4chan standards – without disturbing the operations of other boards, with many of its posters subscribing to the "alt-right" movement and exhibiting characteristics of xenophobia, social conservatism, racism, and, generally speaking, hate.

3.4.1. Dataset

On June 30, 2016, we started crawling 4chan using its JSON API.4 We retrieve /pol/'s thread catalog every 5 minutes and compare the threads that are currently live to those in the previously obtained catalog, then, for each thread that has been purged, we retrieve a full copy from 4chan's archive, which allows us to obtain the full/final contents of a thread.

For each post in a thread, the API returns, among other things, the post's number, its author (e.g., "Anonymous"), timestamp, and contents of the post (escaped HTML). Although our crawler does not save images, the API also includes image metadata, e.g., the name the image is uploaded with, dimensions (width and height), file size, and an MD5 hash of the image.

On August 6, 2016 we also started crawling /sp/, 4chan's sports boards, and on August 10, 2016 /int/, the international board. Table N provides a high level overview of our datasets as of Sept. 12, 2016. We note that for about 6% of the threads, the crawler gets a 404 error: from a manual inspection, it seems that this is due to "janitors" (i.e., moderators) removing threads for violating rules.

We additionally crawl any YouTube comments that are linked in posts.

	/pol/	/sp/	/int/	Total
Threads	216,783	14,402	24,873	256,058
Posts	8,284,823	1,189,736	1,418,566	10,893,125

Table 4. Summary of our 4Chan data through September 12, 2016

3.4.2. Hate Speech

Hate speech. /pol/ is generally considered a “hateful” ecosystem; however, quantifying hate is a non-trivial task. One possible approach is to perform sentiment analysis over the posts in order to identify positive vs negative attitude, but this is difficult since the majority of /pol/ posts (about 84%) are either neutral or negative. As a consequence, to identify hateful posts, we use the hatebase dictionary, a crowdsourced list of more than 1,000 terms from around the world that indicate hate when referring to a third person. We also use the NLTK framework to identify these words in various forms (e.g., “retard” vs “retarded”).

Our dictionary-based approach identifies posts that contain hateful terms, but there might be cases where the context might not exactly be “hateful” (e.g., ironic usage). Moreover, hatebase is a crowdsourced database, and is not perfect. To this end, we manually examine the list and remove a few of the words that are clearly ambiguous or extremely context- sensitive (e.g., “india” is a variant of “indio,” used in Mexico to refer to someone of Afro-Mexican origin, so can often be confused with the country India). Nevertheless, given the nature of /pol/, the vast majority of posts are more likely to use these terms in a hateful manner.

Despite these caveats, we can use this approach to provide an idea of how prevalent hate speech is on /pol/. We find that 12% of /pol/ posts contain hateful terms, which is significantly higher than in /sp/ (6.3%) and /int/ (7.3%). In comparison, analyzing a random sample of tweets reveals that it contains only 2.18% of hateful tweets. In Figure N, we also report the percentage of /pol/ posts in which the top 15 most “popular” hate words from the hatebase dictionary appear. “Nigger” is the most popular hate word, used in more than 2% of posts, while “faggot” and “retard” appear in over 1% of posts as well. To get an idea of the hate magnitude, consider that “nigger” appears in 265K posts, i.e., in about 120 posts an hour. After the top 3 hate words, there is a sharp drop in usage, although we see a variety of slurs. These include “goy,” which is a derogatory word used by Jewish people to refer to non-Jewish people. In our experience, however, we note that “goy” is used in an inverted fashion on /pol/, i.e., posters call other posters “goys” to imply that they are submitting to Jewish “manipulation” and “trickery.”

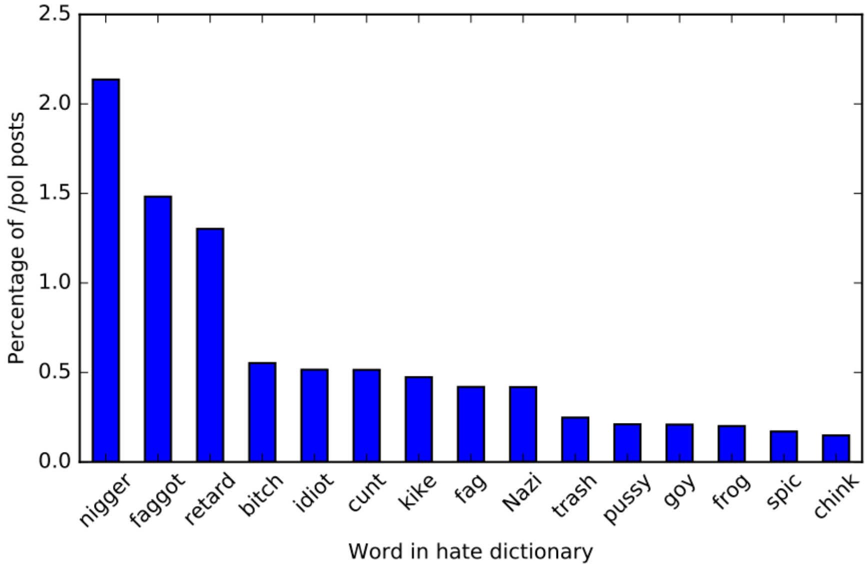


Figure 4. The top 15 most popular hate words on /pol/

Using the flag icons associated with /pol/ posts as a proxy for user location, Figure 5 plots the usage of hate speech per country.

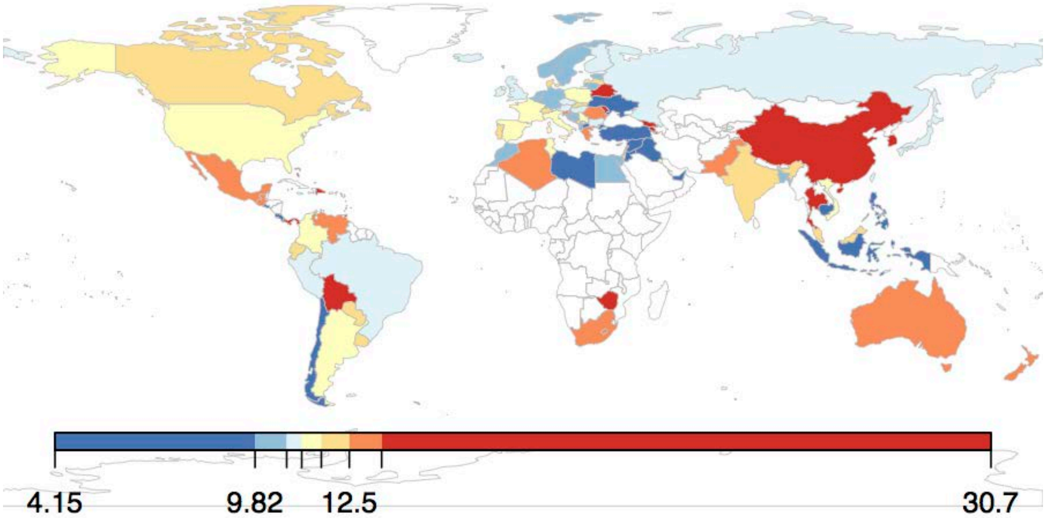


Figure 5. Hate speech usage per country (% posts with at least one hate word)

Overall, our findings with respect to hate indicate that /pol/ is an “excellent” source of data for ground truth and understanding.

3.4.3. Raids (Abusive behaviour)

/pol/ is often used to post links to other sites: some are posted as commentary to the discussion, but others often serve to call /pol/ users on certain coordinated actions, including attempts to skew post-debate polls as well as “raids”. Broadly speaking, a raid is an attempt to disrupt another site, not from a network perspective (as in a DDoS attack), but from a content point of view, aiming to disrupt the community operating on that service. Raids on /pol/ are semi-organized: we observe a number of calls for action consisting of a link to a target – e.g., a Youtube video or a Twitter hashtag – and the text “you know what to do”, prompting other 4chan users to start harassing the target. We also observe that the thread itself often becomes an aggregation point with screenshots of the target’s reaction, sharing of sock pup- pet accounts used to harass, etc. Unlike 4chan’s earliest days, raids are now strictly prohibited, and special mention is made on /pol/’s rules as well, however, we have found evidence they still occur.

We studied how raids work on /pol/. We start with a case study of a very broad-target raid, attempting to mess with anti- trolling tools by substituting racially charged words with company names, e.g., “googles.” Next, we find large scale-evidence of raids and describe an algorithm to detect raids taking place.

“Operation Google”

On September 22, 2016, a thread on /pol/ called for the execution of so-called “Operation Google,” in response to Google announcing the introduction of anti-trolling machine learning based technology and similar initiatives on Twitter. It was proposed to poison these by using, e.g., “Google” instead of “nigger” and “Skype” for “kike,” calling other users to disrupt social media sites like Twitter, and also recommending using certain hashtags, e.g., #worthlessgoogs and #googlehangout. By examining the impact of Operation Google on both /pol/ and Twitter, we aim to gain useful insight into just how efficient and effective the /pol/ community is in acting in a coordinated manner.

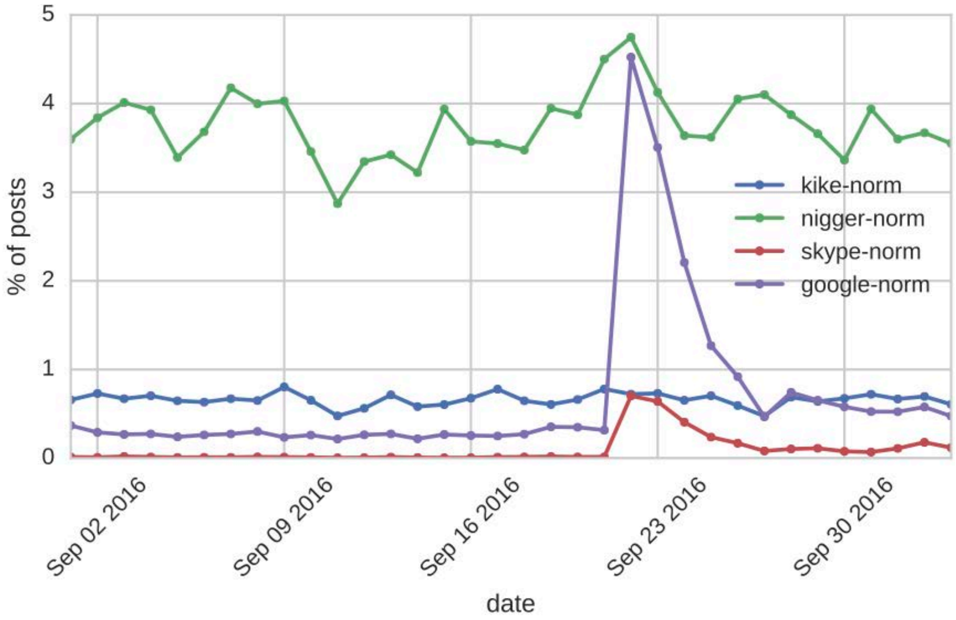


Figure 6. Normalized usage of Operation Google word replacements on 4Chan, over time

In Figure 6, we plot the normalized usage of the specific replacements called for in the Operation Google post. The effects within /pol/ are quite evident: on Sep 22 we see the word "google" appearing at over 5 times its normal rate, while "Skype" appears at almost double its normal rate. To some extent, this illustrates how quickly /pol/ can execute on a raid, but also how short of an attention span its users have: by Sep 26 the burst in usage of Google and Skype had died down. While we still see elevated usages of "Google" and "Skype," there is no discernible change in the usage of "nigger" or "kike," but these replacement words do seem to have become part of /pol/'s vernacular.

Next, we investigated the effects of Operation Google outside of /pol/, counting how many tweets in our 60M tweet dataset (see Section 4) contain the hashtags #worthlessgoogs, #googlehangout, #googleriots, #googlesgonnagoog, and #dumbgoogles. (Recall that our dataset consists of a 1% sample of all public tweets from Sep 18 to Oct 5, 2016.) As expected, the first instances of those hashtags, specifically, #googleriots and #dumbgoogles, appear on Sep 22. On Sep 23, we also see #worthlessgoogs and, on later days, the rest of the hashtags. Overall, Sep 23 features the highest hashtag activity during our observation period. While this does indicate an attempt to instigate censorship evasion on Twitter, the percentage of tweets containing these hashtags shows that Operation Google's impact was much more prevalent on /pol/ itself than on Twitter. For example, on Sep. 23, #dumbgoogles appears in only 5 out 3M tweets (0.00016%) in our dataset for that day, despite it being the most "popular" hashtag (among the ones involved in Operation Google) on the most "active" day. Incidentally, this is somewhat at odds with the level of media coverage around Operation Google.

"YouTube Comments"

We examine the comments from 19,568 YouTube videos linked to by 10,809 /pol/ threads to look for raiding behavior at scale. Note that finding evidence of raids on YouTube (or any other service) is not an easy task, considering that explicit calls for raids are an offense that can get users banned. Therefore, rather than looking for a particular trigger on /pol/, we look for elevated activity in comments to YouTube videos linked in /pol/. In a nutshell, we expect raids to be exhibited by synchronized activity between a YouTube link appearing on /pol/ and the amount of comments it receives on YouTube. We also expect the rate of hateful comments to increase after a link is posted on /pol/.

To model synchronized activities, we use signal processing techniques. In a nutshell, we use cross correlation to compute the “lag” between a /pol/ thread and the YouTube comments it links to. Our hypothesis is that the more synchronized the two comment threads are, the more likely a raid is happening.

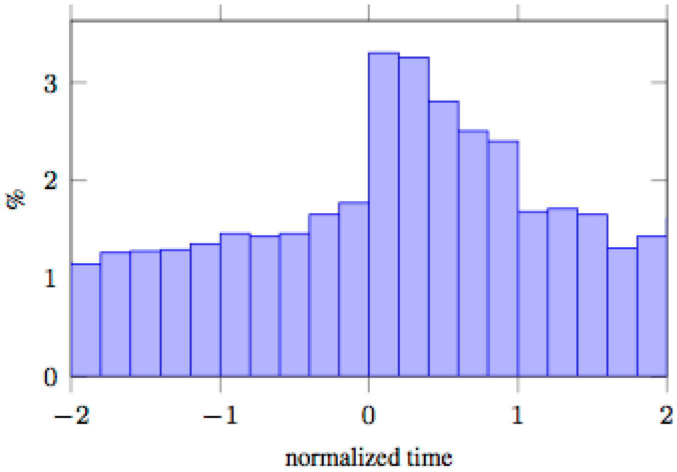


Figure 7. Distribution of activity peaks on YouTube comments linked from /pol/

Figure 7 depicts the distribution of the highest activity peak on YouTube comments linked from /pol/. Note that the distribution is centered around when the /pol/ thread that the YouTube comments were posted in was still “live.” Also, t=0 denotes the time when the video was first linked on /pol/ and t=1 represents the time of the last comment on /pol/ before the thread died.

Just because there is activity on YouTube that coincides with a video being linked on /pol/ does not provide evidence for raids, however. Keeping in mind our hypothesis, we next measured the rate of hate comments on YouTube (defined as a comment having at least one hate word from the hatebase dictionary) as a function of the synchronization lag with the /pol/ thread they were linked on. I.e., is evidence of synchronization correlated with a higher rate of hate speech, indicating /pol/ users were actively posting on YouTube as well? Figure 8 plots the distribution providing ample evidence that yes, /pol/ users are raiding YouTube and posting hateful comments to disrupt YouTube.

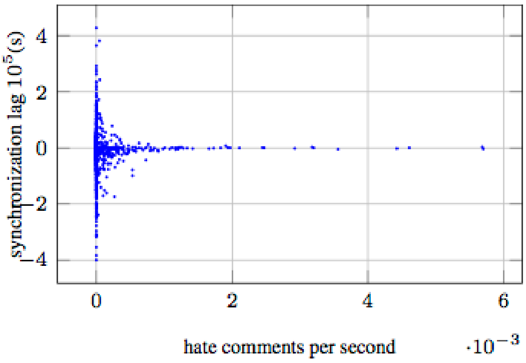


Figure 8. Hate comments per second on YouTube as a function of the synchronization lag with the /pol/ thread they were linked from.

4. Use Cases

In the context of ENCASE, we aim to design and implement a platform that will be able to protect minors and inform their parents when their children face the following:

1. Malicious behaviour, cyberbullying, and sexual grooming
2. False information dissemination, fake identity and activity detection
3. Sensitive content detection and protection

Below, we present indicative use cases for each one of the aforementioned threats, and we also present some use cases for understanding the risks undertaken in the use of social networks for educational purposes.

4.1. Use Case A – Malicious Behaviour Detection

4.1.1. Use Case purpose

The purpose of malicious behaviour detection is to detect and protect minors and especially children from cyber bullies and sexual predators in social networks. OSNs allow sexual predators to contact minors and perform sexual cyber grooming with the aim of sexual advancement. Cyber bullying in OSNs can be done either by a friend to another friend or via random associated mentions. Additionally, another goal is to detect and inform parents when their children are experiencing or are about to experience distressed or aggressive behaviour as a result of being victims of malicious behaviour. Informing parents for such incidences can be very crucial and can prevent undesired incidences like depressions or even worst child suicides.

4.1.2. Scenario 1: Friend to friend cyber bullying detection

Code number	A.1
Name	Friend to friend cyber bullying detection
Author/Partner	CUT, CYRIC
Stakeholders	Marios is a student (age: 13) at a high school in Limassol. He enjoys surfing the web and he has several social network accounts. Melanie is Mario’s mother. She is familiarised with social media and is a Facebook friend with her son. Phanos and Tim are also Mario’s fellow students (age 13). They have created a Facebook group chat with Marios for discussing topics relevant to their school subjects.
High-level Description	Marios is a high school student in Limassol, Cyprus. He enjoys surfing the web. He has several social network accounts since he enjoys chatting with his friends

	and family. Phanos and Tim started making fun of Mario's excessive weight on Facebook. It all started as a joke but Marios couldn't handle being insulted in such a way. He was already trying to lose weight but has not been successful. After several attempts to get them to stop, the jokes continued and Marios became depressed as a result. Mario's mother started worrying about her son being exposed to bullying behaviour online but he refused to discuss it. Melanie found out about **ENCASE** and its ability to detect online malicious behaviour and cyber bullying attacks. Melanie decided to enable **ENCASE** on her son's laptop. Phanos and Tim continued insulting Marios and **ENCASE** immediately detected and notified Melanie, informing her about her son being a victim of cyber bullying. Melanie now is able to take action and discuss with her son how they should handle this type of attack.
Issues	-
Benefits	The **ENCASE** platform enables the parent and the child to be notified on possible cyber bullying so that they are able to take immediate actions.
Notes	-
Services	Web-proxy, Middleware, Data analytics software stack (Back-end)

4.1.3. Scenario 2: Threatening messages cyber bullying detection

Code number	A.2
Name	Threatening messages cyber bullying detection
Author/Partner	CUT
Stakeholders	Serena (age: 14) is a high school student in Barcelona. She has accounts on Facebook and Twitter. She is very outgoing; she enjoys meeting new people especially through social networks. Daniel (age: 17) is also a high school student, who goes to a school near Serena's school. He is very active in posting photos and videos on Facebook and likes meeting people.
High-level Description	Serena, being very active in social networks, she recently attended a seminar at her school related to "Cyber grooming in online social networks". She found out about **ENCASE** and she decided to give it a try. She enabled **ENCASE** on her mobile and laptop browser. Daniel is now browsing through his Facebook account and amongst the suggested friends is Serena and he decides to send her a friend request. Serena, after seeing Daniel's photos, she finds him interesting and she decided to accept the request. As soon as she accepts he drops her friendly message in her inbox. They found each other living near-by and having similar interests and activities. Serena was happy as she found someone so mature and interesting to talk with, yet Daniel had other plans. After a few days of exchanging messages, Daniel asked her to get into a relationship. Since she liked Daniel, she accepts. Everything was fine for the first two weeks but suddenly Daniel sends Serena a message requesting from her to break up. Serena declined and asks persistently from Daniel the reason for that. Instead of giving a reason for his decision, Daniel threatens her that if she sends

	him a message again he will publicly post to Facebook all the messages that Serena sent him during their relationship. At this time, **ENCASE** detected a possible cyber bullying and gives Serena an alert requesting her to stop communicating with this person as he may be a possible cyber bully. At the same time a notification is sent to Serena's mother informing her about the possibility her child is being a victim of cyber bullying. Then, Serena's mother was able to talk with Serena and take the appropriate measures to prevent further harassment of her child.
Issues	-
Benefits	ENCASE add-on protected Serena from being a victim of cyber bullying and Serena's mother was timely informed and able to take the appropriate measures.
Notes	-
Services	Web-proxy, Middleware, Data analytics software stack (Back-end)

4.1.4. Scenario 3: Random associated mentions cyber bullying detection

Code number	A.3
Name	Random associated mentions cyber bullying detection
Author/Partner	CUT, CYRIC
Stakeholders	Ryan (age: 14) is a high school student in Barcelona. He has accounts on Facebook and Twitter. He is a very shy boy and he has some socialization issues because of his weight. Because of that he prefers meeting new people through social networks. Daniel and Brandon (age: 17) are also high school students, who go to the same school as Ryan.
High-level Description	Ryan, a high school student, has some socialization issues due to his weight. Because of that he prefers to meet new people through social networks. Since he is very active in social networks, he recently attended a seminar at her school related to "Cyber bullying in online social networks". During this seminar he found out about **ENCASE** and he decided to give it a try. He talked with his parents and they enabled **ENCASE** on their devices. Daniel and Brandon are close friends going to the same school as Ryan. They are both familiar with OSNs and they used to making fun of people through them. They decided to make fun of Ryan through Facebook by sending him personal messages from a fake account. They created this account pretending a 16 years old girl who goes to a school near their school. All started when Ryan accepted the friend request from the fake account that Daniel and Brandon created. They started sending him personal messages trying to acquire as much information about him as they could. Since this "girl" was one of the few girls that showed interest about Ryan, he started sharing his thoughts with her. A few days later, when they acquired enough information about Ryan the two guys decided to start posting them on Facebook newsfeed from their real Facebook account mentioning Ryan. When they started posting, ENCASE detected possible cyber bullying and notified Ryan and his parents.

