Innopolis University

Secure Systems and Networks
Research project

Submitted By:
Vasiliy Podtikhov,
Bulat Saifullin,
Timur Samigullin

Submitted To:
Rasheed Hussain
Azat Safin
Konstantin Urysov
Kirill Saltanov

December 5, 2016

Contents

Abstract

In modern day social networks become widely used. Practically almost all employers using them. But they can be used to formating public opinion in way not acceptable by company, or by accident share some confidentially information. This often happened because ordinary employee don't unaware of global company goals.

In this work we will try to link a social identity to an IP address by analysis of user traffics. This will help us to establish leakage, find disgruntled employees and change company politics to prevent this situations.

1 Introduction

Mapping IP address to account on social network is generally believed to be difficult for an individual with no dedicated infrastructure or privileged information. Social networks owners such as Vk.com and Facebook.com have this information, but they always hide it except in the case of a legal decision. But this information may be very handy in big corporations. In average 60% of employee actively use social networks [1]. And sometime employees post trade secret in social network, usually they use fake name. But if the employee go in account while he in the corporation's network mapping IP address to account on social network, can help us to find him.

Our project is suitable for enterprise companies that has its own network infrastructure and who care about information security and data losses. Nowadays social network are very popular, most of people has accounts in social networks and companies cannot control the datas, that uploaded by their employee to the Internet. It can be very dangerous for security reasons.

2 Related Work

Today we widely used Netflow analysis for security reasons [2][3]. But only recently science works was introduced whom main goal was determine users action in social networks [4][5]. Unfortunately method who helped us to identify user never was introduced. In this paper we tried to find a solution for this problem.

3 Research question

- Find connection between user traffic and profile changes.
- What sending data affects changes in profile?
- How to analysis user's net-flow traffics?
- How to analysis a profile in social network?

ls

4 Methodologies

NetFlow is a feature that was introduced on Cisco routers that provides the ability to collect IP network traffic as it enters or exits an interface. By analyzing the data provided by NetFlow, a network administrator can determine things such as the source and destination of traffic, class of service, and the causes of congestion. A typical flow monitoring setup (using NetFlow) consists of three main components:[1]

- Flow exporter: aggregates packets into flows and exports flow records towards one or more flow collectors.
- Flow collector: responsible for reception, storage and pre-processing of flow data received from a flow exporter.
- Analysis application: analyzes received flow data in the context of intrusion detection or traffic profiling, for example.

We analysed netflow dumps in corporation environment and tried to check if connection was established in period of time and check presence of person in this time period on site. The main purpose is to find correlations between posted time and netflow traffic.

1. VK

Text about subject

1.1 Instagram

Instagram is an online mobile photo-sharing, video-sharing, and social networking service that enables its users to take pictures and videos, and share them either publicly or privately on the app, as well as through a variety of other social networking platforms, such as Facebook, Twitter, Tumblr, and Flickr. [2]

1.1.1 Getting data from instagram

To find correlation between posted time and netflow traffic, we need to get exact time, when user post photo.

Fisrst, all information about post was getting through API functional of Instagram web-site. For this reason third-part application should be registered at instagram development page, and access-token should be got. Request method for getting info about specific post shown bellow:

```
1 https://api.instagram.com/v1/media/shortcode/BNPjFcrh22_?access_token=3955223166.3a064fe.2562
      f48363ac48f8b002f713fddeae2e
```

With testing accounts everything was fine, but when I tried to access to another real account I have a API error. Instagram API has one speciality: there is a sandbox environment for testing reasons, and for getting data from every page, first, he should be invited to sandbox and he should accept requests from application, even if page open for everybody. I think, application should not ask requests for viewing user's page.

The way out is to parse raw html page, and get data from it. In source of html page I saw, that there is raw json data.

It means, that I can simply get all data from specific web page even without any authorization. For parsing I used xpath and re library from python.

```
1 script = html.xpath('//script[contains(., "window._sharedData")]/text()')[0]
2 data = re.search(r"window._sharedData = (.*?);\$", script).group(1)
3 data = json.loads(data)
```

All data was stored in **data** array.

