



# **NEW BGP ROUTE LEAKS CLASSIFICATION AND DETECTION USING SUPERVISED MACHINE LEARNING TECHNIQUE**

By

**Salma Abdel Monem Abdel Motaleb Mohamed**

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
**MASTER OF SCIENCE**  
in  
**Computer Engineering**

FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2019

NEW BGP ROUTE LEAKS CLASSIFICATION AND DETECTION  
USING SUPERVISED MACHINE  
LEARNING TECHNIQUE

By

Salma Abdel Monem Abdel Motaleb Mohamed

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
MASTER OF SCIENCE in  
Computer Engineering

Under the Supervision of

Prof. Dr. Samir I. Shaheen

Professor of Computer  
Computer Engineering  
Faculty of Engineering, Cairo University

Dr. Ahmed Bashandy

Associate Professor of Computer  
Computer Engineering  
Faculty of Engineering, Some University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2019

NEW BGP ROUTE LEAKS CLASSIFICATION AND DETECTION  
USING SUPERVISED MACHINE  
LEARNING TECHNIQUE

By  
Salma Abdel Monem Abdel Motaleb Mohamed

A Thesis Submitted to the  
Faculty of Engineering at Cairo University  
in Partial Fulfillment of the  
Requirements for the Degree of  
MASTER OF SCIENCE  
in  
Computer Engineering

Approved by the Examining Committee:

---

Prof. Dr. Samir I. Shaheen	Thesis Main Advisor
----------------------------	---------------------

---

Dr. Amr G. Wassal	Internal Examiner
-------------------	-------------------

---

Prof. Dr. Mohamed Z. Abd El Mageed	External Examiner
------------------------------------	-------------------

FACULTY OF ENGINEERING, CAIRO UNIVERSITY  
GIZA, EGYPT  
2019

**Engineer's Name:** Salma Abdel Monem Abdel Motaleb  
Mohamed  
**Date of Birth:** 08/03/1994  
**Nationality:** Egyptian  
**E-mail:** salmacmpe@cu.edu.eg  
**Phone:** 01114302362  
**Address:** 33 El-Zhoor st, El Motamiz District,  
6 October city, Egypt.  
**Registration Date:** 01/10/2016  
**Awarding Date:** 29/7/2019  
**Degree:** Master of Science  
**Department:** Computer Engineering  
**Supervisors:** Prof. Samir I. Shaheen  
Dr. Ahmed Bashandy



**Examiners:** Prof.Dr. Samir I. Shaheen (Thesis Main Advisor)  
Dr. Amr G. Wassal (Internal Examiner)  
Dr. Mohamed Zaki Abdel Mageed (External Examiner)

**Title of the thesis:**

New BGP Route Leaks Classification and Detection Using Supervised Machine Learning Technique.

**Key Words:**

BGP; Route Leaks; Leaks Classification; Machine Learning.

**Summary:**

The route leaks problem is considered one of the unsolved problems in BGP through the previous 15 years. The confidentiality of autonomous systems (ASes) relationships and the lake of advertisement of route leaks incidents are the main two reasons behind this. This thesis solves the route leaks problem relaying on three steps: A new taxonomy for the route leaks types based on their effects to the BGP Traffic not to their ASes relationships is proposed, the first dataset for published real route leaks incidents through the previous years is collected and labeled as route leaks or normal traffic, and the first real-time detection system of route leaks problem using complex features extracted from BGP Update messages only and using Classification algorithms is proposed. The system achieves the best accuracy of 88% and 92% F1 Score, the whole system can detect route leaks upon receiving them in real-time as it runs in less than one second.

## **Disclaimer**

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name:

Date:

Signature:

## **Dedication**

This thesis is dedicated to my family for their unconditional support, help, and encouragement without whom I would never be the person who am I.

## **Acknowledgements**

I would like to thank Prof. Samir Shaheen and Dr. Ahmed Bashandy for their support, advice, and guidance throughout the thesis. I would like to thank Youssef Ghatas and Eman Hossam for their continuous encouragement, help and their valuable comments and ideas.

