Social Network Analysis as an Internet Security Tool

Abstract

Security devices (firewalls, IDS, IPS) produces a huge amount of data by posting each security incident/event into a Syslog database. This (big) data enables the system administrators to identify the source of the largest attacks, and the most frequently victimized/targeted server.

However, due to massive number of records generated by Syslogs, a quicker and more timely analysis is needed. **Social Network analysis** is presented here as an optimal way to quickly analyze and create actionable insights from this huge amount of data – by converting (big) data into graphics format.

Compare the typical incident entry in the syslog database:

2853776 2 2013-04-17 10:56:36.653 202.91.161.254 0 23 6 fa-0-1-7206a-dagupan SEC-6-IPACCESSLOGP 12071018: UTC: list 150 denied tcp 114.122.33.16(2963) (Ethernet5/3 0025.9e5d.d0f7) -> 202.91.171.77(445), 1 packet 114.122.33.16 202.91.171.77 00259e5dd0f7

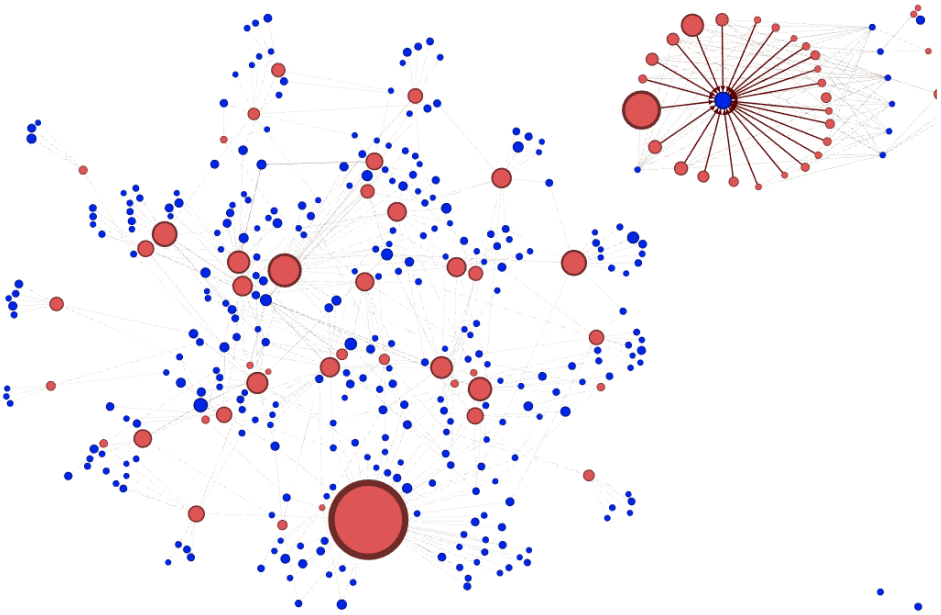
Where the data of interest are shaded:

- in Red (Date and time)
- in Green (Source of the Incident)
- in LightBlue (Destination of the incident)

