PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Mauro Garofalo

XXIX Cycle

Training and Research Activities Report – Second Year

Tutor: Giorgio Ventre

Information

Il sottoscritto Mauro Garofalo, laureato in Ingegneria Informatica presso l'Università di Napoli Federico II, iscritto al secondo anno del Dottorato in Information Technology and Electrical Engineering – ITEE- (XXIX ciclo) sotto la supervisione del Prof. Giorgio Ventre, dichiara di aver partecipato alle attività di formazione e svolto le attività di ricerca esposte nel seguito.

Study and Training activities

Courses

- Project Management per la Ricerca, Guido Capaldo, marzo 2015
- Models, methods and software for Optimization, Antonio Sforza, maggio 2015
- Designing and writing scientific manuscripts for publication in English language scholarly journals and related topics, Parker, giugno 2015
- Cambridge English: Cambridge English: First, G. Thomas, luglio 2015
- Network Security, Simon Pietro Romano, gennaio 2016
- Semantic Web, Piero Bonatti, in corso

Seminars

- Agent with Truly Perfect Recall, Prof Nils Bulling, aprile 2015
- Regularization of two-fold bifurcations in planar piecewise-smooth systems, John Hogan, Giugno 2015
- Memory technologies for Android based systems, Simon Pietro Romano, novembre 2015
- Model Based and Pattern Based GUI Testing, Fasolino, novembre 2015
- Security Operations in una Telco, esperienze e riflessioni dal campo, Fabio Zamparelli, dicembre 2015
- Test and Diagnosis of Integrated Circuits, novembre 2015

Research activity

Obiettivo della mia attività di ricerca in questo secondo anno di dottorato è stato definire e sviluppare metodologie per il rilevamento di anomalie su reti di calcolatori. Con il termine anomalie (descritto spesso con altri nomi come: outliers, eccezioni, aberrazioni, etc.), ci si riferisce in generale a schemi di comportamento che differiscono da ciò che normalmente ci si aspetterebbe. Nelle reti di calcolatori in particolare un'anomalia può essere causata sia da attacchi informatici che da attività non malevole. Nella mia attività di ricerca mi sono interessato ad entrambi i tipi di anomalie. Come caso di studio riguardo le anomalie da attività non malevole, ho valutato le performance di tre dei top video hosting provider, misurate dal

punto di vista degli utenti e da diverse parti del mondo, per rilevare eventuali trattamenti preferenziali del traffico di rete. Ho effettuato delle misurazioni attive simulando l'attività di utenti sparsi per tutto il mondo intenti a guardare un video su ognuno dei provider valutati. Questo tipo di misurazione serve ad identificare eventuali anomalie dovute a politiche di gestione preferenziale del traffico da parte degli operatori della rete (ISP, fornitori dei servizi, governi). Questo tipo di anomalie, se presenti, possono essere dovute a violazioni della cosiddetta neutralità della rete.

Per l'identificazione delle anomalie dovute ad attività malevole, tematica su cui si è concentrata maggiormente la mia attività di ricerca, ho svolto un tipo di misurazione passiva, analizzando le informazioni di livello flusso del traffico di rete provenienti da opportuni punti di osservazione. Nel seguito sono riportati più nel dettaglio tutti gli aspetti di questa tematica.

Title

"Flow-based Network Anomaly Detection on Tbps Traffic"

Study

Anche se in letteratura sono state proposte diverse soluzioni al riguardo, l'aumento delle velocità di trasmissione, la mole di dati prodotta dal numero crescente di dispositivi collegati alla rete, e la concomitante crescita delle attività malevole, rendono ad oggi il rilevamento di anomalie nel traffico di reti Tbps sempre più impegnativo. Dal punto di vista dei dati da analizzare esistono diversi approcci. Alcuni basati sull'analisi di informazioni a basso livello, come quelle dei pacchetti che transitano sulla rete, altri, all'estremo opposto, utilizzano informazioni di alto livello come quelle contenute nei file di log dei servizi attivi sulla rete (come ad esempio un web server).

Il nostro approccio, collocandosi nel mezzo, analizza le informazioni a livello flusso, fornendo una visione di insieme del traffico di rete (ad esempio il numero di pacchetti scambiati, la quantità di byte e la durata della comunicazione).

Dato che il traffico di rete possiede tutte e 3 le V che definiscono i BIG DATA, volume, varietà e velocità, con la nostra metodologia ci proponiamo di sfruttare i framework di Big Data Analysis come strumento per l'esecuzione degli algoritmi di anomaly detection.

Il mio lavoro in questo secondo anno si è quindi concentrato sui seguenti problemi:

1. Studiare lo stato dell'arte delle tecniche di anomaly detection.
2. Studiare lo stato dell'arte dei framework di Big Data Analysis.
3. Trovare una ground truth con la quale poter validare gli algoritmi di rilevamento.
4. Definire un'architettura per l'analisi del traffico di rete a livello flusso.