	Then, Ryan’s parents were able to take actions against the two bullies and prevent any unwanted incidence.
Issues	-
Benefits	-
Notes	-
Services	Web-proxy, Middleware, Data analytics software stack (Back-end)

4.1.5. Scenario 4: Detection and report of distressed behaviour

Code number	A.4
Name	Detection of distressed behaviour
Author/Partner	CUT
Stakeholders	Marios (age: 13) is a high school student in Limassol. He enjoys surfing the web. He has several social network accounts since he enjoys chatting with his friends and family through Facebook. Melanie is Mario’s mother. Melanie is familiarized with online social networks and a Facebook friend with her son. Phanos (age: 13) is Mario’s fellow student. He is also Mario’s best friend. They both have a Facebook account and they use to chat from their discussing everything.
High-level Description	Marios is a high school student in Limassol, Cyprus. He enjoys surfing the web and he enjoys chatting with people and especially with his best friend Phanos through Facebook. Recently Marios has been a victim of bullying by two other students at his school. They started making fun of Mario’s excessive weight. It all started as a joke and he told them to stop but they continued and Marios could not afford being insulted in such a way. As a result Marios got depressed and his mother started worrying about her son’s behaviour. After searching the web, Melanie found out about **ENCASE** and its malicious behaviour detection functionality and that it was able to detect when a minor is experiencing distressed behaviour. Since she knew that her son used to share everything with his best friend through Facebook chat, ENCASE was her best chance to find out if her son’s behaviour was something more than just stress. She decided to enable it on her son’s laptop and on her own devices. A day after, while chatting with Phanos on Facebook, Marios told him that he cannot afford bullying at school anymore and that he wanted to set an end to his life. Immediately, **ENCASE** detected a distressed behaviour and Melanie received a notification, informing her about her son’s distressed behaviour. Melanie now is able to take action and discuss with her son how they should handle this situation.
Issues	-
Benefits	**ENCASE** enables the parent to be notified on possible situations where its child is experiencing distressed behaviour and take immediate action for protecting its child.
Notes	-
Services	Middleware, Data analytics software stack (Back-end)

4.1.6. Scenario 5: Bad reputation for cyber bullying

Code number	A.5
Name	Bad reputation for cyber bullying
Author/Partner	CUT
Stakeholders	Serena (age: 14) is a high school student in Barcelona. She enjoys social media activity and has accounts on various social networks. Daniel (age: 20) also lives in Barcelona and is currently unemployed. He is very active in social networks and has been recently voted by a lot of ENCASE users as a possible cyber bully due to demonstrating malicious behaviour towards other users.
High-level Description	Serena is a 14 year old high school student Barcelona. She enjoys meeting new people especially through Facebook. Being so active in social networks, Serena’s mother enabled **ENCASE** on her laptop. Serena is now browsing through her Facebook account, and receives a friends request from Daniel Stringini. She likes his profile and decides to accept his request. As soon as she accepts, Serena receives a notification that the specific profile has bad reputation within the **ENCASE ecosystem** as a cyber bully or predator. Serena decided to delete Daniel from her Facebook friends and informed her friends and family about the incident.
Issues	-
Benefits	Serena was timely informed on Daniel’s previous malicious behaviour before he got the opportunity to attack her.
Notes	We will employ sophisticated reputation mechanisms to assess the veracity and the weight of each voter and their tags.
Services	Middleware, Data analytics software stack (Back-end)

4.1.7. Scenario 6: Sexual cyber grooming

Code number	A.6
Name	Sexual cyber grooming
Author/Partner	CUT, CYRIC
Stakeholders	Alice (age: 13) is a high school student in London. She recently created an account on Facebook to talk with her friends. Daniel (age: 17) is a high school student in Barcelona. He is very active in posting photos and videos on Facebook and meeting people.
High-level Description	Alice is a high school student in London. All of her friends had a Facebook account and she didn’t because of her parents being so worried of the threats exist in social networks. The last few months Alice was persistently asking from her parents to allow her create an account so that she is able to chat with her friends. In the end they given their permission and Alice created an account on Facebook but in order to be able to protect their child in case of something bad happen they enabled **ENCASE** on Alice laptop and on their devices. After three months Alice received a friend request from a boy called Daniel. After looking on his photos, she found him handsome and she decided to accept his request. A few days later Daniel sent her a message and they started chatting. In the beginning, Daniel showed a lot of interest on Alice and he would listen to all of her problems trying to make her feel good about herself. At some point Alice thought that he cared about her and she did some sexual stuff over camera for

	him as he initially requested. Immediately, **ENCASE** gave Alicia its first yellow notification warning her that there is a danger for sexual cyber grooming but she ignore it as she enjoyed being admired. Then Daniel turned nasty and started threatening her because he wanted to do and send more. At this point, ENCASE detected a possible sexual cyber grooming and sent a second red alert notification directly to Alice’s parents informing them about the incidence along with a detailed report of the conversation.
Issues	-
Benefits	ENCASE protected Alice from being exposed to sexual cyber grooming by a stranger.
Notes	Levels of alert: 1. Yellow alert: the victim should receive this type of alert when the probability for sexual grooming is below a threshold. 2. Red alert: the victim’s parents should receive this type of alert when the probability for sexual grooming is above a threshold.
Services	Middleware, Data analytics software stack (Back-end)

4.1.8. Scenario 7: Sexual advancement as a result of sexual cyber grooming using fake identity

Code number	A.7
Name	Sexual advancement as a result of sexual cyber grooming using fake identity
Author/Partner	CUT, CYRIC
Stakeholders	Andrea (age: 14) is a high school student in Manchester. After their parents’ permission she created on Facebook that uses to communicate with her friends and to meet new people. Archie (age: 38) is a school teacher in Liverpool. He is familiar with social networks and very active especially on Facebook.
High-level Description	Andrea is a high school student in Manchester. She uses her newly created Facebook account to communicate with her friends and to meet new people. To be able to protect Andrea, her parents recently enabled ENCASE on her laptop because they were worried about the threats that she may face on Facebook. A few days ago, Andrea received a friend request from a 17 years old boy called Eduard. He was good looking and the photos of him on his profile attracted Andrea’s interest who decided to accept his request. In fact, this profile was created by Archie, a 38 years old school teacher, aiming to draw Andrea’s attention and approach her. As soon as Andrea accepted the request she received a message from Eduard. They started chatting and Eduard showed so much interest to listen to Andrea’s problems. Andrea was very happy since she found someone cared so much about her, yet Eduard (Archie) had other plans. He started approaching her using sexual hints about her body. At this point, **ENCASE** gave its first yellow notification warning her that there is a danger for sexual cyber grooming. Serena ignored the notification as she enjoyed being admired. After a few days Eduard started asking her to meet and fully enjoy each other. Then **ENCASE** gave a second red alert notification requesting her to stop

	communicating with this person as he is probably trying to sexually advanced her. The same notification was sent to Andrea’s parents informing them about the incidence along with a detailed report of the conversation. Being timely notified, Andrea’s parents was able to talk with their daughter and convince her to stop communicating and delete Eduard from her Facebook friends.
Issues	-
Benefits	ENCASE protected Alice from being exposed to sexual advancement by a stranger who lied about his identity.
Notes	Levels of alert: 1. Yellow alert: the victim should receive this type of alert when the probability for sexual cyber grooming is below a threshold. 2. Red alert: the victim’s parents should receive this type of alert when the probability for sexual cyber grooming is above a threshold.
Services	Middleware, Data analytics software stack (Back-end)

4.2. Use Case B – False Information Dissemination and Fake Identity Detection

4.2.1. Use case purpose

The purpose of false information dissemination and fake identity detection is: a) to detect and protect minors and especially children by notifying them when they are communicating with malicious actors who pretend to be someone else using fake identities in social networks; b) to detect and alert minors when they are communicating with someone who has fraudulent and/or fake activity in social networks; and c) to detect and notify minors when someone is spreading false information about them in social networks.

4.2.2. Scenario 1: Fake identity and activity detection

Code number	B.1
Name	Fake identity and activity detection
Author/Partner	CUT
Stakeholders	Serena (age: 14) is a high school student in Barcelona. She enjoys social media activity and has accounts on various Social Networks. She enjoys meeting new people especially through Facebook. Daniel is 35 year old, working as a taxi driver in Barcelona.
High-level Description	Serena is a high school student who lives in Barcelona. She is very outgoing and she enjoys meeting new people. Being so active in social networks, Serena’s mother was worried about the threats that her daughter may encounter and she decided to enable **ENCASE** on her laptop. Serena is now browsing through her Facebook account, and receives a friend request from a guy called “Jonathan Stringini”. Daniel has created a fake profile using this nickname on Facebook with a fake profile picture pretending a 19 years old guy working as a barman in a well-known club in Barcelona. She liked his profile and decided to accept his request. As soon as she accepted, Daniel dropped a private message in her inbox. She replied to the message and they started chatting for a while. Meanwhile ENCASE analysed Jonathan’s Facebook

	profile and detected that it was a fake account and sent a notification to Serena informing her that she is communicating with a user that has fake identity and advised her to stop communicating with this person and ignore his messages. Then, Serena deleted Jonathan from her Facebook friends and informed her friends and family about the incident.
Issues	-
Benefits	Serena was timely informed on the fake identity of Jonathan and protected herself from fake activity or malicious behaviour.
Notes	-
Services	Middleware

4.2.3. Scenario 2: False information dissemination detection

Code number	B.2
Name	False information dissemination detection
Author/Partner	CUT
Stakeholders	Andreas (age: 28) is a guy from Cyprus who studied Computer Engineering at Cyprus University of Technology (CUT). He is interested in entrepreneurship and he enjoys social media activity. Peter (age: 34) who is also from Cyprus, is a journalist currently working for an online news media company.
High-level Description	Andreas is a 28 years old guy from Cyprus. He is very active in social networks and especially Facebook. He studied Computer Engineering and he worked as a junior developer for a foreign company that operates in Cyprus. Because of his studies he likes to try every new product and as soon as he heard about **ENCASE** he enabled it on his laptop. Besides that he is also very interested in entrepreneurship and he spent the past two years trying to build his own start-up company related his studies with some of his fellow students from CUT. He recently attended and won a start-up competition and their start-up idea gain a lot of popularity. Peter is a journalist who works for an online news media company located in Cyprus. As soon as he heard about the competition that Andreas’ start-up company won he decided to write an article and publish it on Facebook. However, Peter’s intentions were not good because the start-up company that won the second prize in this competition is owned by a friend of him. Peter wrote an article and published it on Facebook, in which he was trying to discredit Peter claiming that he was unable to run a start-up company and that he didn’t even got a bachelor degree on Computer Engineering. After a while, ENCASE detected this article on Facebook News Feeds and immediately sent a notification to Andreas informing him that the Facebook user “Peter Solomou” is spreading false information about him on Facebook. Andreas decided to communicate with the news media company that Peter was working to report the incident and request a replacement of the article in order to recover his reputation.
Issues	-
Benefits	Andreas was timely informed from **ENCASE** about the false information spread

	about him on Facebook and he was able to take measures in order to recover his reputation.
Notes	-
Services	Middleware, Data Analytics Software Stack (Back-End)

4.2.4. Scenario 3: Detection of false information received by minors

Code number	B.3
Name	Detection of false information received by minors
Author/Partner	CUT
Stakeholders	George (age: 26) works as a journalist in a well-known online media sports news company located in Cyprus. Andreas (age: 17) is a high school student from Cyprus. He enjoys social media activity and has accounts on various Social Networks.
High-level Description	Andreas is currently in the last year of his high school studies. He spends a lot of hours at home in front of his laptop reading and preparing for his final exams. Andreas also enjoys social media activity and especially Facebook. Besides that he also likes football and he used to spend some time reading sports news posted from online media sports news pages on Facebook. Being so active, Andreas recently enabled **ENCASE** on his laptop in order to be able to identify and avoid malicious users with fake identities and to be notified of any false information. George is currently working in a well-known online media sports news company which mostly operates on Facebook. He is responsible for editing and posting most of the company's articles on its Facebook page. The last few days George's boss reprimanded him because of the reduction of their page traffic. George in order to increase their page traffic he started posting articles on Facebook with false content using Clickbait title and thumbnail. In most of the cases such titles and thumbnails draw the users' attention to click on the post and read the content of the article. Yesterday, while Andreas was browsing through his Facebook news feeds he saw a post posted by George. The title draws his attention but at the same time **ENCASE** notified him to ignore this post since probably it is a source of false information (Clickbait). Then Andreas was able to ignore it and avoid clicking on a post that has false content.
Issues	-
Benefits	Using **ENCASE** Andreas gets notified when he receives false information of any kind. As a result Andreas is able to decide whether he wants to ignore such information and be sure that he will always read valid information on social networks.
Notes	-
Services	Middleware, Data Analytics Software Stack (Back-End)

4.3. Use Case C – Sensitive Content Detection and Protection

4.3.1. Use case purpose

The purpose of sensitive content detection and protection is to detect any sensitive content that users are about to share with inappropriate audience or any sensitive content that the user is about to receive. Additionally, when sensitive content is detect the ENCASE enables users to protect their content by offering them multiple ways to do so and only share it with the people they are willing to. For the protection of the users’ content steganography, group encryption and attribute-based encryption techniques are used.

4.3.2. Scenario 1: Detection and protection of sensitive photos in OSNs

Code number	C.1
Name	Detection and protection of sensitive photos in OSNs
Author/Partner	CUT, CYRIC
Stakeholders	Serena (age: 16) is a high school student in Barcelona. She enjoys social media activity and has accounts on various Social Networks.
High-level Description	Serena enjoys using Facebook to chat with her friends and she used to share a lot of photos for sharing her photos with them. Being so active in photo-sharing, her mother Marie decided to enable **ENCASE** on her personal devices to protect her. This year Serena visited Mallorca and took a lot of photo of herself and with her family. At some time she tried to upload and set one of her photos as her profile picture, but she received the following notification from **ENCASE**: This photo cannot be uploaded on your profile as it contains more than 80% nude content. Having seen the message, Serena decided to privately share the photo to her friends. Expect the notification, **ENCASE** informed her of being able to share her photo using steganography or the following cryptographic techniques: group encryption and attribute-based encryption. As a result, Serena decided to use Group encryption to protect her photo and ensure that only her friends will be able to access it.
Issues	-
Benefits	ENCASE sensitive content detection and protection functionality protected Serena from publicly sharing a photo that contains nudity on Facebook, providing her with alternatives in sharing the photo properly.
Notes	-
Services	Sensitive content detection and protection, Web-Proxy Server, Middleware

4.3.3. Scenario 2: Detection and protection of sensitive information in OSNs

Code number	C.2
Name	Detection and protection of sensitive information in OSNs
Author/Partner	CUT
Stakeholders	Serena (age: 14) is a high school student in Barcelona. She enjoys social media activity and has accounts on various Social Networks.
High-level Description	Serena enjoys using Facebook for discussing and meeting new friends. Additionally, she used to share her personal information through chat messages or posts on Social Networks and she is unaware of the dangers behind sharing such type of information online. Serena recently participated in a seminar related to e-safety and was informed that sharing this type of data could end in being available to inappropriate

	audience. She was also informed about **ENCASE** sensitive content detection and protection functionality and she decided to enable it on her device. A few days, later while she was trying to publicly share her phone number to her new friend's profile on Facebook, Serena received a notification from **ENCASE** warning her that is not safe to publicly share this type of information online. As a result, Serena changed her mind and decided to share her phone number with her friend using Facebook's private messages.
Issues	-
Benefits	**ENCASE** sensitive content detection and protection functionality protected Serena from publicly sharing personal information that should not be shared with inappropriate audience, providing her with alternatives in sharing the information properly and to her preferred persons.
Notes	-
Services	Sensitive content detection and protection, Web-Proxy Server, Middleware

4.3.4. Scenario 3: Secure sharing of sensitive content in OSNs

Code number	C.3
Name	Secure sharing of sensitive content in OSNs
Author/Partner	CUT
Stakeholders	Serena (age: 18) is a high school student in Barcelona. She enjoys social media activity and has accounts on various Social Networks. She used to share her personal information through chat messages or posts on Social Networks. Marie is Serena's mother and she is .
High-level Description	Serena is very active in social networks and especially Facebook and she use share her personal information through chat messages or public posts. Being so active in photo-sharing, her mother Marie was concerned about her daughter's privacy and she decided to enable **ENCASE** on her personal devices due the sensitive content detection and protection functionality that it offers. Additionally, after a discussion she had with her daughter Serena accepted to use **ENCASE** to securely share photos with her friends. Today Serena is at the beach and she is capturing photos of herself that she wants to share on Facebook with her friends. She is trying to publicly post one of those photos and tag her friends on it, but she immediately received the following notification from ENCASE: This photo should not be publicly shared as it contains more than 80% nude content. The ENCASE add-on also informed her of being able to securely share the photo using steganography, group encryption or attribute-based encryption. Having seen the notification, Serena decided to privately share the photo to her friends using Group Encryption adding her friends in the list of the people who are able to see it. Then, her friends were able to access and see that photo and no other that is not in the list that Serena defined is able to access it.
Issues	-
Benefits	ENCASE sensitive content detection and protection functionality provides Serena with multiple ways for sharing her sensitive photos properly with her preferred persons without being publicly visible to inappropriate audience.
Notes	-
Services	Sensitive content detection and protection, Web-Proxy Server, Middleware

4.4.Use Case D – Educators' Awareness

4.4.1. Use case purpose

The purpose of educator's awareness use case is to raise educators' awareness in understanding the risks undertaken in the use of social media for educational purposes.

4.4.2. Scenario 1: Malicious behaviour detection in educational OSN groups

Code number	D.1
Name	Malicious behaviour detection in educational OSN groups
Author/Partner	CUT
Stakeholders	Mary is a German language instructor in a high school in the UK. She has just completed her studies in German language and culture and her MA in Computer-Assisted Language Learning.
High-level Description	Mary (age: 30) has completed her studies in German language and culture and her MA in Computer-Assisted Language Learning she is appointed as a German language instructor in a high school in United Kingdom. Being a social media savvy, Mary decides to incorporate social media in her teaching methods and creates a Facebook group where her students can use the target language in authentic environments. Yet, Mary accepts negative criticism from her school director for "exposing children in malicious behaviour in social media". Mary is initially insulted by the critique but then decides to find out more about the risks of social media in schools. She came across **ENCASE platform** and she decides to enable it for all of her students so that is able to protect them from malicious behaviours in social networks. In order to do so she requested for the parents' permission and as soon as they agreed she enabled it.
Issues	-
Benefits	Mary is now in place to protect her students from malicious behaviour and demonstrate to her school ways for using safely social media for educational purposes.
Notes	-
Services	Web-Proxy, Middleware, Data Analytics Software Stack (Back-End)

4.4.3. Scenario 2: Fake identity and activity detection in educational OSN groups

Code number	D.2
Name	Fake identity and activity detection in educational OSN groups
Author/Partner	CUT
Stakeholders	Mary (age: 30) is a German language instructor in a high school in United Kingdom. She has just completed her studies in German language and culture and her MA in Computer-Assisted Language Learning. Kevin Kalen has just completed his high school degree and he is currently unemployed. He enjoys surfing in social media and meeting new people.
High-level Description	Mary is a 30 year old German language instructor who works in a high school in United Kingdom. Being a social media savvy, Mary decides to incorporate social media in her teaching methods and creates a Facebook group where her students can use the target language in authentic environments. She creates the group and all of her students join with high level of excitement.

	All of a sudden, a new member appears in the group from someone called Kevin Kalen. Kevin posted that one of the students in the group had passed away and everyone in the group got very upset. Immediately upon posting, Mary received a notification from **ENCASE** that she enabled on her computer. **ENCASE** informed her that the information given by Kevin Kalen is probably fake and that the profile that posted this information was also a fake one. Mary deleted Kevin from the group and informed her students about the incident. She also gave them a short information hand-out on how **ENCASE** can protect them from fake and malicious actions in online social networks.
Issues	-
Benefits	Mary and her class were immediately protected from a fake account.
Notes	-
Services	Middleware, Data Analytics Software Stack (Back-End)

4.4.4. Scenario 3: Sensitive content detection in educational OSN groups

Code number	D.3
Name	Sensitive content detection in educational OSN groups
Author/Partner	CUT
Stakeholders	Mary (age: 30) is a German language instructor in a high school in the UK. She has just completed her studies in German language and culture and her MA in Computer-Assisted Language Learning. Alice is one of Mary’s students. She is very keen in using social media for chatting with her friends and meeting new people.
High-level Description	Mary is a 30 year old German language instructor. Having completed her studies in German language and culture and her MA in Computer-Assisted Language Learning she is working as a German language instructor in a secondary school in United Kingdom. Being a social media savvy, Mary decides to incorporate social media in her teaching methods and created a Facebook group where her students can use the target language in authentic environments. She created the group and all of her students joined with high level of excitement. Peter, one of Mary’s students, tries to post one of his photos from his holidays in Germany during the summer wearing a swimming suite. Immediately, Mary received a notification from ENCASE that sensitive/inappropriate content is being rejected from the group. Peter also notified and informed that 80% nudity is unacceptable and photo will be deleted.
Issues	-
Benefits	-
Notes	-
Services	Sensitive content detection and protection, Web-Proxy Server, Middleware

5. User Stories and Acceptance Criteria

5.1. Use Case A - Malicious Behaviour Detection

5.1.1. Scenario 1: Friend to friend cyber bullying detection

Code number	MBD_FF_1
Title	Notify minor when an OSN friend is performing cyber bullying against him
Description	**As a minor** **I want to** receive a notification when someone of my friends in OSNs is trying to perform cyber bullying against me **so that** I am aware and avoid him/her
Acceptance criteria	1. The ENCASE middleware should analyse the social network data of the user and try to unveil incidences of cyber bullying. 2. The system should first check whether the person who the minor is communicating with is not included in the whitelist that the parent set. 3. The system should detect and send an early notification to the user when there is a suspicion for cyber bullying. 4. A user should receive notification on his browser from the malicious behaviour browser add-on. 5. If the detected malicious user is not in the ENCASE cyber bullies blacklist then he system should flag him in the blacklist for future reference.

Code number	MBD_FF_2
Title	First early notification there is a suspicion for cyber bullying against a minor
Description	**As a minor** **I want to** receive a first early notification when someone of my friends in OSNs is trying to perform cyber bullying against me **so that** I am aware
Acceptance criteria	1. The ENCASE middleware should rarely analyse the social network data of the user and try to unveil incidences of cyber bullying. 2. The system should detect when there is a small probability for cyber bullying and send an early notification to the user. 3. A user should receive notification on his browser from the malicious behaviour browser add-on.

Code number	MBD_FF_3
Title	Notify parent when her child becomes victim of cyber bullying by a friend in OSNs
Description	**As a parent** **I want to** receive a notification when my child has become a victim of cyber bullying by one of its friends in OSNs **so that** I am aware and able to protect my child
Acceptance criteria	1. The ENCASE middleware should rarely analyse the social network data of the user and try to unveil incidences of cyber bullying. 2. The system should send the notification to the parent’s device. 3. The parent should timely receive the notification.

Code number	MBD_FF_4
Title	Flag malicious user as cyber bully
Description	**As a parent** **I want to** be able to flag a malicious user that ENCASE notified me that he

	performed cyber bullying against my child **so that** I prevent him from insulting other minors
Acceptance criteria	1. The ENCASE front-end should allow the parent to flag a malicious user as a cyber bully. 2. The system should add the flagged malicious user in a blacklist that maintains for future reference. 3. ENCASE should prevent the flagged user for communicating with the parent’s child in the future. 4. Each time a user is flagged his bad reputation score should increase.

Code number	MBD_FF_5
Title	Set whitelist for cyber bullying checks
Description	**As a parent** **I want to** be able to set the OSN friends of my child that I trust for cyber bullying **so that** I get notified only for people that I don’t trust to perform cyber bullying on my child
Acceptance criteria	1. The ENCASE front-end should allow the parent to set for each OSN account of her child a whitelist that contains the users that she trust for communicating with her child.

5.1.2. Scenario 2: Threatening messages cyber bullying detection

Code number	MBD_TM_1
Title	Notify parent when her child receives threatening messages
Description	**As a parent** **I want to** receive a notification when someone is sending threatening messages to my child in OSNs **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The ENCASE middleware should analyse the social network data of the user and try to unveil incidences of cyber bullying. 2. The system should first check whether the person who the minor is communicating with is not included in the whitelist that the parent set from before. 3. The system should detect and send a notification to the parent when someone is sending threatening message to its child in OSNs. 4. If the detected malicious user is not in the ENCASE cyber bullies blacklist then t/he system should flag him in the blacklist for future reference.

Code number	MBD_TM_2
Title	Notify child when is becoming a victim of cyber bullying
Description	**As a minor** **I want to** receive a notification when someone is threatening me and there is a possibility for cyber bullying **so that** I am aware and I can avoid him
Acceptance criteria	1. The ENCASE middleware should analyse the social network data of the user and try to unveil incidences of cyber bullying. 2. The system should first check whether the person who the minor is communicating with is not included in the whitelist that the parent set

	from before. 3. The system should detect and send a notification to the minor when someone is sending him threatening messages in OSNs and there is a possibility for cyber bullying. 4. After detecting possible cyber bullying the web-proxy should block any threatening messages sent to the minor. 5. If the detected malicious user is not in the ENCASE cyber bullies blacklist then the system should flag him in the blacklist for future reference.

5.1.3. Scenario 3: Random associated mentions cyber bullying detection

Code number	MBD_RAM_1
Title	Notify minor when is becoming a victim of cyber bullying from associated mentions in OSNs
Description	**As a minor** **I want to** receive a notification when someone is posting sensitive information about me in OSNs in form of cyber bullying **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The ENCASE middleware and/or back-end should analyse the social network data and try to unveil incidences where a malicious user is posting sensitive information about me in OSNS in the form of performing cyber bullying. 2. If the detected malicious user is not in the ENCASE cyber bullies blacklist then the system should flag him in the blacklist for future reference.

Code number	MBD_RAM_2
Title	Notify minor when he is communicating with a possible cyber bully in OSNs
Description	**As a minor** **I want to** receive a notification when I am communicating with a possible cyber bully **so that** I am aware and not share my personal information with him
Acceptance criteria	1. The system should timely notify the minor when he is communicating with an OSN user that has record for cyber bullying. 2. The system should block the minor from sharing personal information with a possible cyber bully. 3. If the detected malicious user is not in the ENCASE cyber bullies blacklist then the system should flag him in the blacklist for future reference.

Code number	MBD_RAM_3
Title	Notify parent when her child is becoming a victim of cyber bullying from associated mentions in OSNs
Description	**As a minor** **I want to** receive a notification when my child is becoming a victim of cyber bullying when someone is posting sensitive information of my child in OSNs **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The ENCASE middleware and/or back-end should analyse the social network data and try to unveil incidences where a malicious user is posting sensitive information about me in OSNS in the form of performing cyber bullying.

	2. When such an incidence is detected ENCASE should block the child from communicating with the malicious user. 3. If the detected malicious user is not in the ENCASE cyber bullies blacklist then the system should flag him in the blacklist for future reference.

Code number	MBD_RAM_4
Title	Block minor from sharing its personal information in OSNs
Description	**As a parent** **I want to** be able to block my child from sharing its personal information with possible cyber bullies in OSNs **so that** I avoid any undesired incidences
Acceptance criteria	1. The ENCASE front-end should allow the parent set whether she is willing or not her child to be blocked when sharing personal information in OSNs. 2. When the child is trying to share its personal information in OSNs the system should block the child if the parent set this as true. 3. When the child is trying to share its personal information in OSNs, If the parent did not set her preference, then the parent should be notified and be able to provide his consent for sharing this information.