In a typical hour, thousands of such entries would be appended to the database. A representative screen shot of such incidents would look like this:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
MsgID	EngineID	DateTime	DeviceIP	Acknowledge	syslogFacility	Severity	Hostname	Message Type	Message	SyslogTag	SourceIP	DestinationIP	MacInMessage
2853777	2	56:36.7	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGRL	12071019: UTC: access-list logging rate-limited or missed 459 packets				0015175ac90c
2853776	2	56:36.7	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071018: UTC: list 150 denied tcp 114.122		114.122.33.16	202.91.171.77	00259e5dd0f7
2853775	2	56:34.5	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071017: UTC: list 150 denied tcp 202.152		202.152.199.151	202.91.166.24	002590312c02
2853774	2	56:34.5	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071016: UTC: list 130 denied tcp 192.0.2.		192.0.2.43	122.52.49.46	0015175ac90c
2853773	2	56:32.2	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071014: UTC: list 150 denied tcp 114.39.1		114.39.138.198	202.91.172.94	002590312c02
2853772	2	56:32.2	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071015: UTC: list 130 denied tcp 192.0.2.		192.0.2.43	122.52.49.46	0015175ac90c
2853771	2	56:31.7	74.115.208.105	0	3	3	74.115.208.105		BITSTOP-2A0044C Security: 529: NT AUTHO		192.169.55.45		00259e5dd0f7
2853770	2	56:30.3	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071013: UTC: list 150 denied tcp 85.94.1		85.94.160.140	202.91.175.120	00259e5dd0f7
2853769	2	56:29.2	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071012: UTC: list 150 denied tcp 70.36.2		70.36.237.56	202.91.174.102	00259e5dd0f7
2853768	2	56:28.2	202.91.161.130	0	3	5	www		Security: 538: DAGUPAN.COM\IMAGESRV$: User Logoff: User Name: IMAGESRV				00259e5dd0f7
2853767	2	56:28.0	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071011: UTC: list 150 denied tcp 212.225		212.225.138.119	202.91.170.122	002590312c02
2853766	2	56:27.0	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071010: UTC: list VMUF-in denied tcp 10.		10.21.11.216	98.138.24.49	002590312c02
2853765	2	56:26.8	202.91.162.7	0	3	6	202.91.162.7		snmpd[1854]: Connection from UDP: [202.9		202.91.161.133		00259e5dd0f7
2853764	2	56:26.7	74.115.208.105	0	3	3	74.115.208.105		BITSTOP-2A0044C Security: 529: NT AUTHO		192.169.55.45		00259e5dd0f7
2853763	2	56:26.7	74.115.208.105	0	3	3	74.115.208.105		BITSTOP-2A0044C Security: 529: NT AUTHO		84.241.36.207		00259e5dd0f7
2853762	2	56:25.9	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071009: UTC: list 150 denied tcp 116.236		116.236.205.250	202.91.166.74	00259e5dd0f7
2853761	2	56:24.9	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071008: UTC: list 130 denied tcp 192.0.2.		192.0.2.43	122.52.49.46	0015175ac90c
2853760	2	56:23.9	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071007: UTC: list 130 denied tcp 192.0.2.		192.0.2.43	95.211.138.143	0015175ac90c
2853759	2	56:22.8	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071006: UTC: list 155 denied udp 202.91.		202.91.161.143	199.165.76.11	0002b3ac026d
2853758	2	56:21.3	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071005: UTC: list 150 denied tcp 189.63.8		189.63.8.193	202.91.160.21	002590312c02
2853757	2	56:20.1	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071004: UTC: list VMUF-in denied tcp 10.		10.21.14.170	31.13.76.8	002590312c02
2853756	2	56:19.1	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071003: UTC: list 150 denied tcp 111.242		111.242.12.46	202.91.170.90	0015175ac90c
2853755	2	56:17.8	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071002: UTC: list 130 denied tcp 192.0.2.		192.0.2.43	121.54.58.220	0015175ac90c
2853754	2	56:16.8	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071001: UTC: list 155 denied udp 202.91.		202.91.161.143	133.100.9.2	0002b3ac026d
2853753	2	56:16.7	74.115.208.105	0	3	3	74.115.208.105		BITSTOP-2A0044C Security: 529: NT AUTHO		192.169.55.45		00259e5dd0f7
2853752	2	56:15.7	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12071000: UTC: list 150 denied tcp 184.106		184.106.114.220	202.91.172.85	00259e5dd0f7
2853751	2	56:14.7	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12070999: UTC: list VMUF-in denied tcp 10.		10.21.21.12	124.106.174.162	002590312c02
2853540	2	53:40.3	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12070861: UTC: list 150 denied tcp 38.69.3		38.69.39.114	202.91.160.118	00259e5dd0f7
2853539	2	53:39.3	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12070860: UTC: list 130 denied tcp 192.0.2.		192.0.2.43	124.83.60.172	0015175ac90c
2853538	2	53:37.9	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12070859: UTC: list 150 denied tcp 36.235.1		36.235.180.127	202.91.162.36	002590312c02
2853537	2	53:36.6	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGP	12070858: UTC: list VMUF-in denied tcp 10.		10.21.21.12	63.150.131.10	002590312c02
2853536	2	53:36.6	202.91.161.254	0	23	6	fa-0-1-7206a-dagupan	SEC-6-IPACCESSLOGRL	12070857: UTC: access-list logging rate-limited or missed 435 packets				0024e844da8f