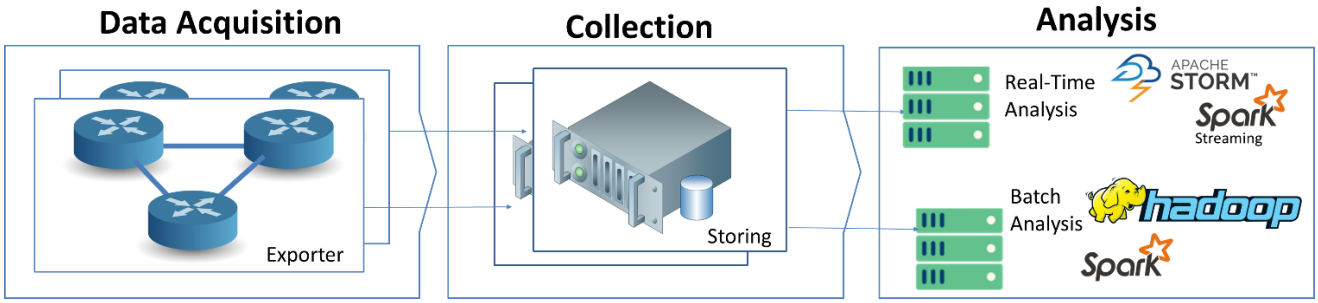

Figura 1 - Architettura di anomaly detection proposta

Per validare la nostra metodologia, una versione embrionale dell'architettura, rappresentato nella figura precedente, è stata sviluppata. Questa è composta da tre componenti:

1. Data Acquisition, a partire dai pacchetti osservati crea ed invia al componente Collection i record dei flussi.
2. Collection, ha il compito di recuperare i flussi, memorizzarli e trasformarli in un formato compatibile con i framework di analisi.
3. Analysis, analizza i flussi, siano essi memorizzati o in arrivo in streaming, utilizzando gli algoritmi di rilevamento.

Come dati di ingresso, anziché utilizzare gli aggregati provenienti da dispositivi (router o switch) costituenti una rete reale, abbiamo utilizzato delle tracce di traffico di rete messe a disposizione pubblicamente dal progetto MAWI. Il progetto MAWI, oltre alle tracce di traffico, mette a disposizione un archivio relativo alle anomalie presenti nelle stesse. Tale archivio è stato utilizzato come ground truth per la validazione degli algoritmi.

I primi due componenti dell'architettura, Data Acquisition e Collection, in questa prima fase sono stati simulati via software. Il primo utilizzando dei software che trasformano le tracce di traffico in record di flussi. Il secondo utilizzando il file system distribuito HDFS, messo a disposizione da Hadoop. Per l'ultimo componente, abbiamo creato un piccolo cluster di 3 nodi virtuali, su cui è stato installato Apache Spark come framework di Big Data Analysis.

l proseguo del lavoro nel breve periodo mira a rendere l'architettura completamente operativa sviluppando i componenti che attualmente sono simulati via software, investigare e confrontare le prestazioni di altri algoritmi di rilevamento ed altri framework di Big Data

Analysis ed infine testare la scalabilità degli stessi utilizzando soluzioni commerciali di PaaS e SaaS come ad esempio AWS.

Products

Publication

List those in preparation

Title: A survey on Big Data and Networking

Title: An architecture for flow-based anomaly detection based on Big Data analysis tools

Title: Internet streaming and network neutrality: comparing the performance of video hosting services (extended)

Conferences and Seminars

Conference Paper

Conference: ICISSP 2016 - The International Conference on Information Systems Security and Privacy. Roma 19-21 febbraio 2016.

Title: Internet streaming and network neutrality: comparing the performance of video hosting services.

Authors: Alessio Botta, Aniello Avallone, Mauro Garofalo and Giorgio Ventre

Abstract

Network neutrality is a hot topic since a few years and involves different aspects of interest (e.g. economic, regulatory and privacy) for a wide range of stakeholders, including policy makers, researchers, economists, and service providers.

When referring to video streaming, a killer web service of the Internet, much has been discussed regarding if and how video providers violate or may violate neutrality principles, in order to give users a ``better'' service compared to other services or to other providers.

In this paper we provide a contribution to this discussion analyzing the performance of three main video hosting providers (i.e. YouTube, Vimeo, and Dailymotion) from an user viewpoint.

We measure the throughput and RTT experienced by users watching real videos of different popularity, at different day hours and at several locations from around the world.

We uncover the performance differences of these providers as a function of the different variables under control and move a step forward to understand what causes such differences.

Our results allow to understand what are the real performance users currently get from these providers and if the performance differences observed can be due or considered as a violation of network neutrality principles, providing a ground for people interested in legal and regulatory issues of web applications and services.

CS Summary

	Credits year 1								Credits year 2								Credits year 3									
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total	Check
Modules	20		3		3		6	**12**	20	3	7	6			6	**22**	4							**0**	**34**	**30-70**
Seminars	8		0,6		2	0,9	2,2	**5,7**	6	0,2	0,2			4,6		**5**	5							**0**	**11**	**10-30**
Research	32	10	6,4	10	5	9,1	1,8	**42**	34	6,8	2,8	4	10	5,4	4	**33**	51							**0**	**75**	**80-140**
	60	10	10	10	10	10	10	**60**	60	10	10	10	10	10	10	**60**	60	0	0	0	0	0	0	**0**	**120**	**180**