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 – Third 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

Summer Schools

Summer School on Computer Security & Privacy - Building Trust in the Information Age, Pula (CA), Italy 5-9 Settembre 2016

Seminars

- Internet: la dimensione immateriale dell'esistenza, Maggio 2016
- MINIX3: A Reliable and Secure Operating System, Novembre 2016
- L'innovazione nel mercato IT, Dicembre 2016

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	15		3		3		6	**12**	20	3	7	6			6	**22**	2							**0**	**34**	30-70
Seminars	6		0,6		2	0,9	2,2	**5,7**	6	0,2	0,2			4,6		**5**	2		0,4		3	1		**4,4**	**15**	10-30
Research	39	10	6,4	10	5	9,1	1,8	**42**	34	6,8	2,8	4	10	5	4	**33**	56	10	9,6	10	7	9	10	**56**	**131**	80-140
	60	10	10	10	10	10	10	**60**	60	10	10	10	10	9,6	10	**60**	60	10	10	10	10	10	10	**60**	**180**	180

Research activity

Al giorno d'oggi, l'utilizzo della rete Internet è diventato indispensabile e pervasivo nelle nostre vite. Infatti, sempre più servizi sono disponibili su Internet, e per alcuni di essi, l'accesso alla rete è diventato indispensabile per il normale funzionamento. Usiamo e-mail e social media per comunicare, compriamo cose su siti di e-commerce e utilizziamo i servizi di e-banking per gestire i nostri soldi.

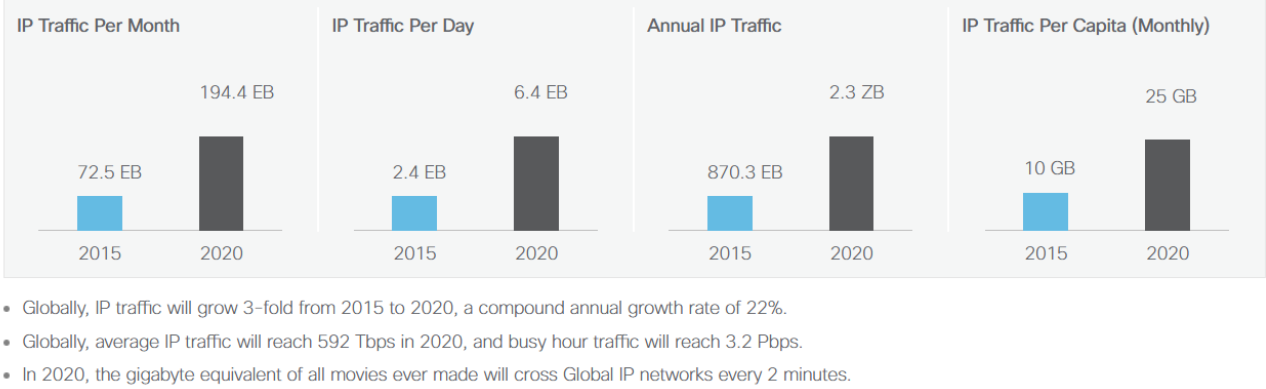
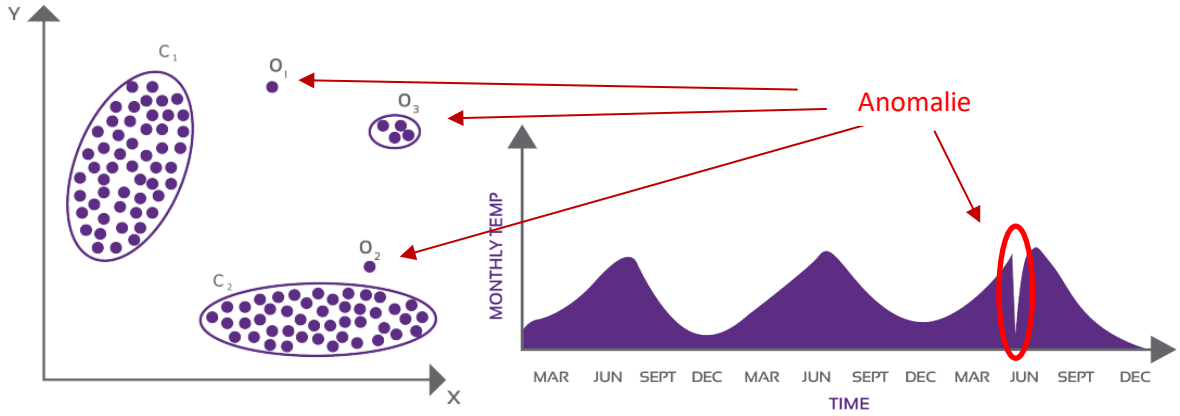