Now compare the above Syslog list with its equivalent output in Gephi:



It is now easier for both Administrators and C-Level Executives to get a 'bird's eye' view of what is happening in their networks. The Red colored Circles represent NODES that are the **sources** of attacks, while the blue colored circles represent the target/**destination** of such attacks. The larger the size of the red circles means that there are more attacks coming out of these nodes. The thicker the lines/edges, the more traffic originates from the same source to same destination nodes.

Four Quick and Easy Steps to Social Network Analysis

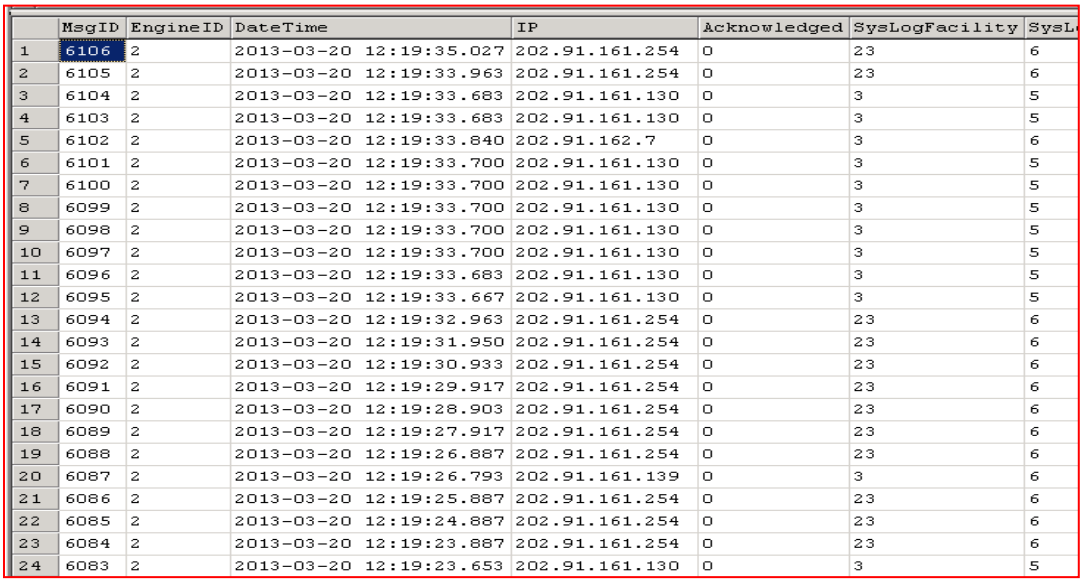
- Extract Syslog SQL database records and save as CSV file format
- Convert CSV into Gephi GEXF format by using **Table2NET** [1]
  http://tools.medialab.sciences-po.fr/table2net/
- Load the GEXF file into Gephi
- Configure Gelphi and Extract Graph as picture.

[1] Table2NET was shared by Paul Alford in this Facebook Post:
https://www.facebook.com/groups/140630009439814/permalink/146327902203358/

KeyStep1: Extract Data

Log in to the SQL Server via SQL MMC or SQL Query and issue the command to extract a single day's worth of syslog data:

```
Select * from syslog where [datetime]>=('4-20-2013') and [datetime]<('4-21-2013')
```