# Table of Contents

Disclaimer .....	xi
Dedication .....	xi
Acknowledgements .....	xi
Table of Contents .....	xiv
List of Tables .....	xi
List of Figures .....	xii
List of Symbols and Abbreviations.....	xiii
Abstract .....	xiv
1 Introduction .....	1
1.1 Motivation .....	1
1.2 Objectives .....	2
1.3 Achievements .....	2
1.4 Organization of the thesis.....	3
2 Literature review .....	4
2.1 Scientific Background .....	4
2.1.1 Introduction to Border Gateway Protocol.....	4
2.1.2 Autonomous systems relationships.....	4
2.1.3 Valley Free Rules.....	6
2.1.4 Route leaks definition .....	7
2.1.5 Route leaks analysis, measurements and verification.....	7
2.1.6 Route leaks effects .....	7
2.1.7 Scientific view of different types of treats for the route leaks problem.....	8
2.1.8 Machine learning.....	8
2.2 Probabilistic models .....	13
2.3 Related Work.....	14
2.3.1 AS inference.....	14
2.3.2 Control plane.....	15
2.3.3 Classification and understanding bgp misconfigurations .....	15
2.3.4 Simple Tier-1 ASes detection .....	16
2.3.5 Detect routing loops .....	16



2.3.6	Detecting large route leaks.....	16
2.3.7	Detecting route leaks using historical data and inference.....	17
2.3.8	Detecting route leaks by investigating local information .....	17
2.4	Summary or literature review .....	18
3	Route leaks .....	19
3.1	Route leaks classification according to AS relationships.....	19
3.2	Large Reported Route leak incidents .....	21
3.2.1	VolumeDrive 2014 incident [21] Figure 3.2.....	21
3.2.2	Telecom Malaysia 2015 incident [22] [23] Figure 3.3 .....	22
3.2.3	Dodo Network 2012 incident [24] Figure 3.4.....	23
3.2.4	Verizon 2014 incident [25] .....	24
3.2.5	Amazon 2015 incident [26] Figure 3.5 .....	24
3.2.6	Google 2015 incident [27] Figure 3.6.....	25
3.2.7	Belarusian prefix hijacking [28] Figure 3.7 .....	26
3.3	Route Leaks causes .....	28
4	Proposed taxonomy to the route leaks .....	29
4.1	The Four Types Taxonomy .....	29
4.1.1	Type (1) More specific route leaks about internal network and customers ....	30
4.1.2	Type (2) More specific route leaks about peers .....	30
4.1.3	Type (3) Leaked routes learned from transit provider to another provider ....	31
4.1.4	Type (4) Full table route leak.....	31
4.2	Statistical analysis of the large reported route leak incidents .....	31
4.2.1	Features for statistical analysis for large route leak incidents .....	32
4.2.2	Conclusion about the previous results: .....	35
5	System Model .....	36
5.1	System blocks diagram.....	37
5.2	System Inputs and outputs.....	38
5.3	Data Gathering, Cleaning and labeling .....	41
5.3.1	Data gathering .....	41
5.3.2	Data labeling .....	42
5.4	Features selection and extraction .....	42
5.4.1	Feature extraction Method 1 (MYSQL database).....	44

5.4.2	Feature extraction Method 2(Patricia tree)	46
5.5	Classification	50
5.5.1	Decision tree classifier (DT)	50
5.5.2	Random Forest classifier (RF)	51
5.5.3	Naïve Bayes classifier (NB)	51
5.5.4	SVM classifier (SVM)	52
5.5.5	NN classifier (NN)	52
5.5.6	XGBoost classifier (XG)	52
5.6	System Algorithm	53
6	Experimental Results	61
6.1	Testing environment and tools:	61
6.2	Population size	61
6.3	Validation	62
6.4	Evaluation metrics	63
6.4.1	Accuracy	64
6.4.2	Recall	64
6.4.3	Precision	64
6.4.4	F1-Score	64
6.4.5	Balanced Accuracy	64
6.4.6	Confusion Matrix	65
6.4.7	Execution Time	65
6.5	Results	65
6.5.1	Decision tree classifier (DT)	66
6.5.2	Random Forest classifier (RF)	67
6.5.3	Naïve Bayes classifier (NB)	68
6.5.4	Support Vector Machines classifier (SVM)	69
6.5.5	XGBoost classifier (XG)	70
6.5.6	Nearest Neighbors Classifier (NN)	71
6.5.7	Average statistics for all classifiers on the more balanced dataset	72
6.5.8	Average statistics for all classifiers on the less balanced dataset	73
6.6	Comparison to previous work	74
7	Discussion	75

7.1	Summary of findings .....	75
7.2	Limitations.....	75
8	Conclusion .....	76
8.1	Conclusion.....	76
8.2	Future work .....	77
	References .....	78
	Appendix A: Statistical Analysis results for large route leaks incidents .....	82
A.1	VolumeDrive 2014 incident [21] .....	83
A.2	Dodo Network 2012 incident [24].....	85
A.3	Verizon 2014 incident [25].....	87
A.4	Amazon 2015 incident [26].....	89
A.5	Con-Edison [38] .....	91
A.6	Google 2015 incident [27].....	93