Figura 1 Cisco Visual Networking Index - Forecast Highlights 2015-2020

Secondo Cisco [1], la crescita di dispositivi collegati alla rete genera e genererà nei prossimi anni una quantità di traffico di rete mai vista prima. I dati di rete aumentano in tre direzioni: in complessità, volume e velocità. Riuscire comprendere ciò che sta realmente accadendo all'interno della rete sta diventando un compito sempre più complesso a causa della crescente lista di applicazioni che modellano traffico Internet oggi. Di conseguenza, il monitoraggio e l'analisi del traffico sono diventati fondamentali per diverse attività che vanno dal rilevamento delle intrusioni all'ingegnerizzazione del traffico, alla pianificazione delle capacità e classificazione del traffico.

In particolare, a causa della enorme aumento di attacchi informatici in tutto il mondo, la ricerca si è concentrata sulla rilevazione anomalie. Tuttavia, in letteratura sono presenti molte proposte che spesso non rispettano alcuni requisiti del mondo reale, come quando devono essere dispiegate su reti di tipo backbone.

Obiettivo della mia attività di ricerca in questo terzo anno di dottorato è stato quello di affrontare alcuni di questi problemi definendo e sviluppando metodologie per il rilevamento di anomalie su questo tipo di reti.



Con il termine anomalia ci si riferisce in generale ad un comportamento che differisce da ciò che normalmente ci si aspetterebbe. Nelle reti di calcolatori un'anomalia può essere causata sia da un'azione malevola (e.g. un attacco informatico) che da una configurazione sbagliata (o in generale da un'azione non malevola). Nella mia attività di ricerca mi sono interessato ad entrambi i tipi di anomalie. Per quanto riguarda lo studio delle anomalie dovute ad attività non malevole, si sono studiati possibili trattamenti preferenziali del traffico di rete (ovvero violazione della network neutrality). Per questo motivo ho valutato le performance di tre dei top video hosting provider (Dailymotion, Vimeo e YouTube), misurate dal punto di vista degli utenti e da diverse parti del mondo. Queste misurazioni di tipo attivo sono state effettuate 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 [2].
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 del traffico di rete da opportuni punti di osservazione. L'analisi del traffico è stata effettuata analizzando le informazioni del traffico a livello flusso. Nel seguito sono riportati più nel dettaglio tutti gli aspetti di questa tematica.

Title

"Flow-based Network Anomaly Detection on High-Speed Network"

Study

Anche se in letteratura esistono 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 rete 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. Un flusso TCP/IP, sulla cui osservazione si basano gli approcci flow-based, è caratterizzato da una quintupla composta da: indirizzo sorgente, indirizzo destinazione, porta sorgente, porta destinazione, protocollo utilizzato. Gli approcci flow-based prevedono inoltre l'analisi di ulteriori informazioni come ad esempio la durata del flusso o il numero di byte e pacchetti trasmessi [3].
Dato che il traffico di rete possiede tutte e 3 le V che definiscono i BIG DATA, ovvero 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 terzo anno si è quindi concentrato sui seguenti problemi:

1. Trovare una ground truth con la quale poter validare gli algoritmi di rilevamento.
2. Definire un'architettura per l'analisi del traffico di rete a livello flusso.
3. Realizzare tale architettura sfruttando i succitati framework di Big Data Analysis per la fase di esecuzione degli algoritmi di rilevamento.
4. Validare algoritmi e architettura.