5.1.4. Scenario 4: Detection and report of distresses behaviour

Code number	MBD_DB_1
Title	Minor is about to experience distressed or aggressive behaviour
Description	**As a parent** **I want to** get notified when my child is about to experience distressed or aggressive behaviour **so that** i can take measures to prevent any undesired incidents
Acceptance criteria	1. The ENCASE middleware should be able to capture and analyse the user’s OSNs’ activity in order to detect if she/he is about to experience distressed or aggressive behaviour. 2. ENCASE malicious behaviour browser add-on should notify (in visual or in textual way) the parent when her child is about to experience distressed or aggressive behaviour. 3. The malicious behaviour add-on should provide hierarchical notifications based on the percentage of distressed or aggressive behaviour extracted from the analysis. 4. The ENCASE platform should learn the child’s behaviour in OSNs through the time so that in the future will be more accurate and able to detect when the child is about to experience distressed behaviour.

5.1.5. Scenario 5: Bad reputation for cyber bullying

Code number	MBD_BR_1
Title	Minor is communicating with a user that has bad reputation for cyber bullying
Description	**As a minor** **I want to** get notified when I am communicating with someone who has bad reputation for cyber bullying **so that** i can avoid him

Acceptance criteria	1. ENCASE should maintain a reputation list with all the flagged for cyber bullying malicious users. 2. ENCASE malicious behaviour browser add-on should notify (in visual or in textual way) the minor is communicating or is about to communicate with a user that has bad reputation for cyber bullying in ENCASE malicious behaviour reputation list. 3. When the minor receives a friend request from a user that has bad reputation the system should notify the minor and advise him to ignore the request.

5.1.6. Scenario 6: Sexual cyber grooming

Code number	MBD_SCG_1
Title	Early notification for sexual cyber grooming
Description	**As a minor** **I want to** get timely notified when there is a possibility for sexual cyber grooming **so that** i can avoid him
Acceptance criteria	1. ENCASE middleware should analyse the minors social network data and try to detect if there is a possibility for sexual cyber grooming. 2. ENCASE malicious behaviour browser add-on should send an early notification (in visual or in textual way) to the minor when there is a small probability that the person that she is communicating is performing sexual cyber grooming. 3. When the minor receives a friend request from a user that has bad reputation the system should notify the minor and advise him to ignore the request.

Code number	MBD_SCG_2
Title	Bad reputation for sexual cyber grooming
Description	**As a parent** **I want to** get notified when I receive a friend request in OSNs from someone with bad reputation for sexual cyber grooming **so that** I am aware and avoid him
Acceptance criteria	1. The ENCASE web-proxy should capture every new friend request that the minor receives in OSNs. 2. The middleware should be able to retrieve and analyse all the available OSN account details of the user that send a friend request to the minor. 3. The middleware should first check against the sexual grooming blacklist whether the captured account has bad reputation for sexual cyber grooming. 4. If the ENCASE malicious behaviour browser add-on should send an early notification (in visual or in textual way) to the minor if the OSN user has previously exposed malicious behaviour for sexual cyber grooming.

Code number	MBD_SCG_3
Title	Notify parent when her child has become a victim sexual cyber grooming
Description	**As a parent**

	I want to get timely notified when my child is becoming a victim of sexual cyber grooming **so that** i can take the appropriate measures
Acceptance criteria	1. ENCASE middleware should analyse the minors social network data and try to unveil incidences for sexual cyber grooming. 2. The parent should receive a notification when her child is communicating with someone who has bad reputation for cyber grooming. 3. ENCASE malicious behaviour browser add-on should send an early notification (in visual or in textual way) to the parent when there is a probability that the person that her child is communicating is performing sexual cyber grooming.

Code number	MBD_SCG_4
Title	Prevent minor from communicating with someone who has bad reputation for sexual cyber grooming
Description	**As a parent** **I want to** be able to prevent my child from communicating with someone who has bad reputation for sexual cyber grooming **so that** I do not have to worry and my child be safe
Acceptance criteria	1. The ENCASE front-end should allow the parent to choose whether she prefers to block or not any incoming to her child message from someone who has bad reputation for cyber bullying. 2. If the user chooses to block her child from communicating with someone with bad reputation for cyber bullying then ENCASE web-proxy should also block any outgoing message from the child to such users.

Code number	MBD_SCG_5
Title	Flag a malicious user for sexual cyber grooming
Description	**As a parent** **I want to** be able to flag a malicious user for sexual cyber grooming **so that** I prevent him from sexually abusing other minors
Acceptance criteria	1. The ENCASE front-end should allow the parent to flag the users that she is being notified that they performed sexual cyber grooming on her child. 2. Anyone that is flagged as sexual predator is being noted in the system using a reputation score. The reputation score increases each time someone is flagged as sexual predator.

5.2. Use Case B - False Information Dissemination and Fake Identity Detection

5.2.1. Scenario 1: Fake identity and activity detection

Code number	FID_FIA_1
Title	Get informed when communicating with someone with fake identity in OSNs
Description	**As a minor** **I want to** get notified when I am communicating with someone with fake identity in OSNs **so that** I am aware and avoid him

Acceptance criteria	1. ENCASE middleware should analyse the minors' social network data and try to detect if she is communicating with someone who has fake identity. 2. ENCASE fake identity and activity detection browser add-on should send a notification (in visual or in textual way) to the minor when the person who she is communicating has a fake identity. 3. ENCASE middleware should regularly capture the OSN friends lists of the minor and analyse each one of the accounts for detecting fake identities. 4. The system should create a blacklist with all the flagged fake identities for future reference.

Code number	FID_FIA_2
Title	Get informed when befriending someone with fake identity in OSNs
Description	**As a minor** **I want to** get notified when I receive a friend request in OSNs from someone with fake identity **so that** I am aware and delete him
Acceptance criteria	1. ENCASE web-proxy should capture every friend request the minor receives in OSNs. 2. First the middleware should check the OSN profile of the captured request against the whitelist that the parent set from before. 3. Then the ENCASE middleware should check if the captured OSN profile is flagged from before as a fake identity. 4. If the captured profile is not in the whitelist or the fake identity blacklist then the middleware should analyse all the available information of this OSN profile. 5. In case a fake identity is detected, the ENCASE fake identity and activity detection browser add-on should notify the minor that the friend request she received is from a fake identity.

Code number	FID_FIA_3
Title	Report a possible fake identity in OSNs
Description	**As a parent** **I want to** be able to report to ENCASE a possible fake identity in OSNs **so that** I prevent the holder of the fake account from performing any fake activity or malicious behaviour
Acceptance criteria	1. ENCASE front-end should allow the parent to report a possible fake identity in OSNs.

5.2.2. Scenario 2: False information dissemination detection

Code number	FID_FID_1
Title	False information dissemination detection in OSNs
Description	**As a user** **I want to** get notified when someone is spreading false information about me in OSNs **so that** I am aware of it
Acceptance	1. ENCASE middleware should analyse the donated by users social network

criteria	data and try to detect false information about the user. 2. ENCASE fake identity and activity detection browser add-on should send a notification (in visual or in textual way) to the user when it detects false information about him spread in OSNs.

Code number	FID_FID_2
Title	Report false information dissemination
Description	**As a user** **I want to** be able to report false information spread about me in OSNs **so that** such information will be removed
Acceptance criteria	1. ENCASE front-end should offer the user the ability to report incidences when false information about him is spread in OSNs. 2. The system should record the user, who posted the false information, for fake activity.

5.2.3. Scenario 3: Detection of false information received by minors

Code number	FID_FIR_1
Title	Clickbait detection in OSNs
Description	**As a minor** **I want to** get notified when I receive false information in OSNs **so that** I am aware and ignore it
Acceptance criteria	1. ENCASE middleware should analyse the minors’ social network data and try to detect Clickbait post. 2. ENCASE fake identity and activity detection browser add-on should send a notification (in visual or in textual way) to the minor when he sees clickbait information in his OSN news feeds.

Code number	FID_FIR_2
Title	Report false information in OSNs
Description	**As a minor** **I want to** be able to report any type of false information that I see in my OSN news feed **so that** other minors will be notified when they see such information
Acceptance criteria	1. ENCASE fake identity and activity browser add-on should allow the minor to report any type of false information that he sees in his OSN news feed. 2. The system should keep a record of the false information that users reported. 3. In the future, ENCASE should notify other minors when they will receive information reported as false.

5.3. Use Case C - Sensitive Content Detection and Protection

5.3.1. Scenario 1: Detection and protection of sensitive photos in OSNs

Code number	SCD_SP_1
Title	A minor is about to share a sensitive photo with inappropriate audience in OSNs

Description	**As a minor** **I want to** get notified when I am about to share a sensitive photo with inappropriate audience **so that** I can take measures to prevent sharing it with inappropriate audience
Acceptance criteria	1. The web-proxy should detect and capture the photo that a minor is trying to share with inappropriate audience in OSNs. 2. The web-proxy should be able to scan and accumulate percentage of nudity in the captured photo that the minor is about to share. 3. The sensitive content detection and protection browser add-on should notify the user of being in danger to share sensitive photo with inappropriate audience. 4. The sensitive content detection and protection browser add-on should prevent sharing photo if the nudity percentage exceeds 75%. 5. The sensitive content detection and protection browser add-on should be able to provide alternative ways (e.g., encryption) for securely sharing sensitive photos.

Code number	SCD_SP_2
Title	Notify parent when her child is about to share a sensitive photo with inappropriate audience in OSNs
Description	**As a parent** **I want to** get notified when my child is about to share a sensitive photo with inappropriate audience in OSNs **so that** i can prevent that.
Acceptance criteria	1. The web-proxy should detect and capture the photo that the child is trying to share with inappropriate audience in OSNs. 2. The web-proxy should be able to scan and accumulate percentage of nudity in a photo that a child is about to share. 3. The sensitive content detection and protection browser add-on should notify the parent of his/her child’s activity -being in danger to share a sensitive photo with inappropriate audience in OSNs.

5.3.2. Scenario 2: Detection and protection of sensitive information in OSNs

Code number	SCD_SI_1
Title	A minor is about to share her personal information with inappropriate audience in OSNs
Description	**As a parent** **I want to** get notified when my child is about to share her personal information with inappropriate audience **so that** I can take measures to prevent sharing it with inappropriate audience
Acceptance criteria	1. The web-proxy should detect when a minor is trying to share her personal information (e.g., home address) with inappropriate audience. 2. The web-proxy should block the child from sharing personal information if the parent set this from before. 3. If the parent did not chose to block the child from sharing personal information then she should be notified that her child is trying to share personal information in OSNs providing her with this information and asking

	for her consent. 4. The sensitive content detection and protection browser add-on should notify the child of the risks to share her personal information with inappropriate audience in OSNs and provide her with alternatives to secure ways to do it.

5.3.3. Scenario 3: Secure sharing of sensitive content in OSNs

Code number	SCD_SS_1
Title	Securely share sensitive content in OSNs using steganography
Description	**As a user** **I want to** be able to share sensitive content in OSNs using steganography **so that** i can prevent it to be shared with inappropriate audience
Acceptance criteria	1. The ENCASE sensitive content detection and protection browser add-on should offer the user the ability to steganographise an image or a text before sending it to the receiver. 2. The receiver should be able using ENCASE to de-steganographise and view the actual content that the user sent to him. 3. Any other user that the sender does not want him to see the actual content should be able to do so.

Code number	SCD_SS_2
Title	Securely share sensitive content in OSNs using Attribute-Based Encryption
Description	**As a user** **I want to** be able to share sensitive content in OSNs using attribute-based encryption **so that** i can prevent it to be shared with inappropriate audience
Acceptance criteria	1. Using the sensitive content detection and protection browser add-on the user should be able to encrypt the content that he wants to share or send in OSNs using the attribute of the receiver. 2. Using the sensitive content detection and protection browser add-on the receiver should be able to decrypt and view the actual content that the user sent. 3. The receiver should be able to view the actual content only if he holds the attribute information that the sender used to encrypt the content. 4. Anyone who do not have the attribute, should not be able to decrypt and the actual content that the sender sent.

Code number	SCD_SS_3
Title	Securely share sensitive content in OSNs using Group Encryption
Description	**As a user** **I want to** be able to share sensitive content in OSNs using group encryption **so that** i can prevent it to be shared with inappropriate audience
Acceptance criteria	1. Using the sensitive content detection and protection browser add-on the user should be able to encrypt the content (photo or text) that he wants to share or sent in OSNs and specify the group of people that he wants to be able to decrypt and view the actual content that the user sent.

	2. A user should be able to decrypt and view the content that the user sent only if he is in the group that the user specified. 3. Anyone who is not in the group that the sender specified, should not be able to decrypt and view the actual content that the sender sent.

5.4. Use Case D - Educators' Awareness

5.4.1. Scenario 1: Malicious behaviour detection in educational OSN groups

Code number	EA_MBD_1
Title	Ease and fast installation of ENCASE for education purposes
Description	**As an educator** **I want to** be able to easily and fast enable ENCASE for my students **so that** i can protect my students from malicious behaviours in educational OSN groups
Acceptance criteria	1. The educator should be able to create a list with the OSN accounts of their students. 2. When the educator adds a student's OSN account to a list the student should be able to authorize ENCASE for accessing his/her account.

Code number	EA_MBD_2
Title	Educator receives notification for malicious behaviour in an educational OSN group
Description	**As an educator** **I want to** receive notifications from ENCASE when a malicious behaviour is detected in my educational OSN group **so that** i can take the appropriate measures and protect my students
Acceptance criteria	1. The educator should timely receive notifications for cyber bullying, threatening messages among group users, and when a member of the group is experiencing is about to experience distressed or aggressive behaviour.

5.4.2. Scenario 2: Fake identity and activity detection in educational OSN groups

Code number	EA_FIFA_1
Title	Educator receives notifications for fake identity in an educational OSN group
Description	**As an educator** **I want to** receive notifications when there is a fake identity among the users of my OSN group for educational purposes **so that** i can protect my students from being exposed to malicious behaviours or fake activity
Acceptance criteria	1. The system should analyse the profile of every new user who joins the educational OSN group trying to determine whether it is a fake identity. 2. The educator should receive a notification when a profile with fake identity is trying to join or joined an educational OSN group.

Code number	EA_FIFA_2
Title	Educator receives notifications for false information dissemination in an

	educational OSN group
Description	**As an educator** **I want to** receive notifications when someone is posting false information in an educational OSN group **so that** i can detect and delete such posts
Acceptance criteria	1. ENCASE should capture and analyse every new post to an educational OSN group trying to determine whether it contains false information or not. 2. When a post that contains false information is detected, the educator should be notified.

5.4.3. Scenario 3: Sensitive content detection in educational OSN groups

Code number	EA_SCD_1
Title	Prevent users from sharing sensitive content in an educational OSN group
Description	**As an educator** **I want to** prevent users from sharing sensitive content to my OSN group for educational purposes **so that** i can prevent any undesired incidence
Acceptance criteria	1. The system should everything that a user tries to post to the OSN group, in order to detect whether it contains sensitive content. 2. When the system detects sensitive content it should block the user from sharing it to the group.

6. Reference Architecture

6.1.General Description

The objectives of ENCASE stem from the need to safeguard the security and privacy of minors against malicious actors in OSNs like cyber bullies. The measurement-driven approach that ENCASE follows to assess the urgency and existence of threats like fake identity and activity, and malicious behaviour (such as sexual cyber grooming, cyber bullying, etc.) in Online Social Networks, will guide the design and the implementation of the mitigation mechanisms. To this end, the ENCASE project aims to design and implement a system that protects minors from the aforementioned threats and will be offered through three browser add-ons, a web-proxy server that captures their OSN activity and a middleware that tries to unveil any incidence of malicious behaviour or fake activity based on machine learning detection rules generated from the back-end of our architecture.

As mentioned above, the functionality of ENCASE will be offered via three browser add-ons. These browser add-ons are the web interface of the ENCASE ecosystem and will not perform any complex processing or analysis. The first browser add-on is responsible for reporting any type of malicious behaviour like cyber bullying or sexual cyber grooming. More specifically, this add-on will inform minors of whether they have befriended or are communicating with a person that is presently attempting to bully or exploit them, or has in the past exhibited aggressive behaviour, or has caused other persons to exhibit emotional distress. The second add-on is responsible for informing users for fake identities and activity in OSNs. This browser add-on will enable users to be aware of whether they are communicating with a person that misrepresents its identity, and therefore its intentions, or are being the receivers of false information, or are themselves the subject of malicious false information that spreads through the network. The third one scans an OSN user’s content that she is about to be share (e.g., a photo) in order to determine if it contains nudity. Subsequently, it provides informative notifications and enables users to protect such content from unwarranted leakage to unwanted recipients with easily learnable and usable interfaces.

The ENCASE Web-Proxy Server is mainly responsible to capture all the traffic of the users. Also it will offer the required functionality for user registration and authentication, and for sensitive content detection and protection. Besides registration, the proxy will be able to terminate TLS connections and block user’s incoming or outgoing traffic when malicious behaviour is detected. In general, the web-proxy will not perform any complex processing and is separated from the back-end of our architecture for scalability, flexibility and privacy.

The Middleware of our architecture is responsible for malicious behaviour and fake identity and activity detection. The detection will be based on rules generated from the Back-end of our architecture. It will also analyze the traffic send from the user to be able to detect when the user is about to experience or is experiencing distressed or aggressive behaviour. Additionally, when malicious behaviour is detected the notification module of the middleware should be able to yield notifications to the minors and their parents through the browser add-ons.

The Back-End of our architecture is what we call Data analytics software stack. The back-end will host the machine learning algorithms that will be implemented during the project’s lifetime for the detection of malicious behaviour, hate speech, cyber bullying, fake identities and false information dissemination. The back-end receives users’ donated data from the web-proxy only with the explicit consent of users. Those machine learning and deep learning algorithms will run using the OSN data donated from the ENCASE users with the purpose to extract features and signatures which are expressed as detection rules that will be used by the middleware of our architecture. In order to reduce computation time and cost, the algorithms will run as per a predefined schedule.

Furthermore, ENCASE will allow users who are willing to contribute OSN data to do so. Such functionality will be offered by the browser add-ons where users will also be able to flag a suspicious communication, user or content and send it to the ENCASE back-end for analysis through the web-proxy. Users who contribute data to ENCASE they contribute their social egonet too and the features of their friends. In the end the back end gets a more holistic view of the network so that it can perform fake account and false information detection. In order to force users to contribute their OSN data we will provide incentives like: a) extra features in the ENCASE system; b)free usage or discounts on the user of our system; and c) personalized and more accurate cyber bullying and cyber grooming detection which will not be no longer rule-based but more self-adaptive.

Figure 16 below depicts the overview of our architecture with all the components interacting with each other under a unified architecture.

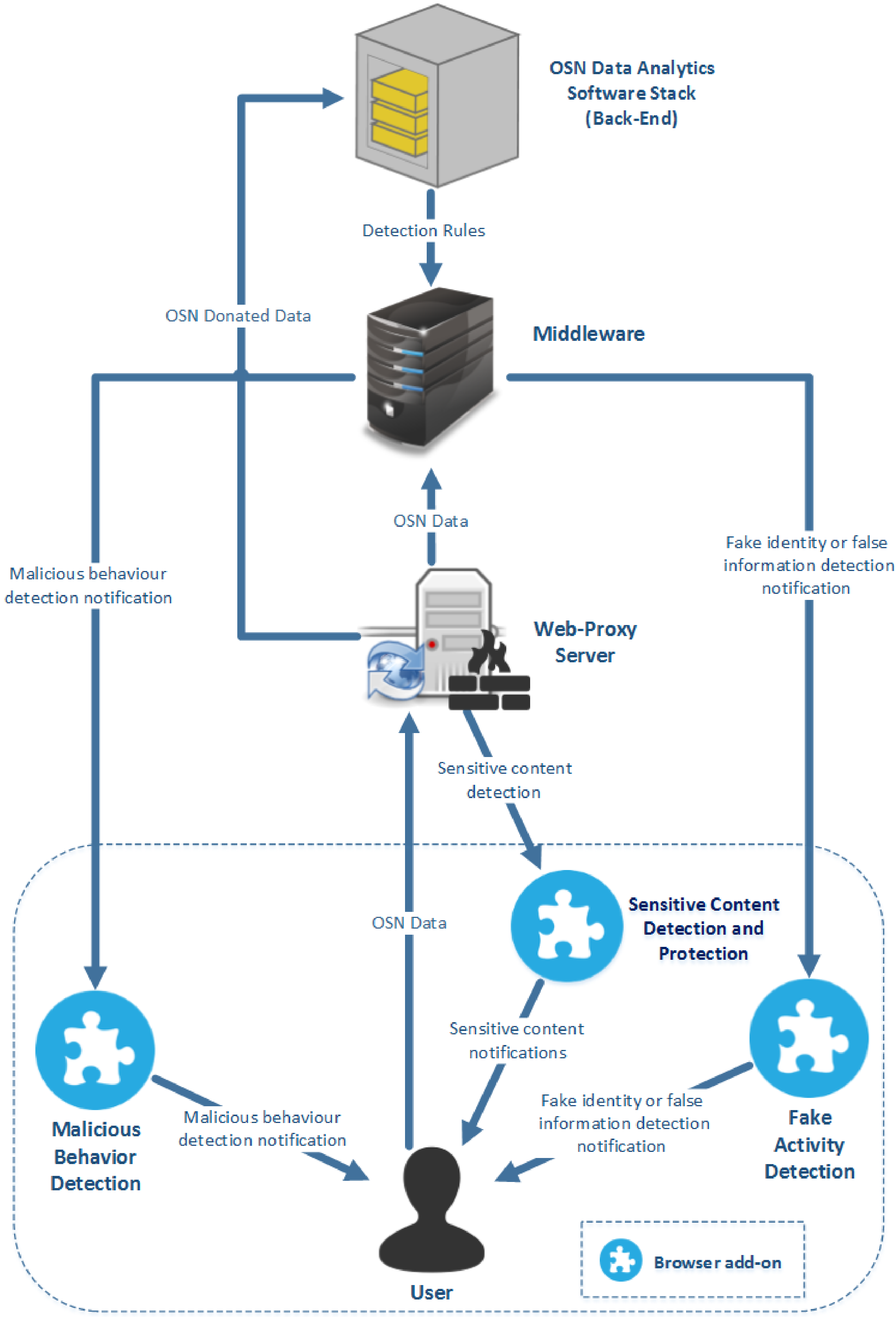

OSN Data Analytics Software Stack (Back-End)
Detection Rules
OSN Donated Data
Middleware
OSN Data
Malicious behaviour detection notification
Fake identity or false information detection notification
Web-Proxy Server
Sensitive content detection
OSN Data
Sensitive Content Detection and Protection
Sensitive content notifications
Malicious Behavior Detection
Malicious behaviour detection notification
Fake identity or false information detection notification
Fake Activity Detection
User
Browser add-on

Figure 9. ENCASE overall Reference Architecture

6.2. Established Architectures

In order to design the high-level architecture of the ENCASE ecosystem a lot of established and state-of-the art architectures have been investigated. Those architectures were the basis for the design of the ENCASE reference architecture. Some of those architectures are listed and explained below.

6.2.1. Proxy plug-in Architecture

Such architecture is applied when you want to re-direct the requests to the application server based on the configurations specified inside plug-in configuration file. The advantages of using the proxy plug-in are as follows:

- Re-direction of requests
- Load Balancing of requests
- Serves the static data (using the proxy cache)

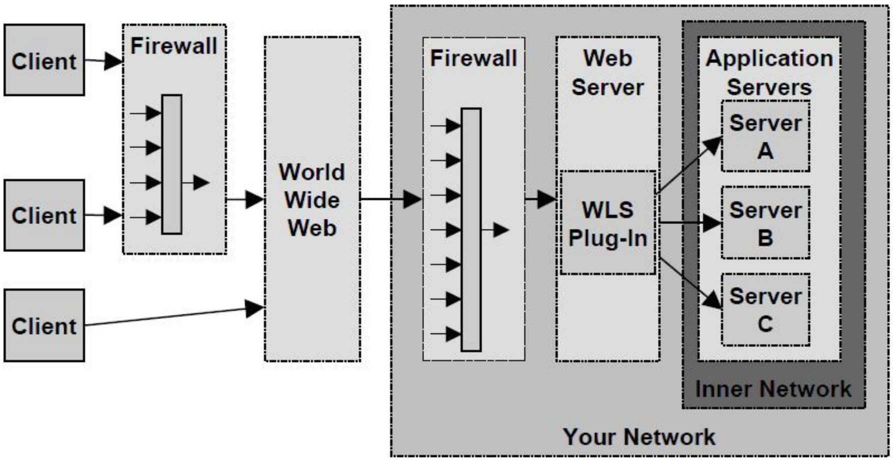


Figure 10. Proxy plug-in inside the dynamic application server architecture

a. WebLogic Server Proxy Plug-ins:

Web-server plug-ins [80] allow requests to be redirected from Oracle HTTP Server, Oracle iPlanet Web-Server, Apache HTTP Server, or Microsoft Internet Information Server (IIS) to Oracle WebLogic Server. In this way, plug-ins enable the HTTP server to communicate with applications deployed on the WebLogic Server. The plug-in enhances an HTTP server installation by allowing Oracle WebLogic Server to handle requests that require dynamic functionality. In other words, you typically use a plug-in where the HTTP server serves static pages such as HTML pages, while Oracle WebLogic Server serves dynamic pages such as HTTP Servlets and Java Server Pages (JSPs). Oracle WebLogic Server may be operating in a different process, possibly on a different host. To the end user—the browser—the HTTP requests delegated to Oracle WebLogic Server still appears to be coming from the HTTP server. In addition, the HTTP-tunneling facility of the WebLogic client/server protocol also operates through the plug-in, providing access to all Oracle WebLogic Server services. The plug-in proxy requests to Oracle WebLogic Server based on a configuration that you specify.