# List of Tables

TABLE 4-1 FEATURES FOR ROUTE LEAKS ANALYSIS .....	32
TABLE 4-2 VERIZON INCIDENT FEATURES DURING LEAK PERIODS WITH RESPECT TO NORMAL PERIODS(X) .....	33
TABLE 4-3 AMAZON INCIDENT FEATURES DURING LEAK PERIODS WITH RESPECT TO NORMAL PERIODS(X) .....	33
TABLE 4-4 VOLUME DRIVE INCIDENT FEATURES DURING LEAK PERIODS WITH RESPECT TO NORMAL PERIODS(X) .....	34
TABLE 4-5 DODO INCIDENT FEATURES DURING LEAK PERIODS WITH RESPECT TO NORMAL PERIODS(X).....	34
TABLE 4-6 COMPARISON BETWEEN INCIDENTS DURING LEAK PERIODS WITH RESPECT TO NORMAL PERIODS(X) .....	34
TABLE 5-1 ROUTE LEAKS DATASET .....	41
TABLE 5-2 CLASSIFICATION FEATURES .....	43
TABLE 5-3 PATRICIA TRIE COMPLEXITY.....	47
TABLE 5-4 COMPARISON BETWEEN METHOD 1 AND 2.....	48
TABLE 6-1 CONFUSION MATRIX.....	65
TABLE 6-2 DECISION TREE CLASSIFIER- CLASSIFICATION RESULTS FOR MORE AND LESS BALANCED DATASET .....	66
TABLE 6-3 RANDOM FOREST CLASSIFIER- CLASSIFICATION RESULTS FOR MORE AND LESS BALANCED DATASET .....	67
TABLE 6-4 NAIVE BAYES CLASSIFIER- CLASSIFICATION RESULTS FOR MORE AND LESS BALANCED DATASET .....	68
TABLE 6-5 SVM CLASSIFIER- CLASSIFICATION RESULTS FOR MORE AND LESS BALANCED DATASET.....	69
TABLE 6-6 XGBOOST CLASSIFIER- CLASSIFICATION RESULTS FOR MORE AND LESS BALANCED DATASET .....	70
TABLE 6-7 NEAREST NEIGHBOR CLASSIFIER- CLASSIFICATION RESULTS FOR MORE AND LESS BALANCED DATASET.....	71
TABLE 6-8 AVERAGE CLASSIFIERS RESULTS FOR THE MORE BALANCED DATASET .....	72
TABLE 6-9 AVERAGE CLASSIFIERS RESULTS FOR THE LESS BALANCED DATASET .....	73
TABLE 6-10 RESULTS OF RLD SYSTEM.....	74
TABLE A-0-1 VOLUME DRIVE STATISTICS .....	83
TABLE A-0-2 DODO STATISTICS.....	85
TABLE A-0-3 VERIZON STATISTICS.....	87
TABLE A-0-4 AMAZON STATISTICS.....	89
TABLE A-0-5 CONED STATISTICS .....	91
TABLE A-0-6 GOOGLE STATISTICS.....	93