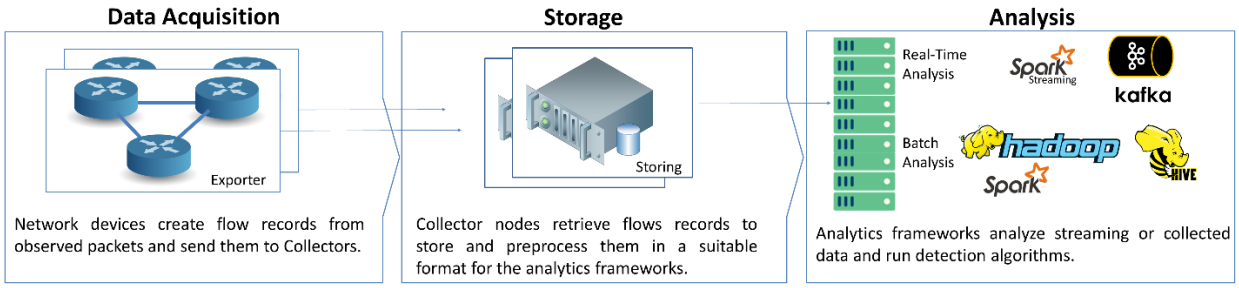


Figura 2 - Architettura di anomaly detection proposta

Per validare la nostra metodologia e la scalabilità dell'architettura rappresentata nella figura precedente sono stati sviluppati due prototipi. Il primo, utilizzando una singola macchina virtuale, e l'altro sfruttando la potenza di calcolo di un'infrastruttura cloud (ovvero Amazon Elastic MapReduce).

Il modello della architettura proposta è composto da tre livelli:

1. Data Acquisition, ha il compito di trasformare i pacchetti osservati in record di flussi ed effettuare le operazioni di pre-processing
2. Storage, ha il compito di recuperare i flussi creati nel livello precedente, trasformarli in un formato compatibile con i framework di analisi e fornisce servizi di memorizzazione e management dei dati.
3. Analysis, analizza i flussi, siano essi memorizzati o in arrivo in streaming, utilizzando gli algoritmi di rilevamento.

I due componenti dell'architettura, Data Acquisition e Storage, sono stati realizzati 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, nel caso di un'analisi di tipo batch, ed il message broker Apache Kafka, nel caso di analisi streaming. Per l'ultimo componente, abbiamo utilizzato il framework di Big Data Analysis Apache Spark.

Per la fase di anomaly detection, abbiamo implementato un algoritmo a soglia che monitora, per ogni indirizzo IP della rete monitorata, il rapporto tra numero di flussi creati e numero di flussi ricevuti. Per validare tale l'algoritmo abbiamo utilizzato come dati di ingresso tracce di traffico messe a disposizione dal progetto MAWI. La scelta di utilizzare queste tracce ha due motivazioni principali. Primo dal 2005 ad oggi, il progetto MAWI mette a un archivio di tracce disponibili pubblicamente relative a traffico reale della durata di 15 minuti catturato giornalmente. Secondo, il progetto MAWILab fornisce un archivio di anomalie di traffico rilevate nelle tracce dell'archivio MAWI. Le anomalie vengono rilevate utilizzando una combinazione di 4 algoritmi di anomaly detection indipendenti.