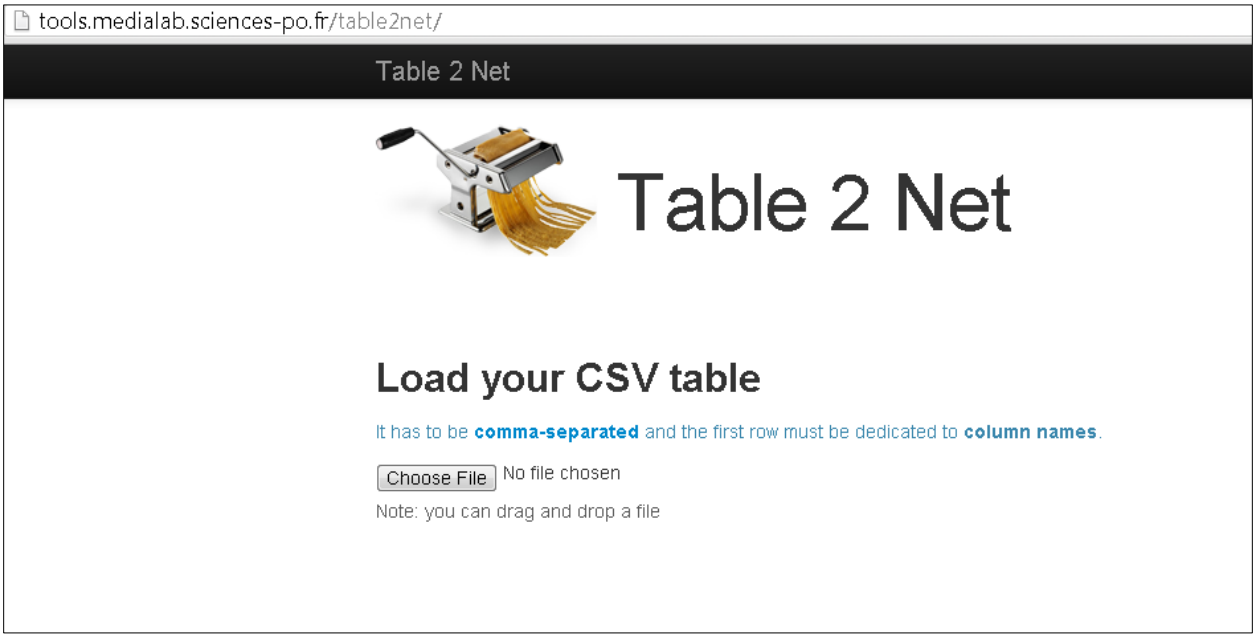


	MsgID	EngineID	DateTime	IP	Acknowledged	SysLogFacility	SysL
1	6106	2	2013-03-20 12:19:35.027	202.91.161.254	0	23	6
2	6105	2	2013-03-20 12:19:33.963	202.91.161.254	0	23	6
3	6104	2	2013-03-20 12:19:33.683	202.91.161.130	0	3	5
4	6103	2	2013-03-20 12:19:33.683	202.91.161.130	0	3	5
5	6102	2	2013-03-20 12:19:33.840	202.91.162.7	0	3	6
6	6101	2	2013-03-20 12:19:33.700	202.91.161.130	0	3	5
7	6100	2	2013-03-20 12:19:33.700	202.91.161.130	0	3	5
8	6099	2	2013-03-20 12:19:33.700	202.91.161.130	0	3	5
9	6098	2	2013-03-20 12:19:33.700	202.91.161.130	0	3	5
10	6097	2	2013-03-20 12:19:33.700	202.91.161.130	0	3	5
11	6096	2	2013-03-20 12:19:33.683	202.91.161.130	0	3	5
12	6095	2	2013-03-20 12:19:33.667	202.91.161.130	0	3	5
13	6094	2	2013-03-20 12:19:32.963	202.91.161.254	0	23	6
14	6093	2	2013-03-20 12:19:31.950	202.91.161.254	0	23	6
15	6092	2	2013-03-20 12:19:30.933	202.91.161.254	0	23	6
16	6091	2	2013-03-20 12:19:29.917	202.91.161.254	0	23	6
17	6090	2	2013-03-20 12:19:28.903	202.91.161.254	0	23	6
18	6089	2	2013-03-20 12:19:27.917	202.91.161.254	0	23	6
19	6088	2	2013-03-20 12:19:26.887	202.91.161.254	0	23	6
20	6087	2	2013-03-20 12:19:26.793	202.91.161.139	0	3	6
21	6086	2	2013-03-20 12:19:25.887	202.91.161.254	0	23	6
22	6085	2	2013-03-20 12:19:24.887	202.91.161.254	0	23	6
23	6084	2	2013-03-20 12:19:23.887	202.91.161.254	0	23	6
24	6083	2	2013-03-20 12:19:23.653	202.91.161.130	0	3	5