1.1.2 Finding range of instagram ip

We ask to give us all university netflow traffic for analysing, it is stored on our computers. The difference of Instagram in comparison with other social network, is that Instagram has very narrow area of usage. In this network users can only add and comment their photos. Every users action connected with photos. It means, that instagram need less computer capacity, than other networks.

And also Instagram now belong to facebook. The problem was to find exact range of Instagram ip addresses.

Instagram hasnt got its own autonomous system, but most number of requests send to 31.13.93.72 or 31.13.92.32 or 31.13.93.54. For the first sight we can assume that we should only restrict 31.13.92.0/24 or 31.13.93.0/24. But that is not a solution, because not every address in this network belongs to Instagram.

So, I deided to find all Instagram ip addresses by myself. I extract all unique destination ip addresses from netflow traffic and get 106 MB file with 6291215 lines. I write a script to revesre-resolve all ip addresses and find instagram string in it. It was bad idea. Script worked for three days, but process was not finished. And during this I find another solution for this. I decided to use all 31.13.0.0/16 network, and resolve Instagram ip addresses after it.

1.1.3 Extracting data from netflow

In my script I used only raw nfdump. You can see the whole string filter bellow:

```
\$ nfdump -R /var/flows/MYROUTER "dst net 31.13.0.0/16 and port 443" -o csv -t
    2016/11/27.22:47:26-2016/11/27.22:47:56 -s record/bytes | head -n -3 | sed '1d'
```

The result of such execution:

```
('2016-11-29 15:57:36', '10.240.20.237', '31.13.72.53', '40166', '255667')
('2016-11-29 15:57:22', '10.91.35.114', '31.13.72.8', '57339', '15368')
('2016-11-29 15:57:24', '10.240.20.133', '31.13.92.11', '45988', '8917')
('2016-11-29 15:57:47', '10.240.16.55', '31.13.92.51', '54943', '8843')
('2016-11-29 15:57:35', '10.240.18.181', '31.13.72.53', '62515', '6779')
('2016-11-29 15:57:33', '10.240.16.208', '31.13.72.53', '37487', '5127')
('2016-11-29 15:57:28', '10.242.1.233', '31.13.72.12', '38472', '5122')
```

After filtering only Instagram ip addresses it became:

```
('2016-11-29 15:57:36', '10.240.20.237', '31.13.72.53', '40166', '255667')
('2016-11-29 15:57:47', '10.240.16.55', '31.13.92.51', '54943', '8843')
('2016-11-29 15:57:35', '10.240.18.181', '31.13.72.53', '62515', '6779')
('2016-11-29 15:57:33', '10.240.16.208', '31.13.72.53', '37487', '5127')
```

The main thing, that I should solve is to find necessary time range. At the moment when user post photo to his instagram account, long tcp connection should occur, so this connection can start early or end later, that exact post time. With empirical analysis I detect that I should take 20 seconds offset before exact time post and 10 sec offset after timepost. This range give valid results.

1.1.4 Experimental results

2. Facebook

2.1 Conclusions

In this work some methods of defining internal IP address of users in corporate environment were presented. Scripts for finding correlations between network traffic in companies and time when posts in social networks were written for Facebook, VK, Instagram social networks. We can conclude that to exactly identify user's ip address we should analyze more than one post. And number of posts should be increasing with traffic growth.

2.2 Acknowledgments

Bibliography

[1] Hofstede, Rick; Celeda, Pavel; Trammell, Brian; Drago, Idilio; Sadre, Ramin; Sperotto, Anna; Pras, Aiko. Flow Monitoring Explained: From Packet Capture to Data Analysis with NetFlow and IPFIX". IEEE Communications Surveys Tutorials. IEEE Communications Society.

[2] http://www.businessinsider.com/instagram-2010-11