- a request can be proxied based on the URL of the request or a portion of the URL. This is called proxying by path.

- a request can be proxied based on the MIME type of the requested file, which is called proxying by file extension.

One can also enable both methods. If both the methods are enabled and a request matches both criteria, the request is proxied by path. One can also specify additional parameters for each of these types of requests that define additional behavior of the plug-in [81].

b. Hola:

Hola [82] is the first community powered (Peer-to-Peer) VPN, where users help each other to make the web accessible for all, by sharing their idle resources. Hola is a freemium web and mobile application which claims to provide a faster, private and more secure Internet. It provides a form of virtual private network services to its users through a peer-to-peer network. It also uses peer-to-peer caching. When a user accesses certain domains that are known to use geo-blocking, the Hola application redirects the request to go through the computers and internet connections of other users in non-blocked areas, thereby circumventing the blocking. This also means that other users might access the internet through one's own computer, and that part of one's upload bandwidth might be used for serving cached data to other users. Paying users can choose to redirect all requests to peers but are themselves never used as peers.

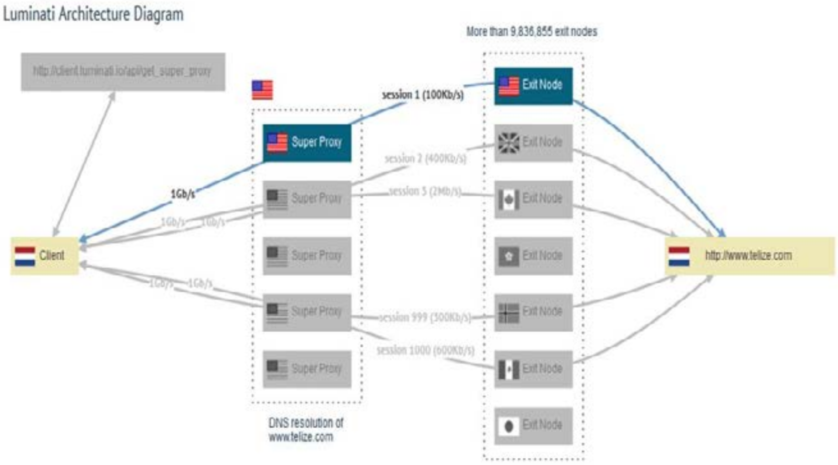


Figure 11. High-level design of Hola Architecture

c. Chromium Plug-ins implementation and architecture:

Chromium [83] supports a multi–process architecture and In-process architecture. Chromium uses separate processes for browser tabs to protect the overall application from bugs and glitches in the rendering engine. They also restrict access from each rendering engine process to others and to the rest of the system. In some ways, this brings to web browsing the benefits that memory protection and access control brought to operating systems. They refer to the main process that runs the UI and manages tab and plugin processes as the "browser process" or "browser". Likewise, the tab-specific processes are called "render processes" or "renderers." The renderers use the Blink open-source layout engine for interpreting and laying out HTML. Following picture depicts the Multi Process Architecture of Chromium.

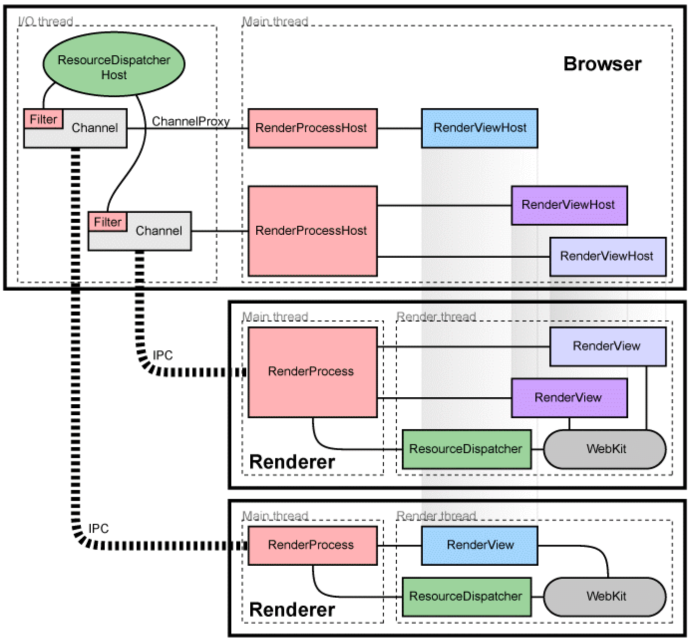


Figure 12. Chromium multi-process architecture

Chromium has the ability to run plugins in process as well as out of process. Both start at their non-multi-process-aware WebKit embedding layer, which expects the embedder to implement the WebKit::WebPlugin interface. This is implemented by WebPluginImpl. The WebPluginImpl talks "up" the chain to a WebPluginDelegate interface, which for in-process plugins is implemented by WebPluginDelegateImpl. This in turn talks to our NPAPI wrapper layer.

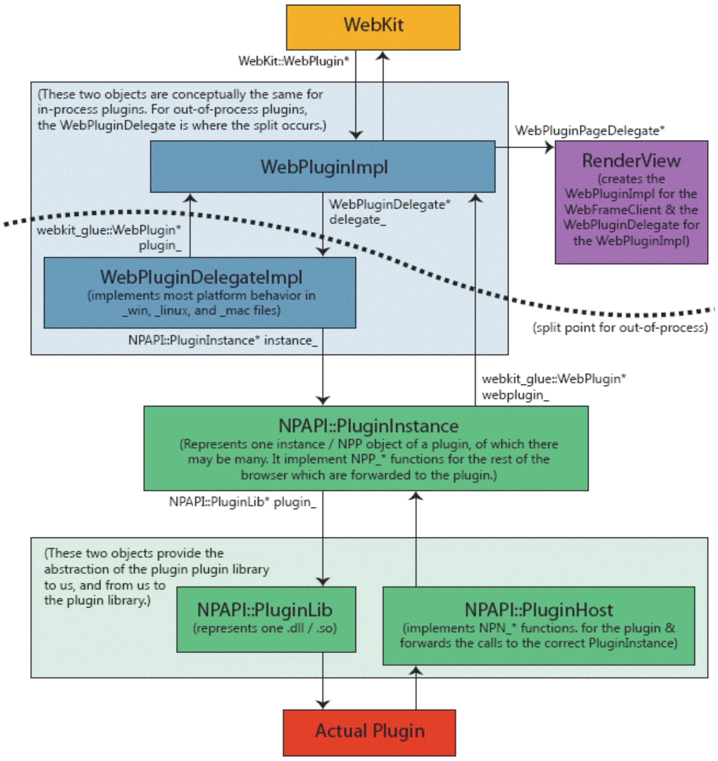

WebKit
WebKit::WebPlugin*
(These two objects are conceptually the same for in-process plugins. For out-of-process plugins, the WebPluginDelegate is where the split occurs.)
WebPluginPageDelegate*
WebPluginImpl
RenderView
(creates the WebPluginImpl for the WebFrameClient & the WebPluginDelegate for the WebPluginImpl)
WebPluginDelegate* delegate_
webkit_glue::WebPlugin* plugin_
WebPluginDelegateImpl
(implements most platform behavior in _win, _linux, and _mac files)
(split point for out-of-process)
NPAPI::PluginInstance* instance_
webkit_glue::WebPlugin* webplugin_
NPAPI::PluginInstance
(Represents one instance / NPP object of a plugin, of which there may be many. It implement NPP_* functions for the rest of the browser which are forwarded to the plugin.)
NPAPI::PluginLib* plugin_
(These two objects provide the abstraction of the plugin plugin library to us, and from us to the plugin library.)
NPAPI::PluginLib
(represents one .dll / .so)
NPAPI::PluginHost
(implements NPN_* functions. for the plugin & forwards the calls to the correct PluginInstance)
Actual Plugin

Figure 13. In-process plugin architecture

6.2.2. Open Source Web-Proxies

For the design of our web-proxy we have investigated some of the existing well-known open source web-proxies. Below we list and describe the most related ones:

a. ENVOY (https://lyft.github.io/envoy/docs/intro/what_is_envoy.html)**:**

ENVOY is an L7 web-proxy server and communication bus designed for large modern service oriented architectures and implemented in C++. It offers Layer 7 Load Balancing that deals with the actual content of the message. Layer 7 also enables the load balancer to make smarter load-balancing decisions, and to apply optimizations and changes to the content (such as compression and encryption). A device that performs Layer 7 load balancing is often referred to as a reverse-proxy server.

In general, ENVOY provides the following features:

i. **ENVOY works with any application programming language.** A single ENVOY deployment can form a mesh between Java, C++, Go, PHP, Python, etc.

ii. **L3/L4 filter architecture.** HTTP filters can be plugged into the HTTP connection management subsystem that performs different tasks such as buffering, rate limiting, routing/forwarding, etc.

iii. **HTTP L7 filter architecture.** HTTP filters can be plugged into the HTTP connection management subsystem that performs different tasks such as buffering, rate limiting, routing/forwarding, etc.

iv. **HTTP L7 routing.** ENVOY supports a routing subsystem that is capable of routing and redirecting requests based on path, authority, content type, runtime values, etc.

v. **Service discovery.** Envoy supports multiple service discovery methods including asynchronous DNS resolution and REST based lookup via a service discovery service.

vi. Health checking

vii. **Front/edge proxy support.** Envoy includes enough features to make it usable as an edge proxy for most modern web applications. This also includes TLS termination.

b. Glype (https://www.glype.com/):

Glype is a web-based proxy script written in PHP which focuses on features, functionality, and ease of use. Webmasters use Glype to quickly and easily set-up their own proxy sites. Glype helps users to defeat Internet censorship and be anonymous while web browsing. A web-based proxy script is hosted on a website which provides a proxy service to users via a web browser. A proxy service downloads requested web pages, modifies them for compatibility with the proxy, and forwards them on to the user. Web proxies are commonly used for anonymous browsing and bypassing censorship and other restrictions. Following are the distinguishable features of Glype:

- Free for personal use and licensing options are available for commercial use.
- Source Viewable and webmasters may modify the source code subject to the terms of the Software License Agreement.
- Plug and Play. Simply upload, configure and go!
- Admin Control Panel for easy management and configuration.
- JavaScript Support provides increased compatibility with websites.
- Skin-able. A theme system allows for customization of your proxy.
- Access Controls blacklist users by IP address and websites by domain name.

- Blocked.com Integration protects the proxy by blocking specified countries, filtering companies, malicious traffic, bots and spiders, and more.
- Unique URLs provide greater privacy by expiring URLs in the browser history at the end of a browsing session.
- Plugins allow for easy installation of site-specific modifications and they are useful for adding new functionality to websites.
- Advanced Options let users change their user-agent and referrer, manage cookies, and remove JavaScript and Flash.

Glype supports the plugin development for particular websites that the proxy may access. For every social media we can create an individual plugin. With the use of cookies, the plugin request and the response from the website we are able to collect the App-Id, User-Id and Content. This information can be transferred to the Backend system / Middleware for analysis and then store the information (encrypted) in the backend database system (if it is necessary, which means if there is malicious content in the response). If malicious content is found, then a page parsing mechanism must exist for blocking the particular section that contains the malicious content without blocking the page. The flow is shown in Figure 13. Glype supports the integration with plugins per social media and it can receive the content for web pages. As soon as Glype is an open source script gives the flexibility for extending the existing functionality that it supports. It is possible to add web services communication libraries written in PHP which can consume web services from the Middleware.

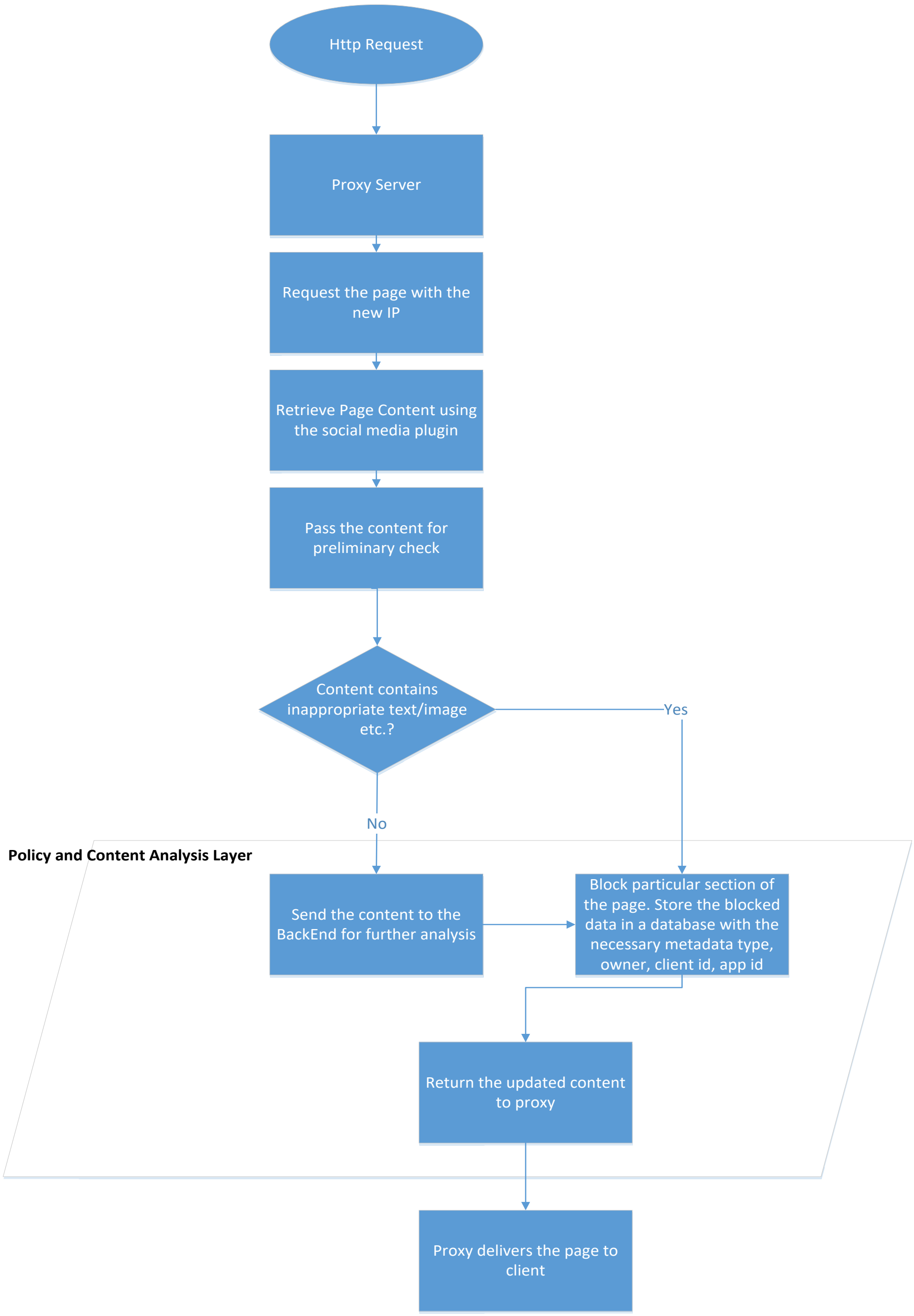

Http Request
Proxy Server
Request the page with the new IP
Retrieve Page Content using the social media plugin
Pass the content for preliminary check
Content contains inappropriate text/image etc.?
Yes
No
Policy and Content Analysis Layer
Send the content to the BackEnd for further analysis
Block particular section of the page. Store the blocked data in a database with the necessary metadata type, owner, client id, app id
Return the updated content to proxy
Proxy delivers the page to client

Figure 14. Probable ENCASE web-proxy server architecture based on Glype

c. PHProxy (http://plrplr.com/47442/what-is-phproxy-and-how-can-i-use-it/, https://github.com/creaktive/PHProxy)**:**

PHProxy is a type of web-proxy that allows users to gain access to websites anonymously. It was made available on 15th of June 2004 and it was continuously growing until it became the most popular web proxy solution after Glype. The latest available version of PHProxy is 0.5 beta 2 and is available since 20th of January 2007.

d. TrotiNet (https://github.com/krys-g/TrotiNet, http://trotinet.sourceforge.net/):

TrotiNet is an open-source proxy library implementation, implemented in C#. Its target is to offer to the public a reusable framework that allows the development of any sort of HTTP proxies.

e. Titanium Web-Proxy (https://github.com/justcoding121/Titanium-Web-Proxy):
Titanium is a lightweight HTTP and HTTPS web-proxy server implemented in C# Some of the salient features of Titanium are:

- Support of HTTP(S) and most of the features of HTTP 1.1
- Support of redirect/block/update requests
- Support of updating response
- Safely relays WebSocket requests over HTTP
- Support of mutual SSL authentication
- Fully asynchronous proxy
- Support of proxy user authentication

f. NGINX (https://www.nginx.com/resources/wiki/):
NGINX is a free, open-source, high-performance HTTP server and reverse proxy, as well as an IMAP/POP3 proxy server. NGINX is known for its high performance, stability, rich feature set, simple configuration, and low resource consumption. NGINX is one of a handful of servers written to address the C10K problem. Unlike traditional servers, NGINX doesn’t rely on threads to handle requests. Instead it uses a much more scalable event-driven (asynchronous) architecture. This architecture uses small, but more importantly, predictable amounts of memory under load. Even if you don’t expect to handle thousands of simultaneous requests, you can still benefit from NGINX’s high-performance and small memory footprint. NGINX scales in all directions: from the smallest VPS all the way up to large clusters of servers.

g. HAProxy Web-Proxy Load Balancer (http://www.haproxy.org/):
HAProxy, which stands for High Availability Proxy, is a popular open source software TCP/HTTP Load Balancer and proxying solution which can be run on Linux, Solaris, and FreeBSD. It is a very fast and reliable solution offering high availability, load balancing, and proxying for TCP and HTTP-based applications. It is particularly suited for very high traffic web sites and powers quite a number of the world's most visited ones. Over the years it has become the de-facto standard open-source load balancer, is now shipped with most mainstream Linux distributions, and is often deployed by default in cloud platforms. HAProxy version 1.7 has been released in November 25, 2016. The native SSL support on both sides with SNI/NPN/ALPN and OCSP stapling, IPv6 and UNIX sockets are supported everywhere, full HTTP keep-alive for better support of NTLM and improved efficiency in static farms, HTTP/1.1 compression (deflate, gzip) to save bandwidth, PROXY protocol versions 1 and 2 on both sides, data sampling on everything in request or response, including payload, ACLs can use any matching method with any input sample maps and dynamic ACLs updatable from the CLI stick-tables support counters to track activity on any input sample custom format for logs, unique-id, header rewriting, and redirects, improved health checks (SSL, scripted TCP, check agent, etc.), much more scalable configuration supports hundreds of thousands of back-ends and certificates without sweating.

h. SQUID Proxy server (http://wiki.squid-cache.org/FrontPage)**:**
Squid is a full-featured open-source web proxy cache server application which provides proxy and cache services for Hyper Text Transport Protocol (HTTP), File Transfer Protocol (FTP), and other popular network protocols.

6.2.3. Firewall Architecture

A firewall is a system designed to prevent unauthorized access to or from a private network. Currently, there are three generation of firewalls and now we are into next generation of firewall. Following are the details of each of one of the three generations.

a. 1st Generation Firewall:
Secure Web Gateway or Proxy-based Network Content Inspection: Proxies have been deployed to provide internet caching services to retrieve objects and then forward them. Consequently, all network traffic is intercepted, and potentially stored. These graduated to what is now known as secure web gateways, proxy-based inspections retrieve and scans object, script, and images.

Proxies, which relies on a fetch the content first if it were not cached, then forwarding the content to the recipient introduced some form of file inspection as early as 1995 when MAILsweeper was released by Content Technologies (now Clearswift), which was then replaced by MIMEsweeper in 2005. 2006 saw the release of the open-source, cross-platform antivirus software ClamAV provided support for caching proxies, Squid and NetCache. Using the Internet Content Adaptation Protocol (ICAP), a proxy will pass the downloaded content for scanning to an ICAP server running an anti-virus software. Since complete files or ‘objects’ were passed for scanning, proxy-based anti-virus solutions are considered the first generation of network content inspection [84, 85].

BlueCoat, WebWasher and Secure Computing Inc. (now McAfee, now a division of Intel), provided commercial implementations of proxies, eventually becoming a standard network element in most enterprise networks.

Limitations: While proxies (or secure web gateways) provide in-depth network traffic inspection, their use is limited because they:

- require network reconfiguration which is accomplished through – a) end-devices to get their browsers to point to these proxies; or b) on the network routers to get traffic routed through these devices
- are limited to web (http) and ftp protocols; cannot scan other protocols such as e-mail
- Proxy architectures which are typically built around SQUID, cannot scale with concurrent sessions, limiting their deployment to enterprises.

b. 2nd Generation Firewall:
Gateway/Firewall-based Network Traffic Proxy-assisted Deep Packet Inspection: The Second generation Network Traffic Inspection solutions were implemented in firewalls and/or UTMs. Given that network traffic is choked through these devices, in addition to DPI inspection, proxy-like inspection is possible. This approach was first pioneered by NetScreen Technologies Inc. (acquired by Juniper Networks Inc.). However, given the expensive cost of such operation, this feature was applied in tandem with a DPI system and was only activated on a-per-need basis, or when content failed to be qualified through the DPI system [86, 87].

c. 3rd Generation Firewall:

Transparent, Application-aware Network Content Inspection, or Deep Content Inspection: The third, and current, generation of Network Content Inspection known as Deep Content Inspection solutions are implemented as fully transparent devices that perform full application level content inspection at wire speed. In order to understand the communication session’s intent —in its entirety—, a Deep Content Inspection System must scan both the handshake and payload. Once the digital objects (executables, images, JavaScript’s, .pdfs, etc. also referred to as Data-In-Motion) carried within the payload are constructed, usability, compliance and threat analysis of this session and its payload can be achieved. Given that the handshake sequence and complete payload of the session is available to the DCI system, unlike DPI systems where simple pattern matching and reputation search are only possible, exhaustive object analysis is possible. The inspection provided by DCI systems can include signature matching, behavioral analysis, regulatory and compliance analysis, and correlation of the session under inspection to the history of previous sessions. Because of the availability of the complete payload's objects, and these schemes of inspection, Deep Content Inspection Systems are typically deployed where high-grade Security and Compliance is required or where end-point security solutions are not possible such as in bring your own device, or Cloud installations. This third generation approach [86, 87] of network content inspection was first pioneered by Wedge Networks Inc. who also coined the term "Deep Content Inspection". Deep Content Inspection is Content-focused instead of analyzing packets or classifying traffic based on application types such as in Next Generation Firewalls. "Understanding" content and its intent is the highest level of intelligence to be gained from network traffic. This is important as information flow is moving away from Packet, towards Application, and ultimately to Content. Examples of Inspection Levels are:

- Packet: Random Sample to get larger picture
- Application: Group or application profiling. Certain applications, or areas of applications, are allowed / not allowed or scanned further.
- Content: Look at everything. Scan everything. Subject the content to rules of inspection (such as Compliance/Data Loss Prevention rules). Understand the intent.

d. Next-Generation Firewall:

A Next-Generation Firewall (NGFW) is an integrated network platform that combines a traditional firewall with other network device filtering functionalities such as an application firewall using in-line deep packet inspection (DPI), an intrusion prevention system (IPS). Other techniques might also be employed, such as SSL and SSH interception, website filtering, QoS/bandwidth management, antivirus inspection and third-party integration.

Deep Content Inspection (DCI) is a form of network filtering that examines an entire file or MIME object as it passes an inspection point, searching for viruses, spam, data loss, key words or other content level criteria. Deep Content Inspection is considered the evolution of Deep Packet Inspection with the ability to look at what the actual content contains instead of focusing on individual or multiple packets. Deep Content Inspection allows services to keep track of content across multiple packets so that the signatures they may be searching for can cross packet boundaries and yet they will still be found. An exhaustive form of network traffic inspection in which Internet traffic is examined across all the seven OSI ISO layers is the application layer. Parallel to the development of Deep Packet Inspection, the beginnings of Deep Content Inspection can be traced back as early as 1995 with the introduction of proxies that stopped malware or spam. Deep Content Inspection can be seen as the third generation of Network Content Inspection, where network content is exhaustively examined [88, 89].

d. Palo Alto Networks Firewall Concepts:

Palo Alto Networks [88] has built a next- generation firewall with several innovation technologies- enabling organizations to empower, enhance and fix some of the shortcomings within traditional firewalls. The limitations of the concepts are that they are implemented on hardware. These innovation technologies bring business relevant elements (application, users, and content) under policy control via high performance firewall architecture. Delivered as a purpose-built appliance every Palo Alto Networks next-generation firewall utilizes dedicated, function specific processing that is tightly integrated with a single-pass software engine. This unique combination of hardware and software maximizes network throughput while minimizing latency. Each of the hardware platforms supports the same rich set of next-generation firewall functions and feature, including it Operation System, the PAN-OS, ensuring consistent operation across the entire line. The Palo Alto device contains separate control plane and date plane the control plane is for management interface for configuration commands for logging and reporting. Has its own dedicated CPU and its own memory both RAM and hard-drive. The data plane has its own CPU and RAM resources which are broken into three general areas. The signature matching engine, security processor, and the network processor. Network processing is for forwarding traffic for NAT (Network address translation), routing lookup, and for QOS (Quality of service), bandwidth shaping, MAC lookup, flow control. Security processor is responsible for Application ID (APP-ID) , User-ID, URL matching and policy matches Vs. signature matching engine is designed to use regular expression and signatures to catch exploits , virus, Spyware and also to search for data leakage like credit cards (CCN) , Social security numbers (SSN) and any sort of pattern matching. Figures 14 and 15 below depict the architecture used by Alto Networks.