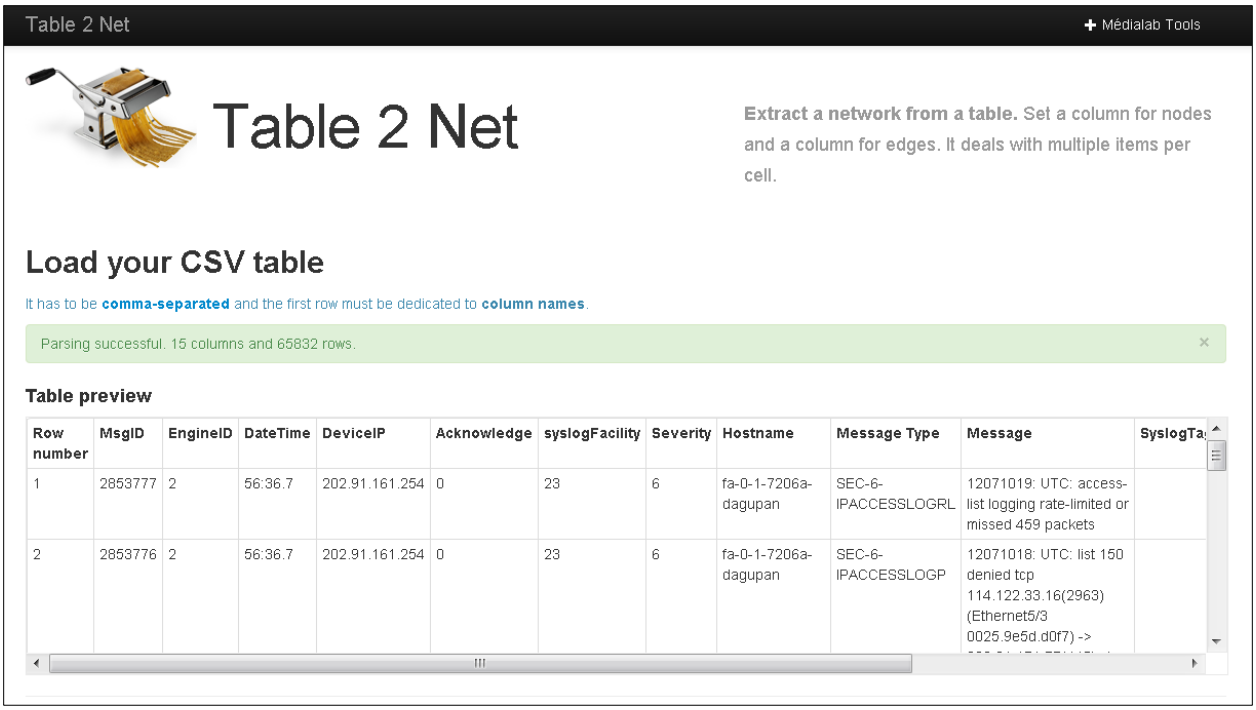
Right click on the output and select [Save AS] and name the file "Syslog-2013-4.csv"