Per validare l'algoritmo abbiamo usato tutte le tracce del mese di Novembre del 2014, uno dei mesi dell'archivio MAWILab con più completo in termini di tracce analizzate (28 giorni su 31). Utilizzando MAWILab come ground truth i risultati sono stati i seguenti:

Day	1	2	4	5	7	8	9	10	11	12	13	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
TP	106	107	47	68	56	60	63	83	58	30	53	60	73	64	37	78	72	71	84	87	73	41	47	59	33	50	54	34
FP	11	22	9	28	45	24	28	32	19	5	13	33	35	41	11	10	34	246	58	47	49	206	203	237	54	53	22	20
PPV	0,91	0,83	0,84	0,71	0,55	0,71	0,69	0,72	0,75	0,86	0,80	0,65	0,68	0,61	0,77	0,89	0,68	0,22	0,59	0,65	0,60	0,17	0,19	0,20	0,38	0,49	0,71	0,63
TP*	111	122	55	94	93	76	87	109	72	34	64	84	99	96	48	86	100	300	133	123	109	246	250	275	76	89	67	46
FP*	6	7	1	2	8	8	4	6	5	1	2	9	6	9	0	2	6	17	9	11	13	1	0	21	11	14	9	8
PPV*	0,95	0,95	0,98	0,98	0,92	0,90	0,96	0,95	0,94	0,97	0,97	0,90	0,94	0,91	1,00	0,98	0,94	0,95	0,94	0,92	0,89	0,996	1,00	0,93	0,87	0,86	0,88	0,85

Nella tabella sono indicati per ogni giorno del mese i True Positive (TP), ovvero il numero di eventi per cui il nostro algoritmo segnala un alert su un flusso che anche MAWILab segnala come anomalo, i False Positive (FP), ovvero il numero di eventi in cui il nostro algoritmo segnala come anomalo un flusso che per MAWILab è normale, e il Positive Predictive Value (PPV, detto anche Precision), metrica che si calcola come: $PPV = \frac{TP}{TP+FP}$

In rosso si possono notare i casi in cui, utilizzando i soli risultati MAWILab come confronto, si riscontrano diverse centinaia di falsi positivi.

Siccome l'etichettatura di MAWILab è un processo di tipo euristico, non è possibile considerarla una ground truth ma piuttosto un gold standard. Per questo motivo abbiamo effettuato un'analisi approfondita sugli eventi segnalati come falsi positivi. Nella seconda parte della tabella sono presenti i risultati di questa analisi. Come si può notare, per i casi peggiori in termini di PPV (quelli in rosso) il numero di FP scendono tanto da raggiungere valori di PPV superiori a 0,9.

Le sperimentazioni volte a misurare la scalabilità della nostra architettura sono ancora in corso.

Bibliography

[1] Cisco (2016). Cisco VNI Forecast and Methodology, 2015-2020. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html. Last Accessed: 2017-02-17

[2] Botta, A., Avallone, A., Garofalo, M., and Ventre, G "Internet streaming and network neutrality: comparing the performance of video hosting services," ICISSP2016

[3] Hofstede, R.; Celeda, P.; Trammell, B.; Drago, I.; Sadre, R.; Sperotto, A.; Pras, A., "Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX," Communications Surveys & Tutorials, IEEE , vol.16, no.4, pp.2037,2064, Fourthquarter 2014

Products

Publication

List those in preparation

Title: Computer Network in the Big Data era.

Title: Big Data Architecture for flow-based anomaly detection

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

Conferences and Seminars

Conference Paper

Conference: AstroInfo16 – AstroInformatics, IAU Symposium No. 325, Sorrento (NA) 20-24 Ottobre 2016.

Title: Astrophysics in the Big Data era: Challenges, Methods, and Tools.

Authors: Mauro Garofalo, Alessio Botta, and Giorgio Ventre.

Abstract

Nowadays there is no field research which is not flooded with data. Among the sciences, astrophysics discoveries have always been driven by the analysis of massive amounts of data. The development of new and more sophisticated observation facilities, from the earth and into space, has led data more and more complex (Variety), and an exponential growth of both data Volume (i.e., in the order of petabytes), and Velocity in terms of production and transmission. Therefore, new and advanced processing solutions will be needed to process this huge amount of data. We investigate some of these solutions, from the machine learning models as well as tools and architectures for the analysis of Big Data that can be exploited in the astrophysics context.