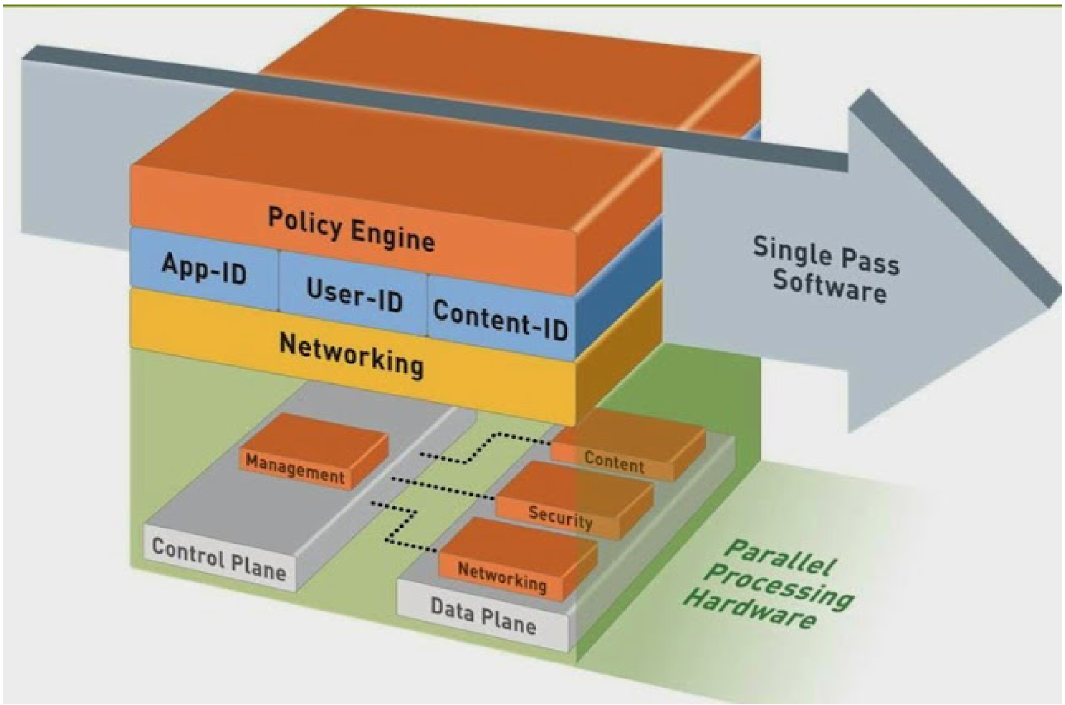


Figure 15. Palo Alto Network architecture

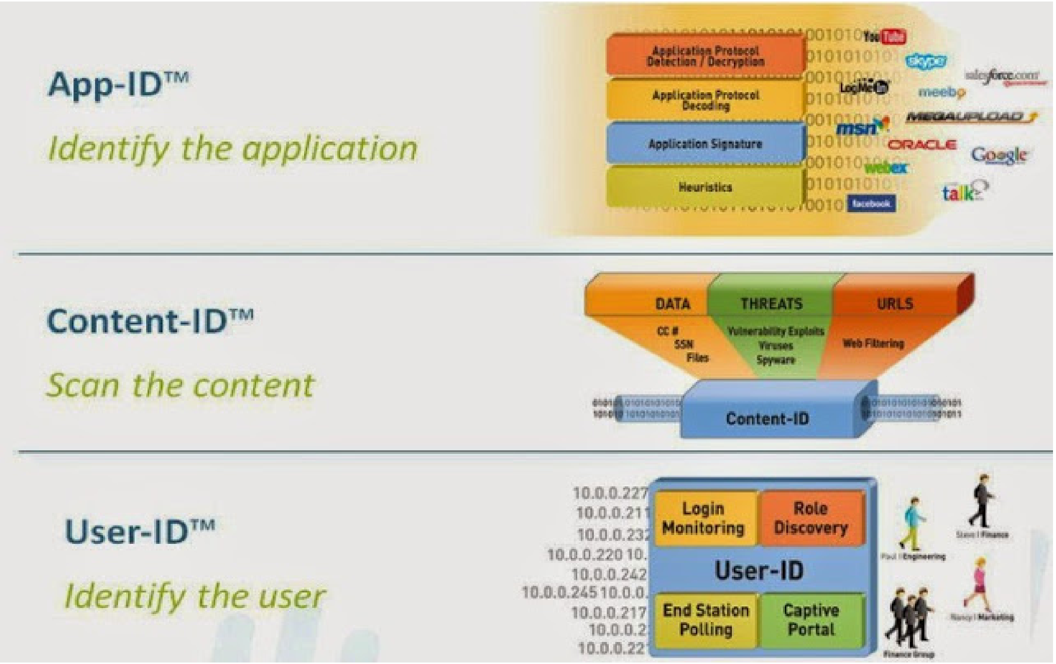


Figure 16. Details of App-ID, Content-ID, and User-ID of Palo Alto Network

The ENCASE reference architecture have been designed based on the above mentioned open-source proxies and firewalls and basically is a consolidation of next-generation firewall and a proxy plug-in architecture. The use of a web-proxy server becomes a necessity due to the number of algorithms that will be executed for the identification of the malicious behavior, fake identity and sensitive content detection. The architecture described in Figure 9 offers the flexibility to create web-proxy server plugins, and with the use of the cached data to produce decisions regarding the routing of the content and the generation of possible notifications.

The overall system design must be able to analyze content and produce decisions based on the semantics something that the Next Generation Firewalls (NGF) are doing and this is generally described as Deep Content Inspection (DPI). The NGF uses programmable hardware to achieve this but this might not be required if we limit our content filtering into the HTTP and HTTPS requests. Other protocols are not of high importance for the time being since we will focus on the browsing history of the users, the exchange of messages and social media content through their browser.

Therefore, if we take into account the three application layers depicted in Figure 14 we need to define the policy layer which contains the policy rules that will define the decisions that the system will take. The App-ID, User-ID and Content-ID layer that will identify the users, the social media and the content from both sides, user and attacker, and finally the Networking layer which will be responsible for blocking or allowing the content delivery according to the algorithms results. In the next section the reference architecture for ENCASE project will be discussed in details.

6.3. Architecture Components

In this subsection we describe in detail the functionalities of each component of the ENCASE architecture. As mentioned before, the ENCASE architecture consist of: a) the front-end, which basically is the three browser add-ons; b) the Web-proxy server component; c) the middleware; and d) the OSN data analytics software stack or back-end. Each component consists of various modules and the appropriate interfaces to be able to communicate with the other components of the architecture.

6.3.1. Web-Proxy Server

This subsection describes the web-proxy server of the ENCASE architecture that will be the main gateway for the users for registering and connecting with the ENCASE platform. In general, the web-proxy is responsible to capture all the user’s incoming and outgoing traffic and also block traffic when malicious behaviour, fake identity or activity is detected. The following figure depicts the various modules of the web-proxy and the interactions between them, and Table 5 below describes the functionality of each one of the modules of the web-proxy. Figure 17 also shows the procedure and the modules that are enabled when a sensitive content is detected.

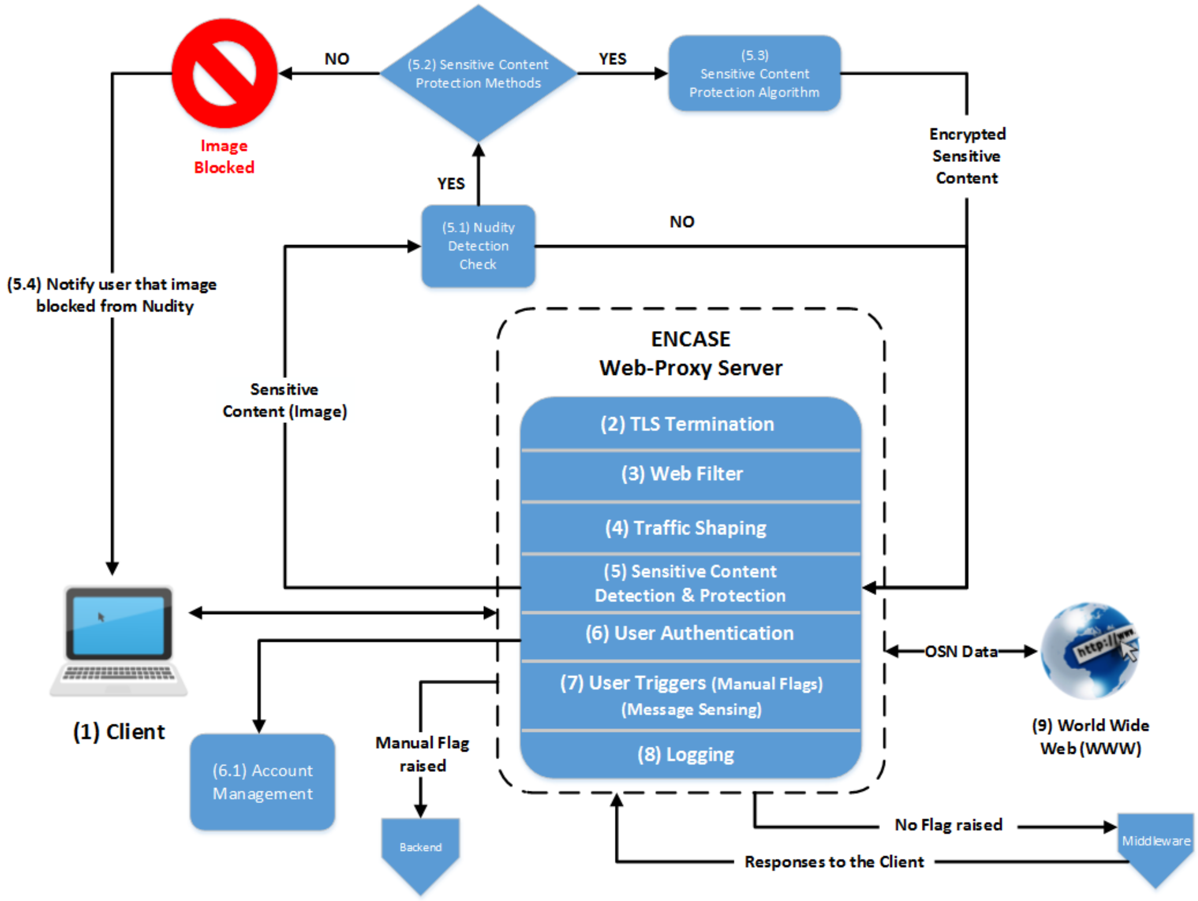

NO
(5.2) Sensitive Content Protection Methods
YES
(5.3) Sensitive Content Protection Algorithm
Image Blocked
Encrypted Sensitive Content
YES
(5.1) Nudity Detection Check
NO
(5.4) Notify user that image blocked from Nudity
Sensitive Content (Image)
ENCASE Web-Proxy Server
(2) TLS Termination
(3) Web Filter
(4) Traffic Shaping
(5) Sensitive Content Detection & Protection
(6) User Authentication
(7) User Triggers (Manual Flags) (Message Sensing)
(8) Logging
(1) Client
OSN Data
(9) World Wide Web (WWW)
(6.1) Account Management
Manual Flag raised
Backend
No Flag raised
Middleware
Responses to the Client

Figure 17. Web-Proxy Server component view

Identifier	Name	Description	Speculated Results
1	**Client**	A Minor	Minors and their parents are notified about any kind of cyber bullying, sexual cyber grooming, and false identity and/or activity in a predefined

			periodic manner.
2	**TLS Termination**	A TLS termination proxy (or SSL termination proxy) is a proxy server that is used by an organization to handle incoming TLS connections, decrypting the TLS and passing on the unencrypted request to the organization's other servers (it is assumed that the institution's own network is secure so the user's session data does not need to be encrypted on that part of the link). TLS termination proxies are used to reduce the load on the main servers by offloading the cryptographic processing to another machine, and to support servers that do not support TLS. **Servers capable of acting as a TLS termination proxy:** • Apache HTTP Server • Envoy • HAProxy • NGINX • Squid Proxy • stunnel TLS Proxy • Internet Information Services	Handling of incoming TLS connections, decrypting the TLS and passing on the unencrypted request to the servers
3	**Web Filter**	A Web filter is a program that can screen an incoming Web page to determine whether some or all of it should not be displayed to the user. The filter checks the origin or content of a Web page against a set of rules provided by company or person who has installed the Web filter. A Web filter allows an enterprise or individual user to block out pages from Web sites that are likely to include objectionable advertising, pornographic content, spyware, viruses, and other objectionable content.	Screening of an incoming Web page to determine whether some or all of it should not be displayed to the user.
4	**Traffic Shaping**	Traffic shaping (also known as packet shaping) is a computer network traffic management	Traffic shaping is used to optimize or guarantee performance, improve

		technique which delays some or all datagrams to bring them into compliance with a desired traffic profile. Traffic shaping is used to optimize or guarantee performance, improve latency, and/or increase usable bandwidth for some kinds of packets by delaying other kinds. The most common type of traffic shaping is application-based traffic shaping. In application-based traffic shaping, fingerprinting tools are first used to identify applications of interest, which are then subject to shaping policies. Some controversial cases of application-based traffic shaping include bandwidth throttling of peer-to-peer file sharing traffic. Many application protocols use encryption to circumvent application-based traffic shaping. Another type of traffic shaping is route-based traffic shaping. Route-based traffic shaping is conducted based on previous-hop or next-hop information.	latency, and/or increase usable bandwidth for some kinds of packets by delaying other kinds.
5	**Sensitive Content Detection and Protection**	Checks all the images that pass through the proxy server for nudity and blocks those that contain nudity. Also it enables the user to protect such sensitive providing him with various options like Steganography, Attribute-based encryption or group encryption.	Blocks images that contain nudity and offers users ways to securely share those images.
5.1	**Nudity Detection Check**	Checks if a given image contains nudity using tools like **"nude.js" library** (https://www.patrick-wied.at/static/nudejs/). If nudity is detected then proceed on choosing method to protect it, or if not then proceed as requested by the user.	Detection of Nudity in an Image
5.2	**Sensitive content protection methods**	Request from the user to choose whether she prefers to protect	Yes/No

		her sensitive content and the method she prefers to use.	
5.3	**Sensitive content protection algorithm**	Sensitive content protection algorithms like steganography, attribute-based encryption and group based encryption should be used for encrypting and securely sharing the Nude photo.	Protect the sensitive content of the user using the method she chose
5.4	**Notify user that image blocked from nudity**	If the user doesn’t opt for a protection method then block the Image and notify her that the Image is blocked due to nudity content	A notification is sent to the user informing her that the image is blocked due to nudity and also informs her parents about the minor’s attempt of sharing images that contains nudity.
6	**User Authentication**	User Log In Credentials are used for activating or deactivating ENCASE.	Activate or deactivate ENCASE.
6.1	**Account Management**	Keep User’s credentials like user id and password for detecting minor user and therefore helping User authentication module to activate or deactivate ENCASE.	Keeps a record with the users’ accounts credentials.
7	**User Triggers (Manual Flags) (Message Sensing)**	If the user/client feels (at any particular point of time) the he/she is being a victim of cyber bullying or sexually harassed or someone is trying to sexually groom him/her in a new way which is not detected by the ENCASE Security system then he/she can Manually Raise a Flag to identify it. For this kind of scenario the Backend will be activated on the predefined time schedule to either upgrade or modify the existing rules/models.	Manual Flag should be raised for the update and modification of the existing rules/models of middleware.
8	**Logging**	Logging the users web activity.	

Table 5. Web-Proxy Server modules description

6.3.2. Middleware

The Middleware component of the ENCASE architecture is responsible for the detection of malicious behaviours like cyber bullying, and for the detection of fake identities and activity and false information dissemination in OSNs. Middleware receives the captured OSN user’s traffic captured from the Web-Proxy component. Initially, this information is checked against some preliminary rules. Those preliminary rules are predefined and are not the outcome of a machine learning algorithm. They signify content inspection by the back-end in the event of a positive evaluation.

Additionally the middleware component is the one responsible for sending notifications to the minors and their parents in the case a threat is detected. Figure 18 below depicts the modules and the functionality of the middleware component and Table 6 provides a description of them.

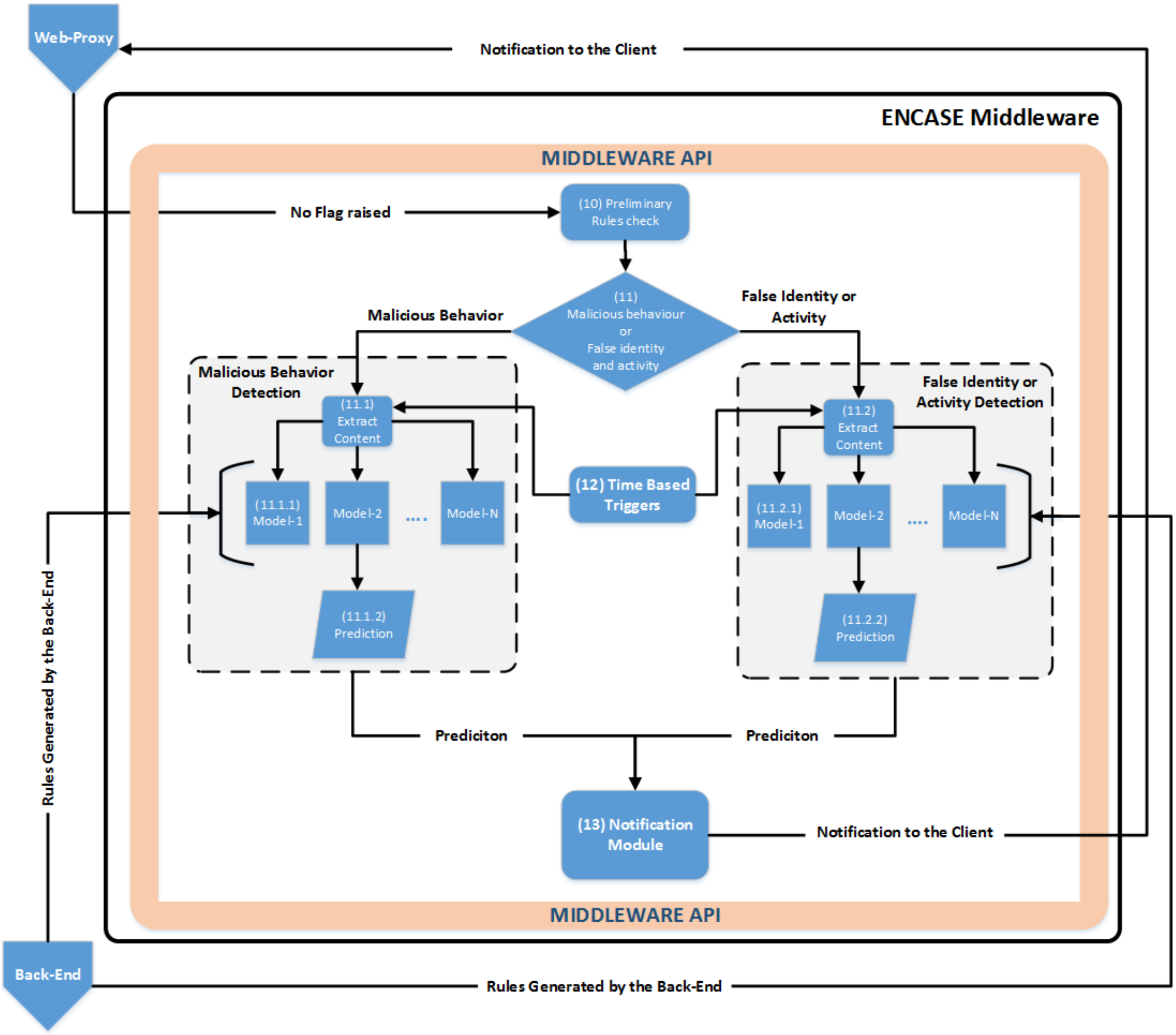


Figure 18. Middleware component view

Identifier	Name	Description	Speculated Results
10	Preliminary Rules check	This functionality should be restricted to predefined rules and not algorithm pre-processing evaluations. These rules should signify content inspection by the back-end in the event of a positive evaluation. Some examples of such rules are: • New friend is chatting with our user. For a period of time i.e. a month new friend chats should be inspected for	Based on these rules only the suspicious content is sent for further analysis to the malicious behaviour detection and fake identity and activity detection modules.

		sexual content, cyber grooming, cyberbullying etc. • Friend with blacklisted history is chatting with our user. • Friend that is an adult and not a direct relative is chatting with out user. • New Friend requests. • Wall posts from new friends. • Wall posts from friends with blacklisting history. • Wall posts from adult friends.	
11	Malicious or Fake identity or activity	This module is responsible to decide whether the suspicious content should checked against malicious behaviour or fake identity or activity.	-
11.1	Extract Content	Extract the content which should be checked for Malicious behaviour	Textual or word level content should be extracted
11.1.1	Model 1….n	Model 1 to n are the rules generated by the backend over the time which should be used for Checking the incoming content for Malicious behaviour	Try to match with existing rules for predicting the Malicious behaviour
11.1.2	Prediction	Prediction with Percentage of probability for the content to be Malicious behaviour type	Prediction with Percentage of Probability
11.2	Extract Content	Extract the content which should be checked for Fake Activity	Opposite side User’s Profile and other account information
11.2.1	Model 1….n	Model 1 to n are the rules generated by the backend over the time which should be used for Checking the opposite side user’s Fake activity or whether that user is a blacklisted one or has a Fake activity history.	Try to match with existing rules for finding whether the opposite end user is Fake.
11.2.2	Prediction	Prediction with Percentage of probability for the user to be Fake	Prediction with Percentage of Probability
12	Time Based Trigger (Cron Jobs)	The software utility Cron is a time-based job scheduler in Unix-like computer operating systems. People who set up and maintain software environments use cron to schedule jobs (commands or	Malicious Behaviour and fake identity and activity detection should run on a predefined time schedule.

		shell scripts) to run periodically at fixed times, dates, or intervals. It typically automates system maintenance or administration—though its general-purpose nature makes it useful for things like downloading files from the Internet and downloading email at regular intervals. • Non-session based timed triggers that retrieve content for inspection, especially with regards to Cyberbullying and Malicious content i.e. all post since last run from Twitter that mention our user or all chats from Facebook for our user. These can be run in the background and inspected at leisure.	
13	Notification Module	Notify the minors and their parents.	Notify in terms of Weekly or Monthly Mail, or Instant notification to the parents if there is high probability for malicious behaviour or fake identity and activity.

Table 6. Middleware modules description

6.3.3. OSN Data Analytics Software stack (Back-End)

The Back-end of the ENCASE architecture is responsible to host and run all the machine learning algorithms. The main purpose of the back-end is to rarely generate and install detection rules in the middleware of our architecture for the detection of malicious behaviour, and fake identity and activity in OSNs. Figure 19 below shows the various modules that consist the back-end component, while Table 7 provides a description of those modules.