# List of Figures

FIGURE 2-1 ASES INTER RELATIONSHIPS .....	5
FIGURE 2-2 VALLEY FREE RULES .....	6
FIGURE 2-3 MACHINE LEARNING CATEGORIES .....	9
FIGURE 2-4 DECISION TREE CLASSIFIER EXAMPLE .....	10
FIGURE 2-5 LINEAR SVM CLASSIFIER EXAMPLE .....	12
FIGURE 2-6 NAÏVE BAYES CLASSIFIER EXAMPLE .....	13
FIGURE 2-7 ROUTE LEAKS SOLUTION CATEGORIES .....	14
FIGURE 3-1 THE FIRST FOUR TYPES OF THE SIX-TYPES TAXONOMY .....	20
FIGURE 3-2 VOLUME DERIVE INCIDENT .....	21
FIGURE 3-3 TELECOM MALAYSIA INCIDENT .....	22
FIGURE 3-4 DODO NETWORK INCIDENT .....	23
FIGURE 3-5 AMAZON INCIDENT .....	25
FIGURE 3-6 GOOGLE INCIDENT .....	26
FIGURE 3-7 BELARUSIAN HIJACKING INCIDENT .....	27
FIGURE 5-1 CONDITIONS FOR PRACTICAL SOLUTION TO ROUTE LEAKS PROBLEM .....	36
FIGURE 5-2 SYSTEM BLOCK DIAGRAM .....	38
FIGURE 5-3 BGP UPDATES FILE IN MRT FORMAT-WITHDRAWAL MESSAGE .....	39
FIGURE 5-4 BGP UPDATES FILE IN MRT FORMAT-UPDATES MESSAGE .....	40
FIGURE 5-5 FEATURES EXTRACTION METHOD 1(MYSQL) .....	44
FIGURE 5-6 FINDING MORE SPECIFIC PREFIXES IN COMPARISON STEP -ALGORITHM .....	45
FIGURE 5-7 BINARY TRIE EXAMPLE .....	46
FIGURE 5-8 PATRICIA TRIE EXAMPLE .....	47
FIGURE 5-9 FEATURES EXTRACTION METHOD 2(PATRICIA TRIE) .....	48
FIGURE 5-10 SEARCH AND INSERT PREFIXES IN PATRICIA TRIE ALGORITHM .....	49
FIGURE 5-11 SYSTEM ALGORITHM-PART1 .....	53
FIGURE 5-12 SYSTEM ALGORITHM-PART2 .....	54
FIGURE 5-13 SYSTEM ALGORITHM -PART3 .....	55
FIGURE 5-14 SYSTEM ALGORITHM -PART4 .....	56
FIGURE 5-15 SYSTEM ALGORITHM -PART5 .....	57
FIGURE 5-16 SYSTEM ALGORITHM -PART6 .....	58
FIGURE 5-17 SYSTEM ALGORITHM -PART7 .....	59
FIGURE 5-18 SYSTEM ALGORITHM -PART8 .....	60
FIGURE 6-1 DISTRIBUTION OF CLASSES IN THE LESS BALANCED DATASETS .....	62
FIGURE 6-2 DISTRIBUTION OF CLASSES IN THE MORE BALANCED DATASET .....	62
FIGURE 6-3 THE 10 CROSS FOLD VALIDATION PROCESS .....	63

## List of Symbols and Abbreviations

AS	Autonomous System
AS-Path	Autonomous Systems Path
ASN	Autonomous System Number
BGP	Border Gateway Protocol
DAG	Direct Acyclic Graph
DT	Decision Tree classifier
FN	False Negative
FP	False Positive
GBoost	Gradient Boosting classifier
ISP	Internet Service Provider
KNN	K-Nearest Neighbor classifier
NB	Naïve Bayes classifier
RF	Random Forest classifier
SVM	Support Vector Machine classifier
TCP	Transmission Control Protocol
TN	True Negative
TP	True Positive

# Abstract

The route leaks problem is considered one of the unsolved BGP problems since more than fifteen years ago. It has a large negative impact on the global internet stability and reliability. A route leak happens when an autonomous system advertises reachability information with a violation of autonomous inter-relationships policies. This problem is hard to prevent due to human errors and misconfigurations, and hard to detect due to the confidentiality of autonomous systems relationships and lack of publicly-advertised datasets.

Traditional solutions to the route leaks problems use one of the following methods: the autonomous systems relationships inference, adding information about connection types to the BGP Update messages, or gathering and comparing information from global vantage points. These solutions were not widely applicable because they may be incomplete, have high time and processing cost, depend on third-parties, or require modification to the existing protocols.

In this study we address the route leaks problem from three aspects, firstly we propose a new taxonomy to the different types of route leaks depending on their effects of the BGP update messages other than existing taxonomy that depends only on the relationships between autonomous systems which could not help much as these types of information are confidential.

Secondly, we gather and label the first route leaks dataset which consists of all the valid route leaks incidents published on the internet. We use this dataset to generate features list which we used as training and testing data to our classifiers. To accelerate the features generation process we use a memory structure called “Patricia Trie” inspired by the structures used in the router’s lookup tables. This structure has reduced the average overall searching and updating run time from several hundreds of minutes to very few seconds.

Lastly, we propose a complete real-time widely applicable detection system to the route leaks updates using supervised learning method, and using the classification technique that can be applied to each border router without adding high time or computational overhead. In our study, we apply and compare between six different classifiers namely (Decision Tree, Random Forest Trees, Support Vector Machines, Naïve Bayes, Nearest Neighbor, and Gradient Boosting) classifiers. Our system proves that it can detect and classify route leaks from normal updates on two datasets the “less balanced” and the “more balanced”. The system achieves the best accuracy of 88%, 83% and F1 Score of 92%, 98% for the two datasets respectively. These results can be achieved with a system that runs in less than one second.