KeyStep2: Convert into Gephi Format

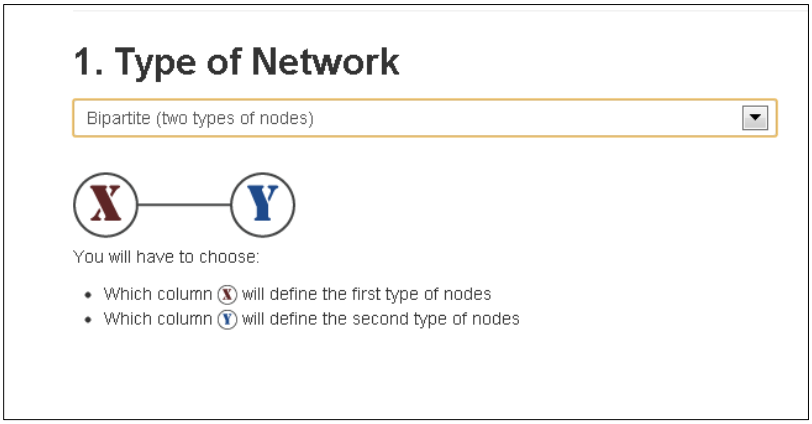
Using the CSV file extracted from Step1, I opened a browser and visited this website: http://tools.medialab.sciences-po.fr/table2net/ to upload and convert the CSV file into gephi format.



Click on “Choose File” and upload the syslog2013-4.csv. Once the upload is finished, you will see a screen similar to this:



Scroll down this screen and select [Bipartite (Two Types of Nodes)] from the Type of Network dropdown.



Then select **SourceIP** for the First Type of nodes and then select **DestinationIP** as the Second Type of Nodes. The screen should be similar to this:

2. Nodes

Which column defines the *first type* of nodes?

SourceIP

Comma-separated ","

Sample of nodes extracted with these settings: (sample)

212.59.28.8 112.198.77.241 124.107.246.180 125.60.240.234 112.206.53.127

Do you want attributes for the *first type* of nodes?

Select one or several columns

Which column defines the *second type* of nodes?

DestIP

Comma-separated ","

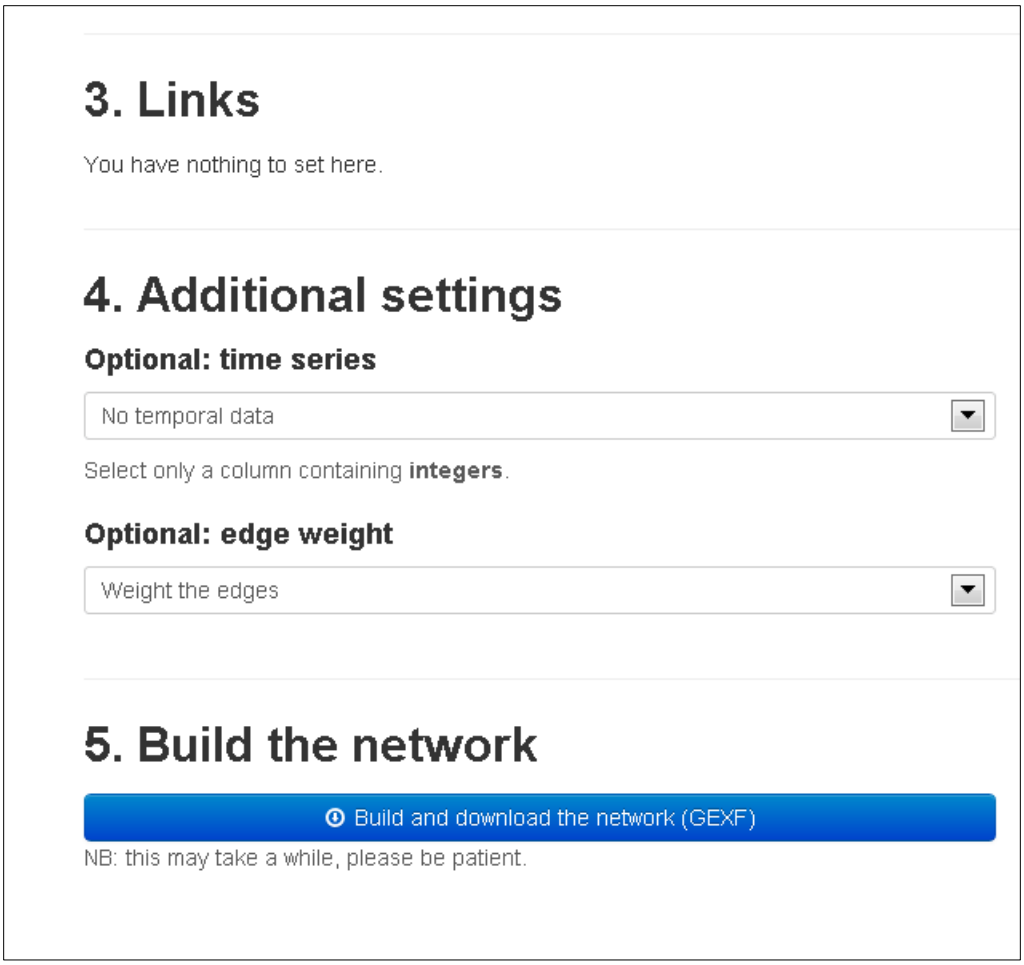
Sample of nodes extracted with these settings: (sample)