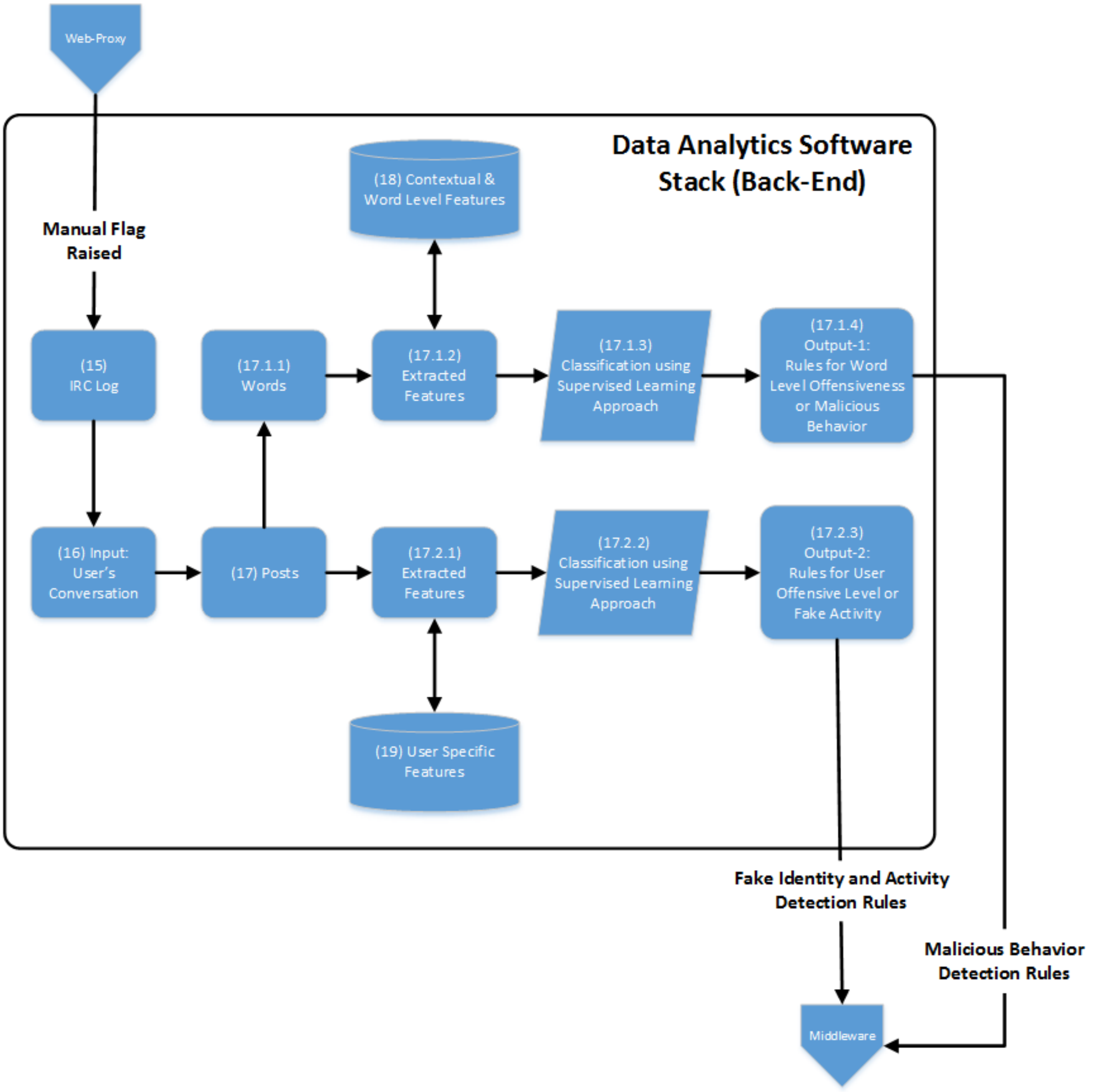

Web-Proxy
Data Analytics Software Stack (Back-End)
Manual Flag Raised
(18) Contextual & Word Level Features
(15) IRC Log
(17.1.1) Words
(17.1.2) Extracted Features
(17.1.3) Classification using Supervised Learning Approach
(17.1.4) Output-1: Rules for Word Level Offensiveness or Malicious Behavior
(16) Input: User's Conversation
(17) Posts
(17.2.1) Extracted Features
(17.2.2) Classification using Supervised Learning Approach
(17.2.3) Output-2: Rules for User Offensive Level or Fake Activity
(19) User Specific Features
Fake Identity and Activity Detection Rules
Malicious Behavior Detection Rules
Middleware

Figure 19. Data Analytics Software Stack (Back-End) component view

Identifier	Name	Description	Speculated Results
15	IRC Log	IRC- Internet Relay Chat, an Application Layer Protocol that facilitates communication in form of Text	Text
16	Input: User's Conversation	User's conversations in OSNs in the form of Text	Text
17	Posts	Extract the entire content of the captured OSN Posts	Text
17.1.1	Words	Separate individual words extracted from the captured OSN	Words

		data	
17.1.2	Extracted features	Extract Features from the Words	Word level features
17.1.3	Classification using Supervised Learning Approach	Classify the words, based on sophisticated and supervised Machine Leaning Techniques	Training of new Model
17.1.4	Output-1: Rules for Word Level Offensiveness or Malicious Behaviour	Generation of New rules or upgrade of the existing in the middleware	New rule or update existing rule
17.2.1	Extracted features	Extract features for the detection of fake identities and fake activity in OSNs	Opposite User’s profile
17.2.2	Classification using Supervised Learning Approach	Classify the user’s profile as a fake one and their activity as fake activity using sophisticated and supervised Machine Leaning Techniques	Training of New Model
17.2.3	Output-2: Rules for User Offensive Level or Fake Activity	Generation of New rules or update of the existing ones	New Model/Rule or Upgrade existing Model/Rule
18	Contextual & Word Level Features	Contextual and word level Ground truths	This is an existing ground truth database
19	User Specific Features	Fake User Specific and fake activity specific Ground Truth	This might be an existing ground truth database

Table 7. Data Analytics Software Stack (Back-End) modules description

6.4. Secure Sensitive Content Sharing Protocols

ENCASE will offer to its users the option to securely share their sensitive contents (images or text) with the audience of their preference preventing any unauthorised access to it. Users will have the option to do so using steganography-related techniques or cryptographic techniques like group encryption and attribute-based encryption. Such functionality and the use of those techniques will enable users to specify groups of people that can access and view certain sensitive content like the images they share in OSNs.

6.4.1. Steganography

Steganography is the art of hiding information in ways that prevent the detection of hidden messages. Steganography, derived from Greek, literally means "covered writing" (Greek words "stegos" meaning "cover" and "gratia" meaning "writing") [74, 75]. It comes under the assumption that if the feature is visible, the point of attack is evident, thus the goal here is always to conceal the very existence of the embedded data. Therefore Steganography gets a role on the stage of information security. Steganography is the science that involves communicating secret data in an appropriate multimedia carrier, e.g., image, audio and video files. The media with or without hidden information are called Stego Media and Cover Media, respectively. Steganography can meet both legal and illegal interests, e.g., civilians may use it for protecting privacy while terrorists may use it for spreading terroristic information [75].

Steganography and Cryptography are cousins in the spy craft family. Cryptography scrambles a message so it cannot be understood. Steganography hides the message so it cannot be seen. A message in cipher text, for instance, might arouse suspicion on the part of the recipient while an "invisible" created messaged with steganographic techniques will not [76].

Furthermore, another technique that is related to Steganography is Watermarking. Watermarking is a protecting technique which protects (claims) the owner's property right for digital media (i.e. images, music, video and software) by some hidden watermark [76]. Therefore, the goal of Steganography is the secret messages while the goal of watermarking is the cover object itself. The main objective of Steganography is to communicate securely in such a way that the true message is not visible to the observer. That is unwanted parties should not be able to distinguish in any sense between cover-image (image not containing any secret message) and stego-image (modified cover-image that contains secret message). Thus the stego-image should not deviate much from the original cover-image. Today Steganography is mostly used on computers with digital data being the carriers and networks being the high speed delivery channels [76].

For ENCASE project the initial idea was to protect both the sensitive image and text content. For this purpose steganographic method suits very well where the sensitive content will be embedded inside a normal, unsuspected cover image, which may be a basic scenery image as well. At the receiver’s end the stego-decryption algorithm will run to extract the sensitive content once the receiver right clicks on the received image and chooses the “Show Original” option. For these approach both the sender and receiver side needs to have ENCASE tool installed into their device.

6.4.2. Group Encryption

Group encryption, introduced by Kiayias et al. [77], is a novel cryptographic primitive that is the encryption analogue of a group signature. In general group encryption is used when we need to conceal a recipient (decryptor) within a group of legitimate receivers. In other words group encryption provides receiver anonymity. A group encryption scheme allows a sender to prepare a ciphertext and convince a verifier that it can be decrypted by a member of the PKI group.

Group encryption involves a public-key encryption scheme with special properties, a group joining protocol that involves public-key certification, and a message space that may have a required structure. The three security protocols that pertain to Group Encryption schemes are:

i. *Security:* protect the sender from a hostile environment that tries to either extract information about the information that he tries to share.

ii. *Anonymity:* protects from extracting information about who the recipient of the message is.

iii. *Soundness:* protects the verifier from a hostile environment in which the sender, the group manager and the recipient may collude against him in order to accept a ciphertext that either does not have the required structure it cannot be decrypted by a registered group member.

In ENCASE we make use of group encryption schemes to enable minors to choose the users (recipients) of the sensitive images that she is willing to share.

6.4.3. Attribute-Based Encryption (ABE)

Attribute-Based Encryption (ABE) [78] is a cryptographic technique that reconsiders the concept of public-key cryptography. In the traditional public-key cryptography scheme, a message is encrypted for a specific receiver using the receiver’s public-key. Identity-Based Encryption [79] changed the idea of public-key cryptography by allowing the sender to encrypt a message using a public-key that is an arbitrary string (e.g., the email address of the recipient). Furthermore, Attribute-based encryption takes this one step further by defining the identity of a user as a set of attributes and the sender can encrypt a message using subsets of attributes (Key-Policy ABE) or using policies defined over a set of attributes (Ciphertext-Policy ABE). Then, the recipient is able to decrypt the ciphertext only if he holds a key for matching attributes where user keys are always issued by a trusted party.

An important security aspect of Attribute-Based Encryption is collusion-resistance in which an adversary that holds multiple keys should only be able to access data if at least one individual key grants access.

In ENCASE we will use Attribute-Based Encryption to enable minors to share their sensitive content, text and images, in Online Social Networks and only the users of their preference will be able to access it.

6.5. Infrastructure Design

This section describes the infrastructure design that will implement the reference architecture defined in this deliverable. More precisely, is a possible implementation of the Middleware and the Data analytics software stack (Back-end) components of the ENCASE architecture. As seen in Figure 20, we envision a complete Big Data Lambda architecture [98] which supports both real time and batch processing aspects of the ENCASE platform. Lambda is a generic, scalable and fault t-tolerant data processing architecture. The infrastructure design is generic enough so that it will support all the machine learning algorithms that will be implemented in the context of ENCASE and run in the Back-end of the architecture.

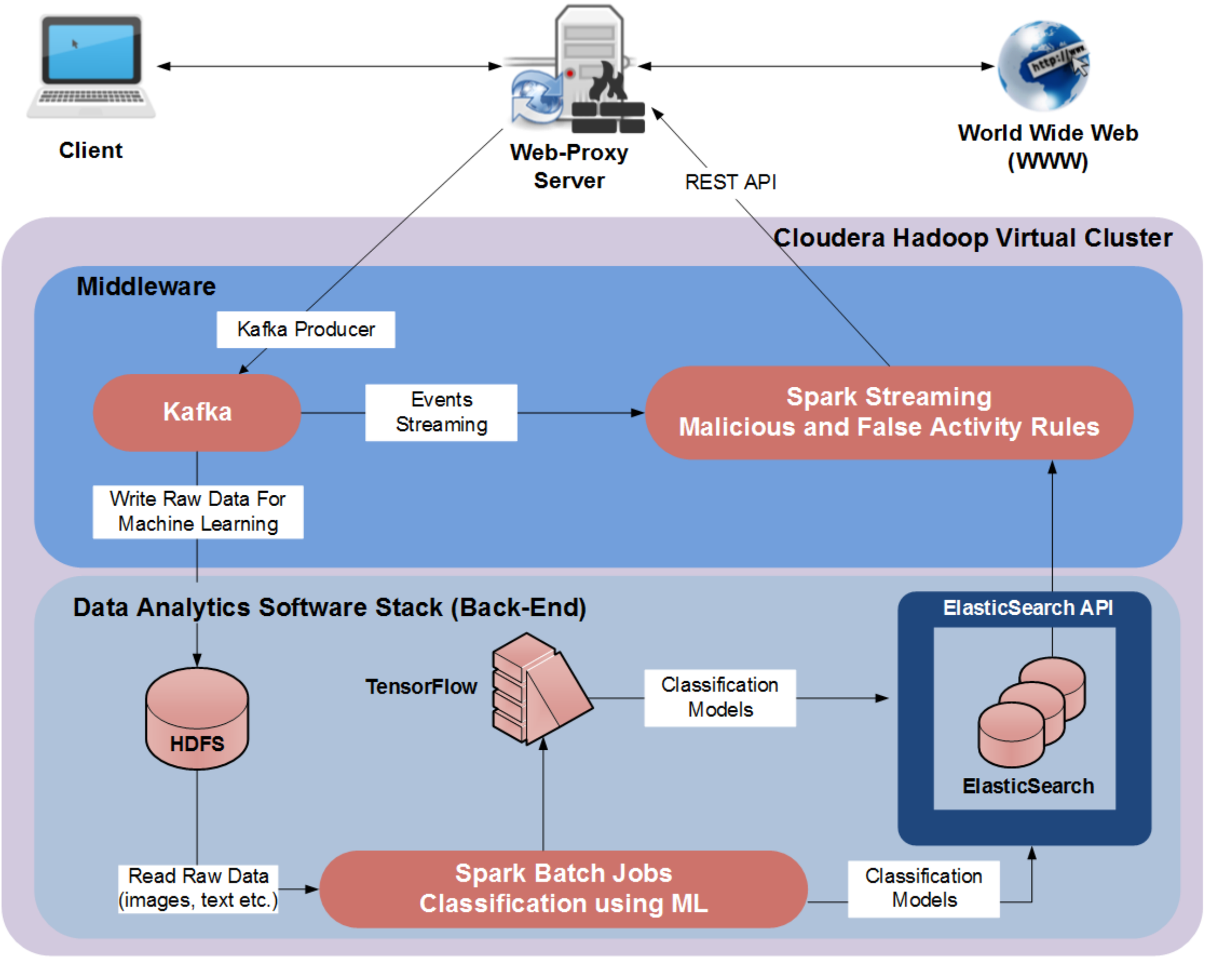

Client
Web-Proxy Server
World Wide Web (WWW)
REST API
Cloudera Hadoop Virtual Cluster
Middleware
Kafka Producer
Kafka
Events Streaming
Spark Streaming Malicious and False Activity Rules
Write Raw Data For Machine Learning
Data Analytics Software Stack (Back-End)
ElasticSearch API
TensorFlow
Classification Models
HDFS
ElasticSearch
Read Raw Data (images, text etc.)
Spark Batch Jobs Classification using ML
Classification Models

Figure 20. ENCASE Infrastructure design

Following is a description of the functionality of each one of the components that comprise the proposed infrastructure design, as depicted in Figure 20.

6.5.1. Middleware

Apache Kafka and Spark Streaming modules comprise the Middleware component of the ENCASE architecture. Both of them offer real-time processing of a large amount of events with only 1-2 seconds latency or even less.

6.5.1.1. Apache Kafka

Apache Kafka [99] is an open-source distributed streaming processing platform that aims to provide a unified, high-throughput, low-latency platform for handling real-time processing of data feeds. In ENCASE, Kafka will be used for accepting and handling of the incoming, from the Web-Proxy, captured OSN data. Its primary job is to store the received data in the HDFS for later batch processing as well as sending them as streams to Spark Streaming for real-time processing. Real-time processing happens using the rules generated from the Back-end of our architecture for the detection of malicious behaviour and fake identity or activity.

6.5.1.2. Apache Spark Streaming

Apache Spark Streaming [100] is an extension of the core Spark API that offers real-time processing of data received from various sources like Kafka, Flume, and HDFS. Its key abstraction is a Discretized Stream or, in short, a DStream, which represents a stream of data divided into small batches. DStreams are built on RDDs, Spark’s core data abstraction. This allows Spark Streaming to seamlessly integrate with any other Spark components like MLlib and Spark SQL. The rules that will be used for the detection of malicious behaviour or fake identity and activity detection are generated from the back-end of our architecture and can be written using PySpark (Python running on Spark). Furthermore, the PySpark code is capable of using external modeling classification frameworks like TensorFlow as well ElasticSearch for fuzzy word queries. Optionally real-time streaming Machine Learning models can also be implemented using Spark’s MLlib. In the end, the result of the processing, which primarily is whether to block or not a specific traffic, should be sent as a REST Response to the Web-Proxy.

6.5.2. Data Analytics Software Stack (Back-End)

The Data analytics software stack, or back-end, of our architecture is comprised of the HDFS module which is the storage of the captured information (images or texts) that will be then used as input for the machine learning classification algorithms (written in PySpark) for the generation of rules that will be send stored in the Middleware. Optionally, an external framework can be used from PySpark for the execution of the classification algorithms and for answering to queries generated from the Spark Streaming.

Finally, the back-end also includes the ElasticSearch module, a real-time distributed NoSQL database, which has a lot of out-of-the-box functionalities for running fuzzy text and natural language processing queries, with support for over 20 languages.

6.5.2.1. HDFS

HDFS [101], which stands for Hadoop Distributed File System, is a file system capable of storing large amount of structured and unstructured data. It is designed to reliably store very large files across distributed machines in a large cluster.

6.5.2.2. Apache Spark

Apache Spark [100] provides programmers an Application Programming Interface (API) centered on a data structure called the resilient distributed dataset (RDD), a read-only multi-set of data items distributed over a cluster of machines that are maintained in a fault-tolerant way. Spark's RDDs function as a working set for distributed programs that offers a (deliberately) restricted form of distributed shared memory. In the context of ENCASE, jobs classification models will be build using PySpark [102].

6.5.2.3. ElasticSearch

ElasticSearch [103, 104] is a distributed, scalable and Restful search and analytics engine that is able and can be used to solve growing number of use cases. ElasticSearch can be used to search all kind of documents that will be stored in HDFS. It uses Lucene and serves all of its functionality through a JSON and Java REST API. Also, it supports real-time GET requests, which makes it suitable for a NoSQL database. In ENCASE, we make use of ElasticSearch to host the classification models build by

Spark, and to support the text analysis of the captured OSN data, due to its real-time and text analysis functionalities.

6.5.2.4. TensorFlow

TensorFlow [105] is an open-source library for machine learning in various kinds of numerical computation using data flow graphs. In ENCASE can be used for perceptual and language analysis algorithms. Currently it is used for both research and production by more than 50 different teams in dozens of commercial Google products, such as speech recognition, Gmail, Google Photos, and search operation.

7. System Technical Requirements

This section contains the technical requirements of the ENCASE platform as they extracted from the use cases and the user stories. The technical requirements of each one of the four major architectural components of the ENCASE platform are listed and categorised according to Functional, Security & Privacy and Operational. Below we list a non-exhaustive list of requirements that are written in the form of user stories. Also, note that since we haven’t started the implementation yet and because we will employ an Agile methodology to the development of the platform the below list can be refined/updated during the course of the development.

7.1. Front-End

The Front-End of the ENCASE platform includes all the tools and web interfaces that will be made available to the users in order to enable and set up ENCASE on their devices and on their children’ devices. The front-end of ENCASE also includes the three browser add-ons that will be developed in the context of ENCASE for: a) malicious behaviour detection; b) fake identity and activity, and false information dissemination detection; and c) sensitive content detection and protection.

7.1.1. Functional Requirements

Register with the ENCASE platform

Code number	FE_RE_1
Title	Implementation of parent’s registration with ENCASE
Description	**As a parent** **I want to** be able to register with the ENCASE platform **so that** I can benefit from all the features that ENCASE offers
Acceptance criteria	1. The user should provide her email address and a hard-to-guess password in order to register. 2. Upon successful registration, the system displays a success message and sends a verification email to the user’s provided email address. 3. In order to proceed, the user should verify her email address using the link included in the verification email sent by the system.

Code number	FE_RE_2
Title	Implementation of child’s registration with ENCASE by parent
Description	**As a parent** **I want to** be able to perform registration in ENCASE for my children **so that** I can protect them from the threats exist in OSNs

Acceptance criteria	1. In order to do so, a parent should register herself first with ENCASE. 2. Child registration happens through the parent’s account management console. 3. The parent declares the email address of her child and the system registers the child generating a random password and linking its account with the parent’s account.

Code number	FE_RE_3
Title	Implementation of parent’s account login with ENCASE
Description	**As a parent** **I want to** be able login to the ENCASE platform from any device **so that** I can manage my account from anywhere
Acceptance criteria	1. In order to login, the user should provide her email address and her account’s password. 2. The system checks the provided credentials against the stored ones in the web-proxy’s account management module and logs the user in her account management console. 3. Upon successful login, the system displays a success message to the user.

Code number	FE_RE_4
Title	Implementation of parent’s account logout from ENCASE
Description	**As a parent** **I want to** be able to logout from the ENCASE platform **so that** I can prevent anyone else from accessing my account in a device
Acceptance criteria	1. The user should be able to logout from her ENCASE account management console with a single click. 2. Upon successful log out, the system displays a success message to the user.

Code number	FE_RE_5
Title	Implementation of deletion of parent’s account
Description	**As a parent** **I want to** have a delete account option **so that** I can delete my account and all the collected information about me when I want to
Acceptance criteria	1. The parent is able to completely delete her account and all of her children’ accounts. 2. Any personal details that ENCASE collected and stored about her or her children are also deleted. 3. Upon successful deletion, the system displays a success message to the user.

Code number	FE_RE_6
Title	Connect child’s device with the ENCASE Web-Proxy
Description	**As a parent** **I want to** be able to connect my child’s device with ENCASE web-proxy **so that** the system can monitor its OSN activity and notify me
Acceptance criteria	1. The parent can connect her child’s device with the web-proxy via: a) system-wide proxy settings on a desktop or laptop; or b) a VPN connection on any other device. 2. Upon successful connection, the system should display a success message and all the traffic on the connected device should pass through the web-proxy and captured.

ENCASE parent account preferences

Code number	FE_AP_1
Title	Implementation of authorization of ENCASE to access a child’s OSN account
Description	**As a parent** **I want to** be able to authorize ENCASE to access an OSN account of my child **so that** I ensure that my child will be protected in this OSN
Acceptance criteria	1. The authorization happens using OAuth2.0 2. Upon authorization, the system should retrieve all the available OSN account information of the child. 3. The authorization token obtained by the system should be stored to be used later.

Code number	FE_AP_2
Title	Update child’s OSN accounts credentials
Description	**As a parent** **I want to** update the credentials of my child’s OSN credentials **so that** ENCASE will still have access to the account when my child changes his account’s credentials
Acceptance criteria	1. When the parent has updated the credentials she has to re-authorize ENCASE to access the OSN account. 2. The process should call the account management to replace the old authorization token for this OSN account with the new on.

Code number	FE_AP_3
Title	De-association of child’s OSN accounts from ENCASE
Description	**As a parent** **I want to** have a management console for my children’s OSN accounts **so that** I am able to de-associate one or more OSN accounts from ENCASE and stop being monitored
Acceptance criteria	1. When the parent de-associates an OSN account of her child from ENCASE, the stored authorization token for this account should be invalidated and deleted. 2. From now on, ENCASE should not monitor or check this OSN account.

Code number	FE_AP_4
Title	Declaration of web filters for the web-proxy
Description	**As a parent** **I want to** be able to set the web pages that I don’t like **so that** ENCASE will block my child from visiting them
Acceptance criteria	1. The list of the declared web-pages should be stored in the parents’ account. 2. The minors’ associated with the parent’s account should not be able to access those web pages and 3. When the minors’ associated with the parent’s account try to access one of the blacklisted web-pages the web-proxy should block them and display the appropriate message informing them that the page is blocked due to parent request.

Code number	FE_AP_5
Title	Declaration of whitelist for the web-proxy
Description	**As a parent** **I want to** be able to set a whitelist with all the persons that I trust **so that** ENCASE will not send notifications to me for those persons
Acceptance criteria	1. The user should provide the OSN account names of the people that he trust to communicate with her children. 2. The declared whitelist should be stored and associated with the parent’s ENCASE account. 3. For each one of the whitelisted users, the parent should be able to choose among the following: a. Monitor and notify in severe threats, or b. Do not monitor in any case.

Code number	FE_AP_6
Title	Implementation of easy installation of browser add-ons on child’s device
Description	**As a parent** **I want to** be able install all the browser add-ons of ENCASE with one click on my child’s device **so that** the installation is easy and fast
Acceptance criteria	1. When the user is logged in, with one click all three browser add-ons should be downloaded and installed in the browser. 2. Upon successful installation, the system should display a success message.

Code number	FE_AP_7
Title	Prevent minors from sharing personal information with strangers
Description	**As a parent** **I want to** be able to set my child’s personal information **so that** ENCASE will block such information (like address, school, etc.) when my child is trying to share them with a stranger in OSNs
Acceptance	1. When a minor tries to share her personal information the system

criteria	should check whether this information matches those set from the parent from before and if yes then block and display the appropriate message to the minor. 2. If the information that the minor tries to share is not included in those defined by the parent, then the parent should be notified and the system should ask her whether she prefers to block this information from sharing or not.

Code number	FE_AP_8
Title	Implementation of preference definition for OSN data donation
Description	**As a parent** **I want to** be able to set whether I wish or not to donate my child’s OSN data to ENCASE **so that** I can benefit from the extra features that ENCASE offers
Acceptance criteria	1. User preference should be stored to the account management module. 2. If the parent wishes to donate her child’s OSN data to ENCASE then the web-proxy should capture all of the child’s OSN traffic and send them to the back-end.

User Notifications

Code number	FE_UN_1
Title	Implementation of Notification module
Description	**As a parent** **I want to** get notified when my child is being exposed to a threat in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. Upon detection of a threat, the notification module should be called providing it with the necessary information that should be included in the notification and with the information of the user that should receive the notification. 2. The notification generates the notification with the provided information and sends it to the declared receiver. 3. The notification should be send via email and if she is logged into the system it should be displayed as a message in her browser. 4. The notification should also include a list of help lines that the parent can call depending on the problem and the country. 5. The notification should also include a list of suggested actions depending on the problem.

Code number	FE_UN_2
Title	Notification when minor creates a new OSN account
Description	**As a parent** **I want to** get notified when my child creates a new account in an Online social network **so that** I authorise ENCASE to access and monitor the activity on this account
Acceptance criteria	1. When the system detects the name of the child in an OSN that the system does not have in its record it should send a notification the parent.

	2. The middleware will check for this in the traffic that the web-proxy captures. 3. When detected, the middleware should call the notification module to generate and send the notification.

Code number	FE_UN_3
Title	Notification when ENCASE loses access to one of a minor’s OSN accounts
Description	**As a parent** **I want to** get notified when the system loses access to one of my child’s OSN accounts **so that** I am aware
Acceptance criteria	1. When the middleware failed to access one of a minor’s OSN accounts it should call the notification module to generate and send the appropriate notification the minor’s parent.

Code number	FE_UN_4
Title	Notification when a minor’s device disconnects from the ENCASE Web-Proxy server
Description	**As a parent** **I want to** get notified when my child’s device has been disconnected from the ENCASE web-proxy **so that** I am aware and re-connect it
Acceptance criteria	1. When the web-proxy stops receiving traffic from a minor’s device for a certain amount of time it should call the notification module to generate and send a notification to the minor’s parent.

Code number	FE_UN_5
Title	Notification when a minor becomes a victim of cyber bullying in an OSN
Description	**As a parent** **I want to** get notified when my child has become a victim of cyber bullying in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The middleware’s malicious behaviour detection module should analyse all of the user’s captured OSN traffic to detect incidences when the minor is being a victim of cyber bullying. 2. Detection can also happen by periodically checking the minors’ OSN account information. 3. Upon detection the notification module should be called to generate and send a notification to the parent. 4. The notification should include a list of cyber bullying help lines that the parent can contact. 5. The account that performed the cyber bullying should be stored in the cyber bullying reputation list.

Code number	FE_UN_6
Title	Notification when a minor is committing cyber bullying in an OSN

Description	**As a parent** **I want to** get notified when my child is committing cyber bullying in an OSN **so that** I am able to take the appropriate measures and stop it
Acceptance criteria	1. The middleware’s malicious behaviour detection module should analyse all of the user’s captured OSN traffic to detect incidences when the minor is committing cyber bullying. 2. Detection can also happen by periodically checking the minors’ OSN account information. 3. Upon detection the notification module should be called to generate and send a notification to the parent.

Code number	FE_UN_7
Title	Notification when a child becomes a victim of sexual cyber grooming in an OSN
Description	**As a parent** **I want to** get notified when my child has become a victim of sexual cyber grooming in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The middleware’s malicious behaviour detection module should analyse all of the user’s captured OSN traffic to detect incidences when the minor has become a victim of sexual cyber grooming. 2. Detection can also happen by periodically checking the minors’ OSN account information. 3. Upon detection the notification module should be called to generate and send a notification to the parent. 4. The account that performed the sexual cyber grooming should be stored in the sexual abusers reputation list.

Code number	FE_UN_8
Title	Notification when a child threatening messages in an OSN
Description	**As a parent** **I want to** get notified when my child has become a victim of cyber bullying by receiving threatening messages in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The middleware’s malicious behaviour detection module should analyse all of the user’s captured OSN traffic to detect incidences when the minor is being a victim of cyber bullying by receiving threatening messages. 2. Detection can also happen by periodically checking the minors’ OSN account information. 3. Upon detection the notification module should be called to generate and send a notification to the parent. 4. The notification should include a list of cyber bullying help lines that the parent can contact. 5. The account that performed the cyber bullying should be stored in the cyber bullying reputation list.

Code number	FE_UN_9

Title	Notification when someone mentions a minor in a cyber bullying post in an OSN
Description	**As a parent** **I want to** get notified when someone mentions my child in a cyber bullying post in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The middleware’s malicious behaviour detection module should analyse all of the user’s captured OSN traffic to detect when he is mentioned in a post that contains cyber bullying. 2. Detection can also happen by periodically checking the minors’ OSN account information. 3. Upon detection the notification module should be called to generate and send a notification to the parent. 4. The notification should include a list of cyber bullying help lines that the parent can contact. 5. The account that performed the cyber bullying should be stored in the cyber bullying reputation list.

Code number	FE_UN_10
Title	Report of distressed or aggressive behaviour
Description	**As a parent** **I want to** get notified when my child is experiencing or is about to experience distressed or aggressive behaviour **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The middleware should analyse the minor’s OSN activity trying to detect when he is experiencing or is about to experience distressed or aggressive behaviour. 2. Upon detection the notification module should be called to generate and send a notification to the parent.

Code number	FE_UN_11
Title	Notification when a minor is communicating with someone who has bad reputation for cyber bullying
Description	**As a parent** **I want to** get notified when my child is communicating with someone who has bad reputation for cyber bullying **so that** I am aware and take the appropriate measures
Acceptance criteria	1. Every user that a minor is communicating with in OSNs should be checked against the ENCASE cyber bullying reputation list. 2. If the system detects that the minor is communicating with a user who is included in this list then it should call the notification module to generate and send a notification to the parent.

Code number	FE_UN_12
Title	Notification when a minor is communicating with someone with fake identity in an OSN
Description	**As a parent**

	I want to get notified when my child is communicating with someone with fake identity in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. The middleware should analyse the profile of the users that a minor is communicating with. 2. When one of those OSN profiles is detected to be fake, the notification module should be called to generate and send a notification to the minor’s parent. 3. The detected fake identity should also be stored to the ENCASE fake identity reputation list. 4. From the notification, the parent should be able to report this user.

Code number	FE_UN_13
Title	Notification when a minor receives a friend request in an OSN from someone with a fake identity
Description	**As a minor** **I want to** get notified when I receive a friend request in an OSN from someone who has fake identity **so that** I am able to ignore the request and report the account
Acceptance criteria	1. The middleware should periodically check the OSN account of the minor to detect any new friend request. 2. If a new friend request is detected then the system should collect the available information for this profile and check if it is a fake identity. 3. If the profile of the user that sent the friend request to the minor is detected to be fake, then the notification module should be called to generate and send the appropriate notification to the minor. 4. The detected fake identity should also be stored to the ENCASE fake identity reputation list. 5. From the notification, the minor should be able to report this user.

Code number	FE_UN_14
Title	Report of false information dissemination
Description	**As a minor** **I want to** get notified when someone is posting false information about me in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. Upon detection, the fake identity and false information detection module of the middleware should call the notification module to generate and send the appropriate notification to the minor.

Code number	FE_UN_15
Title	Notification when a minor receives false information
Description	**As a minor** **I want to** get notified when I receive a false information in OSNs **so that** I am able to ignore it
Acceptance criteria	1. The user’s OSN traffic should be captured by the web-proxy and

	analysed by the middleware. 2. When the fake identity and false information detection module detects false information in the minor’s captured traffic then it should call the notification module to generate and send the appropriate notification to the minor.

Code number	FE_UN_16
Title	Notification for sensitive content detection
Description	**As a parent** **I want to** get notified when my child is about to share an image or text with sensitive content in an OSN **so that** I am aware and take the appropriate measures
Acceptance criteria	1. All the outgoing traffic of a minor should be captured and analysed by the web-proxy to check whether it contains sensitive images. 2. When an image is detected, the nudity detection module should be called and if the images contains over 80% nudity then the notification module should be called to generate and send the appropriate notification to the minor’s parent.

User Triggers

Code number	FE_UT_1
Title	Implementation of reporting a malicious behaviour
Description	**As a minor** **I want to** be able to flag a conversation **so that** I am able to report a malicious behaviour
Acceptance criteria	1. The malicious behaviour detection browser add-on should allow the user to mark the conversation that contains malicious behaviour. 2. The flagged conversation should be captured and sent to the back-end for further analysis.

Code number	FE_UT_2
Title	Implementation of reporting a fake identity
Description	**As a minor** **I want to** be able to flag a user in an OSN **so that** I am able to report a suspicious fake identity
Acceptance criteria	1. The fake identity and activity detection browser add-on should allow the user to flag an OSN account that she thinks is fake. 2. The information of the flagged account should be sent to the back-end for further analysis and if it is indeed a fake identity then the flagged account should be added to the ENCASE fake identity reputation list.

7.1.2. Operational Requirements

#	Operational Requirements	Rational / Comments	Mandatory (YES/NO)
O1.1	The ENCASE web-interface should be	-	YES

	intuitive in its use by both advanced social media users and amateurs.		
O1.2	The browser add-ons should avoid complicated calculation in its interface, to avoid any performance issues.	All the analysis of the collected information should be performed all the other components of the ENCASE ecosystem.	YES
O1.3	The user must register with her email address.	-	YES
O1.4	The user (parent) should be able to link her account with her child’s account.	-	YES
O1.5	The user should be able to login with ENCASE.	-	YES
O1.6	Authentication should be fast (e.g., < 1 second).	-	NO
O1.7	The user should be able to delete her account.	-	YES
O1.8	The parent should be able to modify her account preferences anytime.	-	YES
O1.9	The system should have authorization to access a user’s OSN account.	-	YES
O1.10	Installation of browser add-ons should be easy (with one-click).	-	YES
O1.11	The browser add-ons should be able to show notifications to the user’s screen.	-	YES
O1.12	The user’s browser must be Google Chrome or Firefox.	-	YES
O1.13	The parent should be able to access her account management console from both mobile and desktop browser.	-	YES
O1.14	The parent should be able to easily modify her notification preferences.	-	YES
O1.15	The user should be able to easily recover from errors.	The user should be able to disconnect a device that by mistake requested link.	YES

7.1.3. Security and Privacy Requirements

#	Security & Privacy Requirements	Rational / Comments	Mandatory (YES/NO)
S&P1.1	The system should not leak any user’s personal information to other browser add-ons or to social networks.	-	YES
S&P1.2	All communication channels within ENCASE framework must use communication protocols that provide communication security (e.g., SSL or TLS).	-	YES
S&P1.3	The system should not leak any user’s personal information to other browser add-ons or to social networks.	-	YES
S&P1.4	All users personal information will not be linked to a physical person	-	YES
S&P1.5	Browser add-ons will not store any user’s personal information locally.	-	YES
S&P1.6	The system should monitor minors’ OSN activity based on their age to avoid violating their privacy.	-	YES

7.2. Web-Proxy Server

7.2.1. Functional Requirements

Code number	WPS1
Title	Connections to the Web-Proxy
Description	**As a web proxy** **I want to** be able to accept connections from users **so that** I all of their traffic passes through me and I am able to analyse it and protect them
Acceptance criteria	1. A user can connect to the web-proxy using system-wide proxy settings or through a VPN connection. 2. When connected, the user’s traffic will pass through the web-proxy. 3. When a connection is successful, the web-proxy should be able to monitor and capture the user’s traffic.

Code number	WPS2
Title	Implementation of TLS termination
Description	**As a web proxy** **I want to** be able to terminate TLS connection **so that** i can decrypt and process the traffic in order to be able to block any incoming or outgoing traffic when I detect malicious behaviour, distressed

	behaviour, etc. in a conversation
Acceptance criteria	1. The web-proxy should handle TLS connections. 2. The web-proxy should decrypt a TLS and inspect the unencrypted request.

Code number	WPS3
Title	Implementation of Web filtering
Description	**As a web proxy** **I want to** have a list of web sites **so that** i can perform web filtering and block them when a minor tries to access them
Acceptance criteria	1. The web-proxy should screen an incoming web page to determine whether it is included in the list of web pages that the parent of the user, who requested the web page, declared. 2. The web-proxy should block the rendering of the web page if it included in the list of the web pages that should be blocked. 3. Instead of the web-page, the web-proxy should display a message informing the minor that the web page is blocked due to her parent’s request.

Account Management

Code number	WPS_AM_1
Title	Implementation of user authentication
Description	**As a web proxy** **I want to** be able to authenticate users when they connect to me **so that** i can uniquely identify each user
Acceptance criteria	1. In order to authenticate the user should provide her email address and password declared during registration. 2. Upon successful login, the system should display a success message.

Code number	WPS_AM_2
Title	Implementation of user registration in ENCASE
Description	**As a web proxy** **I want to** be able to receive the user’s basic information and preferred credentials **so that** i can perform user registration
Acceptance criteria	1. In order to register the user should provide her email address and a hard-to-guess password. 2. Before registration, the system should check if the provide email address is not already registered with ENCASE. 3. Upon successful registration, the system should display a success message to the user.

Code number	WPS_AM_3
Title	Implementation of age periodical check

Description	**As a web proxy** **I want to** periodically check and update a user’s age **so that** i can yield the appropriate notifications and not violate his privacy
Acceptance criteria	1. The system should periodically take the user’s declared date of birth and compute his current age. 2. If a user at some time is over 18 years old then the web proxy should not notify her parents of her web activity.

Code number	WPS_AM_4
Title	Implementation of account deletion
Description	**As a web proxy** **I want to** be able to delete all the information that ENCASE stored and captured about a user **so that** i can completely delete a user’s account when this is requested by him
Acceptance criteria	1. A parent should be able to request account deletion. 2. When the web-proxy deletes a parent’s account it should also delete all of her children’s accounts. 3. Any personal details that ENCASE collected and stored about the parent or her children are also deleted. 4. Upon successful deletion, the system displays a success message to the user.

Traffic Monitoring and capturing

Code number	WPS_TMC_1
Title	Incoming traffic monitoring for malicious behaviour detection
Description	**As a web proxy** **I want to** be able to capture all the incoming traffic before it is delivered to the user **so that** i can detect cyberbullying, sexual assault, etc.
Acceptance criteria	1. The web proxy should capture and parse all the incoming traffic in the network. 2. It should use the pre-installed rules to detect any type of malicious behaviour. 3. In the case that malicious behaviour is detected in a conversation the web-proxy should block the traffic before it is delivered to the user and terminate the TLS connection and yields the appropriate warnings.

Code number	WPS_TMC_2
Title	Outgoing traffic monitoring for distressed or aggressive behaviour detection
Description	**As a web proxy** **I want to** be able to capture all the outgoing traffic in a user **so that** i can detect whether a user in the network exhibits distressed behaviour
Acceptance criteria	1. The web proxy should capture and parse all the going traffic in the network. 2. It should use the pre-installed rules to detect whether a user in the network exhibits distressed behaviour.

	3. In the case when distressed behaviour is detected the web proxy should terminate the appropriate TLS connection of the user and yields the appropriate warnings.

Code number	WPS_TMC_3
Title	Implementation of OSN data donation to Back-end
Description	**As a web proxy** **I want to** know whether the user wishes to contribute her OSN data **so that** i can capture her traffic and send it to the back-end
Acceptance criteria	1. The web-proxy should be able to communicate with the back-end to send the user’s captured OSN data. 2. The back-end should associate the donated data with the appropriate registered ENCASE user. 3. When a user donates her OSN data, she should have access to extra features of the platform.

Code number	WPS_TMC_4
Title	Implementation of Back-end API call to send users’ OSN donated data
Description	**As a web proxy** **I want to** be able to call the back-end **so that** i can submit the users’ captured OSN data
Acceptance criteria	1. The web-proxy should call the back-end API in order to send the user’s OSN captured data. 2. Calls to the back-end API should be authenticated.

Code number	WPS_TMC_5
Title	Implementation of traffic submission to the Middleware
Description	**As a web proxy** **I want to** be able to call the middleware **so that** i can submit the users’ captured traffic for analysis
Acceptance criteria	1. All the traffic that is not flagged by the users should be sent to the middleware. 2. The web-proxy should call the middleware API in order to send the user’s captured traffic for malicious behaviour and fake identity and activity detection. 3. Calls to the middleware API should be authenticated.

Sensitive content detection and protection

Code number	WPS_SCDP_1
Title	Implementation of sensitive content detection in incoming traffic
Description	**As a web proxy** **I want to** be able to detect nudity in the photos sent to a user in the network **so that** i can prevent their rendering.
Acceptance criteria	1. The web proxy should determine the level of nudity included in the photos sent to a user in the network.

	2. Nudity detection will performed using existing JavaScript nudity detection libraries. 3. When the web-proxy detects a photo that contains over 80% nudity in a photo then it should notify the sensitive content detection and protection browser add-on to prevent the rendering of that photo.

Code number	WPS_SCDP_2
Title	Implementation of sensitive content detection in outgoing traffic
Description	**As a web proxy** **I want to** be able to detect nudity in the photos shared by a user in the network **so that** i can notify the user to protect it.
Acceptance criteria	1. The web proxy should determine the level of nudity included in the photos shared by a user in the network. 2. Nudity detection will performed using existing JavaScript nudity detection libraries. 3. When the web-proxy detects a photo that contains over 80% nudity in a photo then it should notify the user through the sensitive content detection and protection browser add-on to protect the photo before it is shared.

Code number	WPS_SCDP_3
Title	Implementation of sensitive content protection using steganography
Description	**As a web proxy** **I want to** be able to steganographise a sensitive content (image or text) **so that** i prevent unauthorised access to it
Acceptance criteria	1. When selected by the user, the web-proxy should call a steganography algorithm that encrypts the sensitive image or text with steganography. 2. The receiver of the sensitive content should have the appropriate software in his device in order to de-steganographise it and access the actual content.

Code number	WPS_SCDP_4
Title	Implementation of sensitive content protection using attribute-based encryption
Description	**As a web proxy** **I want to** be able to encrypt a sensitive content (image or text) using an attribute **so that** i can perform attribute-based encryption and prevent any user who is not the holder of the attribute to decrypt and view the actual content
Acceptance criteria	1. The web-proxy should have the appropriate algorithm to perform attribute based encryption and encrypt the user’s sensitive content using a specific attribute. 2. All the other users should be the holders of this attributes to be able to decrypt and view the actual content.

Code number	WPS_SCDP_5
Title	Implementation of sensitive content protection using group encryption

Description	**As a web proxy** **I want to** be able to encrypt a sensitive content (image or text) using a group signature **so that** i can perform group encryption and prevent any user who is not in the group to decrypt and view the actual content
Acceptance criteria	1. The web-proxy should have the appropriate algorithm to perform group encryption. 2. In order to user group encryption, a user should specify the group of users that she wishes to be able to access and view the actual content. 3. When selected by the user, the web-proxy should encrypt the sensitive content of the user and only users who are in the group specified by the user should be able to decrypt and view the actual content.

Code number	WPS_SCDP_6
Title	Implementation of sensitive images blocking
Description	**As a web proxy** **I want to** be able to block sensitive images of passing through unencrypted **so that** i can protect it from unauthorised access when it is not secured
Acceptance criteria	1. The web-proxy should analyse the outgoing traffic of the user before going out of the network in order to detect sensitive images that contain nudity. 2. When an image detected the process should call the nudity detection module to determine the level of nudity. 3. If there is nudity in the image and the image is not encrypted then the web-proxy should block the traffic from not going out of the network. 4. Instead, the system should offer to the user ways to encrypt the sensitive image before sharing it.

Code number	WPS_SCDP_7
Title	Implementation of personal information blocking
Description	**As a web proxy** **I want to** be able to detect when a minor tries to share personal information **so that** I block them from being shared with strangers in OSNs
Acceptance criteria	1. The web-proxy should analyse the outgoing traffic of the user before send out of the network in order to detect any personal information about him. 2. When a minor tries to share her personal information the system should check whether this information matches those set from the parent from before and if yes then block and display the appropriate message to the minor. 3. If the information that the minor tries to share is not included in those defined by the parent, then the parent should be notified and the system should ask her whether she prefers to block this information from sharing or not.

7.2.2. Operational Requirements

#	Operational Requirements	Rational / Comments	Mandatory (YES/NO)
O2.1	The web-proxy should be able to handle TLS connections.	-	YES
O2.2	The web-proxy should be able to perform TLS decryption.	-	YES
O2.3	The web-proxy should be able to perform TLS encryption.	-	YES
O2.4	The web-proxy should be able to terminate TLS connections.	-	YES
O2.5	The web-proxy should be able to enable or disable OCSP responses.	-	YES
O2.6	The web-proxy should be able to handle SSL/TLS ciphers.	-	YES
O2.7	The web-proxy should be able to perform certificate management.	-	YES
O2.8	The web-proxy should be able to manage chained certificates.	-	YES
O2.9	The web-proxy should monitor and capture all the incoming traffic.	-	YES
O2.10	The web-proxy should monitor and capture all the outgoing traffic.	-	YES
O2.11	The web-proxy should be able to block incoming traffic.	-	YES
O2.12	The web-proxy should be able to block outgoing traffic.	-	YES
O2.13	The web-proxy should perform web filtering.	-	YES
O2.14	The web-proxy should perform traffic shaping.	-	YES
O2.15	The web-proxy should be able to authenticate users.	-	YES
O2.16	The account management module of the web-proxy should keep a record of all the registered users’ credentials.	-	YES

O2.17	The web-proxy should be able to identify which traffic belongs to which user of ENCASE.	-	YES
O2.18	All users should register to the web-proxy with their email address.	-	YES
O2.19	The account management web interface should be user friendly.	-	YES
O2.20	All users’ web activity should be logged.	-	YES
O2.21	The log files should be treated in accordance with the data protection laws.	-	YES
O2.22	The web-proxy should be able to detect nudity in images included in the captured traffic.	-	YES
O2.23	Nudity detection should be fast and accurate.	-	YES
O2.24	The web-proxy should be able to encrypt sensitive images with steganography.	-	YES
O2.25	The web-proxy should be able encrypt sensitive images using attribute-based encryption.	-	YES
O2.26	The web-proxy should be able to encrypt sensitive images using group encryption.	-	YES
O2.27	Using account management the parent should be able to use all the features of ENCASE and choose what ENCASE should monitor and report about her children.	-	YES
O2.28	The web-proxy should be able to manage high network traffic load.	-	YES

7.2.3. Security and Privacy Requirements

#	Security & Privacy Requirements	Rational / Comments	Mandatory (YES/NO)
S&P2.1	The web-proxy should be security-hardened so that it is very difficult to compromise and so that no unauthorized entity is able to access and steal the user’s account data.	-	YES

S&P2.2	All communication between other modules must occur over a secure channel.	-	YES
S&P2.3	All the users’ donated OSN data should be encrypted before transferred to the back-end.	-	YES
S&P2.4	User authentication module should detect and reject malicious authentication attempts.	-	YES
S&P2.5	The web-proxy should obtain explicit consent from the user to capture her OSN traffic.	-	YES
S&P2.6	The web-proxy should obtain explicit user consent to send and store her OSN traffic to the back-end.	-	YES
S&P2.7	The web-proxy should securely capture user’s social network incoming and outgoing traffic.	-	YES
S&P2.8	The account management module should secure store users’ personal information	-	YES

7.3. Middleware

7.3.1. Functional Requirements

Code number	MW_1
Title	Receive captured traffic from the web-proxy
Description	**As a middleware** **I want to** receive the captured traffic from the web-proxy when the user does not flag the content **so that** i can analyse it for malicious behaviour or fake activity
Acceptance criteria	1. Based on those rules the system should decide whether the specific traffic is suspicious or not. 2. The system should decide whether a suspicious content should be checked for malicious behaviour or for fake identity or activity.

Code number	MW_2
Title	Implementation of predefined rules traffic check
Description	**As a middleware** **I want to** have predefined rules installed **so that** i can check all the received users’ traffic against them
Acceptance criteria	1. Based on those rules the system should decide whether the specific traffic is suspicious or not. 2. The system should decide whether a suspicious content should be

	checked for malicious behaviour or for fake identity or activity.

Code number	MW_3
Title	Implementation of content extraction from suspicious captured traffic
Description	**As a middleware** **I want to** be able to extract the content from captured suspicious traffic **so that** i can analyse it for malicious behaviour or fake activity
Acceptance criteria	1. The middleware should have the appropriate software to extract the html from the user’s OSN captured traffic.

Code number	MW_4
Title	Implementation of cyber bullying detection
Description	**As a middleware** **I want to** have cyber bullying detection rules installed **so that** i can detect cyber bullying in suspicious content
Acceptance criteria	1. The middleware should use the detection rules received from the back-end in order to analyse a suspicious content for cyber bullying. 2. When cyber bullying is detected in a minor’s OSN traffic, the notification module should be called to send a notification to the minor’s parent.

Code number	MW_5
Title	Implementation of sexual cyber grooming detection
Description	**As a middleware** **I want to** have sexual cyber grooming detection rules installed **so that** i can detect sexual cyber grooming in suspicious content
Acceptance criteria	1. The middleware should use the detection rules received from the back-end in order to analyse a suspicious content for sexual cyber bullying. 2. When sexual cyber grooming is detected in a minor’s OSN traffic, the notification module should be called to send a notification to the minor’s parent.

Code number	MW_6
Title	Implementation of malicious behaviour reputation list
Description	**As a middleware** **I want to** have a malicious behaviour reputation list **so that** i can refer to it when I am performing malicious behaviour detection
Acceptance criteria	1. Each time a malicious behaviour is detected the perpetrator’s OSN account should be added in the appropriate malicious behaviour reputation list. 2. When the middleware analyses a suspicious content for malicious behaviour, this reputation list should be take into account.

Code number	MW_7
Title	Implementation of fake identity detection
Description	**As a middleware** **I want to** have fake identity detection rules installed **so that** i can detect fake identities in OSNs
Acceptance criteria	1. The middleware should use the detection rules received from the back-end in order to check whether a suspicious OSN profile is a fake identity. 2. When a fake identity is detected the notification module should be called to send a notification to the minor.

Code number	MW_8
Title	Implementation of fake identity reputation list
Description	**As a middleware** **I want to** have a fake identity reputation list **so that** i can refer to it when I am performing fake identity detection
Acceptance criteria	1. Each time a fake identity is detected it should be added in the fake identity reputation list. 2. When an OSN profile is being analysed for fake identity or activity the system should first check if it is included in the fake identity reputation list.

Code number	MW_9
Title	Implementation of false information dissemination detection
Description	**As a middleware** **I want to** have false information detection rules installed **so that** i can detect false information dissemination in OSN captured traffic
Acceptance criteria	1. The middleware should use the detection rules received from the back-end in order to check whether a suspicious content contains false information. 2. Each time false information is detected in a minor’s captured OSN traffic the notification module should be called to send a notification to the minor.

Code number	MW_10
Title	Receive detection rules from the back-end
Description	**As a middleware** **I want to** receive and store detection rules from the back-end **so that** i can use them while processing the html text for malicious behaviour detection, distressed behaviour detection, etc.
Acceptance criteria	1. The web proxy should periodically receive new or updated detection rules from the back-end. 2. Those rules should be stored in a secure directory so that it cannot be accessed and modified by malicious users.

Code number	MW_11
Title	Implementation of periodic check of children OSN accounts
Description	**As a middleware** **I want to** have authorization from parents **so that** i can periodically check the children OSN accounts for malicious behaviour, and fake identity and activity detection
Acceptance criteria	1. The middleware should use the corresponding API offered by each social network to periodically retrieve the minor’s OSN accounts’ information and activity. 2. The retrieved information is then analysed for malicious behaviour, fake identity and activity and false information detection.

Code number	MW_12
Title	Implementation of notification module
Description	**As a middleware** **I want to** be able to call the notification module **so that** i can create and send notifications to parents and minors
Acceptance criteria	1. The module should receive all the required information in order to generate a notification, including the email and id of the receiver. 2. The notification should contain the appropriate information to be informative enough but at the same time not violating the minor’s privacy.

7.3.2. Operation Requirements

#	Operational Requirements	Rational / Comments	Mandatory (YES/NO)
O3.1	All requests to the middleware API should be authenticated and authorized.	-	YES
O3.2	The middleware should be able to receive the captured traffic from the web-proxy.	-	YES
O3.3	The middleware should be able to extract the actual content from the captured traffic.	-	YES
O3.4	The middleware should have predefined preliminary rules.	-	YES
O3.5	The middleware should check all the received captured traffic against the preliminary rules.	-	YES
O3.6	The middleware should receive and store detection rules generated from the back-end.	-	YES
O3.7	The middleware should be able to detect	-	YES

	malicious behaviour in suspicious content using detection rules.		
O3.8	The middleware should be able to detect fake identities using detection rules.	-	YES
O3.9	The middleware should be able to detect false information in suspicious content using detection rules.	-	YES
O3.10	The middleware should periodically check the children’ OSN accounts for malicious behaviour detection.	-	YES
O3.11	The detection of a threat should be as accurate as possible.	-	YES
O3.12	The notification module should be able to generate notifications.	-	YES
O3.13	The generated notifications should be informative enough.	-	YES
O3.14	The parents should timely receive notifications about an incidence (e.g., malicious behaviour)	Upon detection of a threat the notification should be send fast (e.g., less than 5 sec.)	YES

7.3.3. Security and privacy Requirements

#	Security & Privacy Requirements	Rational / Comments	Mandatory (YES/NO)
S&P3.1	All communication between other modules must occur over a secure channel.	-	YES
S&P3.2	The middleware must obtain explicit authorization from the user in order to collect her personal OSN account information.	-	YES
S&P3.3	The middleware should securely collect users’ social network information.	-	YES
S&P3.4	The exchange of information between the middleware and the web-proxy must always be encrypted, with protocols such as SSL.	-	YES
S&P3.5	The middleware should protect the content that receives from the web-proxy	-	YES

	from leakage.		
S&P3.6	The generated notifications must not contain sensitive personal information in order to avoid violating minors' privacy.	-	YES

7.4. Data Analytics Software stack (Back-End)

7.4.1. Functional Requirements

Code number	BE_1
Title	Install detection rules in the middleware
Description	**As the data analytics software stack** **I want to** be able to call the middleware **so that** i can install newly generated detection rules in the middleware
Acceptance criteria	1. The back-end should use the middleware API to communicate with the middleware and send the generated detection rules. 2. This process should happen in a predefined schedule.

Code number	BE_2
Title	Cyber bullying detection machine learning algorithm
Description	**As the data analytics software stack** **I want to** have a machine learning algorithm that detects cyber bullying **so that** i can generate cyber bullying detection rules
Acceptance criteria	1. The algorithm will use the user's OSN data to be trained. 2. The algorithm should run in a predefined schedule. 3. The generated detection rules should be sophisticated enough to produce more accurate cyber bullying predictions.

Code number	BE_3
Title	Sexual cyber grooming detection machine learning algorithm
Description	**As the data analytics software stack** **I want to** have a machine learning algorithm that detects sexual cyber grooming **so that** i can generate sexual cyber grooming detection rules
Acceptance criteria	1. The algorithm will use the user's OSN data to be trained. 2. The algorithm should run in a predefined schedule. 3. The generated detection rules should be sophisticated enough to produce more accurate sexual cyber grooming predictions.

Code number	BE_4
Title	Sexually abusive behaviour detection machine learning algorithm
Description	**As the data analytics software stack** **I want to** have a machine learning algorithm that detects sexually abusive behaviour

	so that i can generate sexually abusive behaviour detection rules
Acceptance criteria	1. The algorithm will use the user’s OSN data to be trained. 2. The algorithm should run in a predefined schedule. 3. The generated detection rules should be sophisticated enough to produce more accurate sexually abusive behaviour predictions.

Code number	BE_5
Title	Fake identity and activity detection machine learning algorithm
Description	**As the data analytics software stack** **I want to** have a machine learning algorithm that detects fake identities and activity **so that** i can generate fake identity and activity detection rules
Acceptance criteria	1. The algorithm will use the user’s OSN data to be trained. 2. The algorithm should run in a predefined schedule. 3. The generated detection rules should be sophisticated enough to produce more accurate fake identity and activity predictions.

Code number	BE_6
Title	False information detection machine learning algorithm
Description	**As the data analytics software stack** **I want to** have a machine learning algorithm that detects false information **so that** i can generate false information detection rules
Acceptance criteria	1. The algorithm will use the user’s OSN data to be trained. 2. The algorithm should run in a predefined schedule. 3. The generated detection rules should be sophisticated enough to produce more accurate false information predictions.

Code number	BE_7
Title	Aggressive or distressed behaviour detection machine learning algorithm
Description	**As the data analytics software stack** **I want to** have a machine learning algorithm that detects hate speech **so that** i can generate aggressive or distressed behaviour detection rules
Acceptance criteria	1. The algorithm will use the user’s OSN data to be trained. 2. The algorithm should run in a predefined schedule. 3. The generated detection rules should be sophisticated enough to produce more accurate aggressive or distressed behaviour predictions.

7.4.2. Operational Requirements

#	Operational Requirements	Rational / Comments	Mandatory (YES/NO)
O4.1	The back-end should generate cyber bullying detection rules.	-	YES
O4.2	The back-end should generate sexual cyber	-	YES

	grooming detection rules.		
O4.3	The back-end should generate distressed behaviour detection rules.	-	YES
O4.4	The back-end should generate aggressive behaviour detection rules.	-	YES
O4.5	The back-end should generate fake identity detection rules.	-	YES
O4.6	The back-end should generate false information detection rules.	-	YES
O4.7	The back-end should receive and store suspicious traffic from the web-proxy.	-	YES
O4.8	The machine learning algorithms should periodically run to generate new more sophisticated detection rules.	-	YES
O4.9	The machine learning algorithms running in the back-end should be able to recover from errors.	-	YES

7.4.3. Security and Privacy Requirements

#	**Security & Privacy Requirements**	**Rational / Comments**	**Mandatory (YES/NO)**
S&P4.1	All communication between other modules must occur over a secure channel.	-	YES
S&P4.2	All the stored users’ OSN donated data should be securely stored.	-	YES
S&P4.3	All the stored users’ OSN donated data should not leave the back-end.	-	YES
S&P4.4	The Back-end should be security-hardened so that it is very difficult to compromise.	-	YES
S&P4.5	The user should be able to delete all collected and stored information about him from the platform.	-	YES
S&P4.6	All the users’ OSN data should be encrypted before stored.	-	YES

8. Conclusion

This deliverable is fundamental to the subsequent stages of ENCASE for several reasons: a) it provides a survey of existing security and privacy enhancing web-based tools related to ENCASE and a survey of the research state-of-the-art; b) it provides a revision of the usage scenarios provided in D2.1 which are expressing the functionality of the ENCASE platform that is to be implemented; c) provides a wide range of user stories that highlight the functionalities of the ENCASE platform and the offered security and privacy enhancing solutions that it offers; d) it provides the description of the reference architecture of the entire ENCASE platform, including a breakdown into different functional components connected by clearly identified interfaces; and (e) it defines the technical requirements of the platform derived from all the described user stories along with the reference architecture that aligns with them and will be the basis for the ENCASE implementation outlining the work to be done in each of the technical Work Packages.

9. References

1. C. Williams, et al., Perils and Possibilities: Growing up online. UNICEF. https://www.unicef.org/endviolence/endviolenceonline/files/UNICEF_Growing-up-online.pdf
2. Wright E.R. & Lawson A.H. (2004). Computer-Mediated Communication and Student Learning in Large Introductory Sociology Courses. Paper presented at the Annual Meeting of the American Sociological Association, Hilton San Francisco&Renaissance Parc 55 Hotel, San Francisco, CA. Available at: http://citation.allacademic.com/meta/p_mla_apa_research_citation/1/0/8/9/6/pages108968/p108968-1.php
3. Green H. & Hannon C. (2007). TheirSpace: Education for a Digital Generation. Demos, London. Available at: http://dera.ioe.ac.uk/23215/1/Their%20space%20-%20web.pdf
4. Wolak J., Finkelhor D., Mitchell K.J. & Ybarra M.L. (2008) Online ‘predators’ and their victims: myths, realities and implications for prevention and treatment. American Psychologist 63, 111–128.
5. Dooley, J J, D Cross, L Hearn and R Treyvaud (2011). The Protection of Children Online,OECD. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=dsti/iccp/reg(2010)5/final&doclanguage=en
6. Protect, understand and manage your kids’ internet activity with Qustodio. https://www.qustodio.com/en/
7. SocialShield: Avira Social Network Protection for your child. http://www.thewindowsclub.com/socialshield-review
8. Know which sites to trust. https://www.mywot.com/
9. Computer & Mobile monitoring software. http://www.webwatcher.com/?refID=lnkshr&siteID=Cty0dj6o3sg-GHtU.M9eT5Zlm7qQ5Ms1ig
10. Web Security Service. http://www.cloudacl.com/webfilter/
11. A. C. Squicciarini, J. Dupont, R. Chen, Online Abusive Analytics through Visualization
12. The Parental control for Firefox. https://addons.mozilla.org/en-US/firefox/addon/foxfilter/
13. Parental Controls & and Web Filter. https://chrome.google.com/webstore/detail/parental-controls-web-fil/dpfbddcgbimoafpgmbbjiliegkfcjkmn
14. MetaCert Security API. https://metacert.com/
15. eSafely protects you where your Web filter doesn’t. http://www.esafely.com/
16. Nudity detection with JavaScript and HTMLCanvas. https://github.com/pa7/nude.js
17. ReThink. http://www.rethinkwords.com/
18. PureSight Online child safety. http://puresight.com/puresight-prevents-cyberbullying.html
19. MM Guardian Parental Control. https://play.google.com/store/apps/details?id=com.mmguardian.childapp
20. Funamo Parental Control. https://play.google.com/store/apps/details?id=funamo.funamo
21. Kids Place - Parental Control. https://play.google.com/store/apps/details?id=com.kiddoware.kidsplace
22. AppLock. https://play.google.com/store/apps/details?id=com.domobile.applock
23. Screen Time Parental Control. https://play.google.com/store/apps/details?id=com.screentime.rc&hl=en_GB
24. Net Nanny. http://purchmarketplace.com/pc-software-net-nanny-7-0-download-only-1/?&ICID=ttr-cid|544|pid|49840|pos|
25. Safe Eyes - Parental control software. http://www.internetsafety.com/safe-eyes-parental-control-software.php
26. Elite Keylogger. https://www.elitekeyloggers.com/elite-keystroke-recorder-info

27. Kidlogger. https://kidlogger.net/
28. Spyrix Personal Monitor, Parental & Employees Monitoring Software. www.spyrix.com
29. Zoodles Kid Mode. https://www.zoodles.com
30. Crook, C., & Harrison, C. (2008). Web 2.0 technologies for learning at key stages 3 and 4: summary report.
31. Naidoo, T., Kritzinger, E., & Loock, M. (2013, June). Cyber Safety Education: Towards a Cyber-Safety Awareness Framework for Primary Schools. In International Conference on e-Learning (p. 272). Academic Conferences International Limited.
32. Sharples, M., Graber, R., Harrison, C., & Logan, K. (2009). E-safety and Web 2.0 for children aged 11–16. Journal of Computer Assisted Learning, 25(1), 70-84.
33. Searson, M., Hancock, M., Soheil, N., & Shepherd, G. (2015). Digital citizenship within global contexts. Education and Information Technologies, 20(4), 729-741.
34. Waters, J. K. (2011). Social Networking: Keeping It Clean. The Journal, 38(1), 52.
35. Orech, J. (2012). How it’s done: Incorporating digital citizenship into your everyday curriculum. Tech and Learning, 33(1), 16-18.
36. Ramnath, S. (2015). How schools can keep students safe, and on Facebook. eSchool News, 18(4), 16.
37. Campbell-Wright, K. (2013). E-safety. NIACE.
38. Sharples, M., Graber, R., Harrison, C., & Logan, K. (2009). E-safety and Web 2.0 for children aged 11–16. *Journal of Computer Assisted Learning*, *25*(1), 70-84.
39. Wespieser, K. (2015). Young People and E-Safety: The Results of the 2015 London Grid for Learning E-Safety Survey. National Foundation for Educational Research.
40. Lorenz, B., Kikkas, K., & Laanpere, M. (2011, November). Social Networks, E-learning and Internet Safety: Analysing the Stories of Students. In Proceedings of the 10th European Conference on e-Learning ECEL-2011: 10th European Conference on e-Learning ECEL-2011 Brighton, UK (pp. 10-11).
41. Lorenz, B., Kikkas, K., & Laanpere, M. (2012). Comparing Children's E-Safety Strategies with Guidelines Offered by Adults. Electronic Journal of e-Learning, 10(3), 326-338.
42. Moreno, M. A., Egan, K. G., Bare, K., Young, H. N., & Cox, E. D. (2013). Internet safety education for youth: stakeholder perspectives. BMC public health, 13(1), 543.
43. Cranmer, S. (2013). Listening to excluded young people's experiences of e-safety and risk. Learning, Media and Technology, 38(1), 72-85.
44. A SAFER DIGITAL WORLD.
45. Sharples, M., Graber, R., Harrison, C., & Logan, K. (2008). E-safety and Web 2.0: Web 2.0 technologies for learning at Key Stages 3 and 4.
46. Guidelines for children on child online protection. http://www.itu.int/en/cop/Pages/guidelines.aspx
47. Ofcom report on internet safety measures: Strategies of parental protection for children online. http://stakeholders.ofcom.org.uk/binaries/internet/internet-safetymeasures.pdf
48. Ofcom report on internet safety measures: Strategies of parental protection for children online. http://stakeholders.ofcom.org.uk/binaries/internet/fourth_internet_safety_report.pdf
49. M. Robinson. (2015, March) Koreas internet addiction crisis is getting worse, as teens spend up to 88 hours a week gaming. Business Insider. http://www.businessinsider.com/south-korea-online-gaming-addiction-rehab-centers-2015-3
50. A. Lenhart. (2015, April) Teens, social media & technology overview 2015. Pew Recearch Center: Internet, Science & Tech. http://www.businessinsider.com/south-korea-online-gaming-addiction-rehab-centers-2015-3
51. S. Livingstone, L. Haddon, J. Vincent, G. Mascheroni, and K. Olafsson. (2014) Net children go mobile: The uk report. London: London School of Economics and Political Science.

https://www.lse.ac.uk/media@lse/research/EUKidsOnline/EUn%20Kidsn%20III/Reports/NCGMUKReportfinal.pdf

52. S. Livingstone, K. Cagiltay, and K. Olafsson, “Eu kids online ii dataset: A cross-national study of children’s use of the internet and its associated opportunities and risks,” British Journal of Educational Technology, vol. 46, pp. 988–992, August 2015.
53. T. Woda. (2015) Digital parenting: Understanding the risk of snapchat. uknowkids.com. http://resources.uknowkids.com/blog/digital-parenting-understanding-the-risk-of-snapchat
54. ACMA. (2009, July) Click and connect: Young australians use of online social media. Australian Communications and Media Authority. http://www.acma.gov.au/webwr/aba/about/recruitment/click_and_connect-01_qualitativen_report.pdf
55. Developments in internet filtering technologies and other measures for promoting online safety. Australian Communications and Media Authority. http://www.acma.gov.au/webwr/nassets/main/lib310554/developments_in_internet_filters_2ndreport.pdf
56. P. Hindley, J. Hurn, and S. Stringer, “Ward: Child protection concerns,” in Psychiatry: Breaking the ICE - Introductions, Common Tasks and Emergencies for Trainee. John Wiley & Sons, 2016
57. R. Thompson, “Social support and child protection: Lessons learned and learning,” Child Abuse & Neglect, vol. 41, pp. 19–29, 2015
58. The Protection of Children Online: Report on risks faced by children online and policies to protect them. https://www.oecd.org/sti/ieconomy/childrenonline with cover.pdf
59. S. Livingstone, “A rationale for positive online content for children,” Communication Research Trends, vol. 28, pp. 12–16, 2008.
60. J. Dooley, D. Cross, L. Hearn, and R. Treyvaud. (2009) Review of existing australian and international cyber-safety research. Child Health Promotion Research Centre, Edith Cowan University, Perth. http://unpan1.un.org/intradoc/groups/public/documents/apcity/unpan046312.pdf
61. A. Marwick, D. Murgia-Diaz, and J. Palfrey, “Youth, privacy and reputations,” Berkman Center Research, Tech. Rep. 2010-5, 2010. http://papers.ssrn.com/sol3/papers.cfm?abstract id=1588163
62. R. Slonje, P. Smith, and A. Frisn, “The nature of cyberbullying, and strategies for prevention,” Computers in Human Behaviort, vol. 29, pp.26–32, 2013.
63. Child Online Protection - Statistical Framework and Indicators. https://www.itu.int/dms pub/itu-d/opb/ind/D-IND-COP.01-11-2010-PDF-E.pdf
64. M. Ybarra and K. J. Mitchel, “How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs,” Pediatrics, vol. 121, no. 2, pp. e350–e357, Feb 2008.
65. Implementing the Childrens Online Privacy Protection Act: A Report to Congress. http://www.ftc.gov/reports/coppa/07COPPA Report to Congress.pdf
66. E. Bartoli, “Children’s data protection vs. marketing companies,” International Review of Law, Computers & Technology, vol. 23, no. 1–2, pp.35–45, July 2009.
67. Guidelines for Policy Makers of Child Online Protection, 2009. http://www.itu.int/en/cop/Documents/guidelines-policy%20makerse.
68. Dirty sex dictionary. http://www.cltampa.com/home/article/20750335/dirty-sex-dictionary
69. Prashant Ravi, *Detecting Insults in Social Commentary*, Overleaf Publication, https://www.overleaf.com/articles/detecting-insults-in-social-commentary/gkvrrwryjxhr#.V8fpVph95PY
70. Ben Ismail, Mohamed Maher and Bchir, Ouiem, "*Insult Detection in Social Network Comments Using Possibilistic Based Fusion Approach*", Computer and Information Science, "Springer International Publishing" pp. 15-25, 2015, doi="10.1007/978-3-319-10509-3_2"

71. Kaggle dataset link: https://www.kaggle.com/c/detecting-insults-in-social-commentary/data
72. Aneesha Tool. https://github.com/aneesha/cbd
73. aUDA Foundation. https://www.audafoundation.org.au/grant-recipients/2013-grant-recipients/university-of-technology-sydney/
74. N. F. Johnson and S. Jajodia, "Exploring steganography: Seeing the unseen," in IEEE Computer, vol. 31, no. 2, pp. 26-34, Feb. 1998.
75. A. Cheddad, J. Condell, K. Curran, P. Mc Kevitt, “Digital image steganography: Survey and analysis of current methods”, Elsevier Journal of Signal Processing, Volume 90, Issue 3, March 2010, Pages 727-752.
76. R. Das and T. Tuithung, "A novel steganography method for image based on Huffman Encoding," 2012 3rd National Conference on Emerging Trends and Applications in Computer Science, Shillong, 2012, pp. 14-18.
77. A. Kiayias, Y. Tsiounis, M. Yung, "Group Encryption", IACR, January 2007.
78. A. Sahai, B. Waters, "Fuzzy Identity-Based Encryption", IARC, 2004.
79. A. Shamir. Identity-based cryptosystems and signature schemes. In Proceedings of CRYPTO 84 on Advances in cryptology, pages 47–53. Springer-Verlag New York, Inc., 1985.
80. Oracle WebLogic Server Proxy PlugIn. https://docs.oracle.com/middleware/1221/webtier/develop-plugin/overview.htm#PLGWL4336
81. N. Nahata, Understanding the use of "WebLogic PlugIn Enabled", http://www.ateam-oracle.com/wls-plugin-enabled/
82. Hola!VPN. http://hola.org/
83. Chromium architecture. https://www.chromium.org/developers/design-documents/plugin-architecture
84. The Evolution to the Next Generation Firewall. http://stonesoft-security.co.uk/pdf/whitepapers/evolution_to_ngfw_whitepaper.pdf
85. Firewall. https://en.wikipedia.org/wiki/Firewall_(computing)
86. Intrusion Detection & Response - Leveraging Next Generation Firewall Technology. https://www.sans.org/reading-room/whitepapers/firewalls/intrusion-detection-response-leveraging-generation-firewall-technology-33053
87. Understanding the Next Generation Firewall and its Architecture. https://www.alliedtelesis.com/sites/default/files/aw_how_to_understanding_ngfw_architecture.pdf
88. Palo Alto Networks, "Next Generation Firewall". https://www.paloaltonetworks.com/products/secure-the-network/next-generation-firewall
89. Network Intelligence Consulting, "Next Generation Firewall (NGFW)". https://www.niiconsulting.com/solutions/next-generation-firewalls.html
90. Bhat, S. Y. (2013). Community-based features for identifying spammers in online social networks. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks (pp. 100-107). Niagara, Ontario, Canada: ACM.
91. Castillo, C. A. (2011). Information Credibility on Twitter. (p. Proceedings of the 20th International Conference on World Wide Web). Hyberabad, India: ACM.
92. Cha, M. A. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM), (pp. 10-17).
93. Yang, C. A. (2012). Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. Proceedings of the 21st International Conference on World Wide Web (pp. 71-80). ACM.
94. Grier, C. A. (2010). @spam: The underground on 140 characters or less. Proceedings 17th ACM Conference on Computer & Communications Security (CCS), (pp. 27-37).

95. Xianghan, Z. a. (2015). Detecting spammers on social networks. Neurocomputing, (pp. 27-34).
96. Heartfield, R. a. (2015). A taxonomy of attacks and a survey of defense mechanisms for semantic social engineerings attacks. ACM Computing Surveys (pp. 37-39). ACM.
97. Thomas, K. a. (2011). Suspended accounts in retrospect: An analysis of twitter spam. Proceedings ACM SIGCOMM Conference on Internet Measurement Conference (IMC), (pp. 243-258).
98. Lambda Architecture. http://lambda-architecture.net
99. Apache Kafka. https://kafka.apache.org/intro
100. Apache Spark Streaming. http://spark.apache.org/streaming/
101. Hadoop Distributed File system (HDFS). https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
102. Deep Learning with Apache Spark and TensorFlow. https://databricks.com/blog/2016/01/25/deep-learning-with-apache-spark-and-tensorflow.html
103. ElasticSearch. https://www.elastic.co/products/elasticsearch
104. Suggestersin ElasticSearch. https://www.elastic.co/blog/found-fuzzy-search
105. TensorFlow. https://www.tensorflow.org/