202.91.163.100 202.91.163.60 202.91.163.31 202.91.163.31 202.91.163.31

Do you want attributes for the *second type* of nodes?

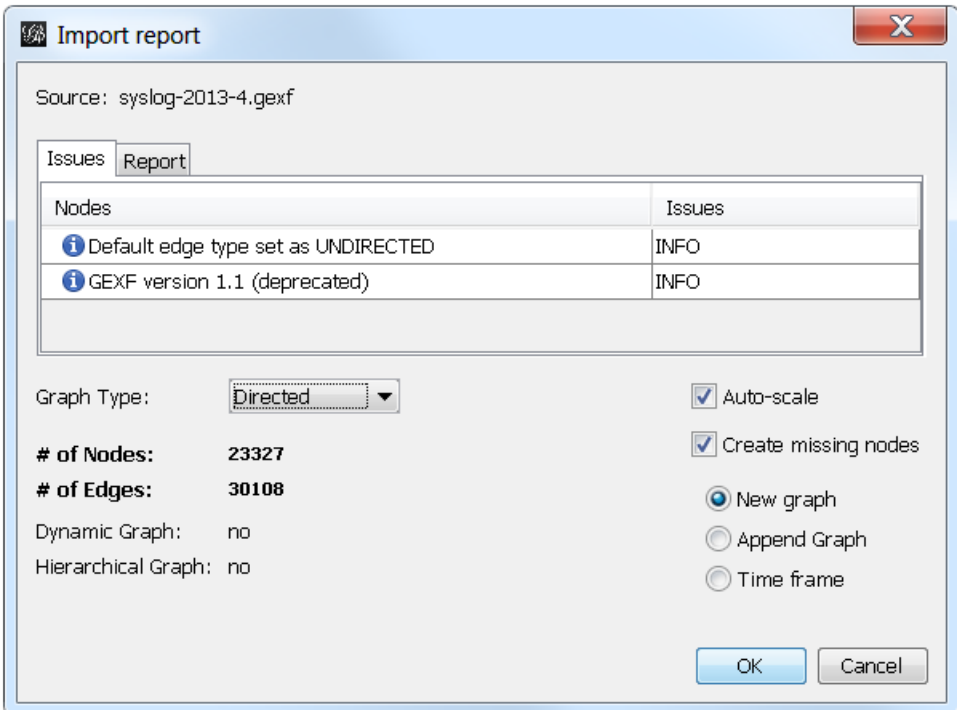
Select one or several columns

Then in the optional items, I choose to enable the option [weight the edges]. Then hit Build and download the network (GEXF) as show below:

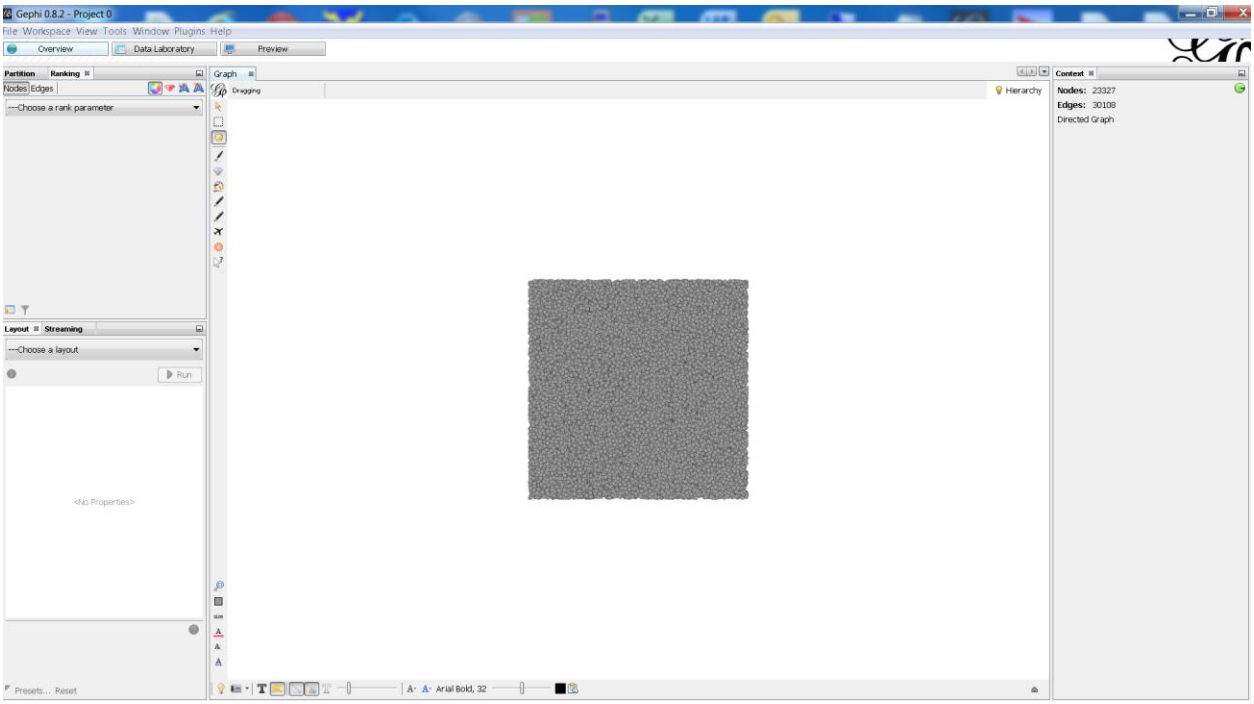


KeyStep3: Load GEXF into Gephi

Run Gephi and open the downloaded Syslog-2013-4.GEXF file.



I am presented with this screen once the GEXF file is loaded:



Data cleansing/Exclusions:

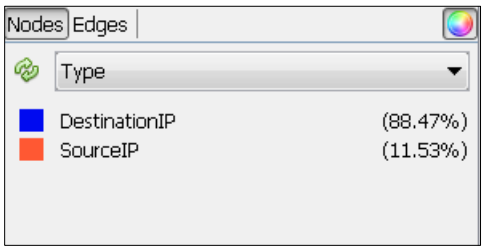
I excluded records that had the following source IPs:

Source IP	Occurence
(blank)	10036
192.169.55.45	8894
192.0.2.43	6744
202.91.161.143	1733
202.91.161.153	1035
0.0.0.0	855

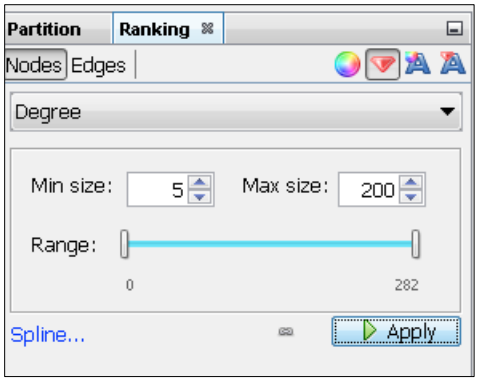
Non Actionable source IPs :
The Blank IP address and 0.0.0.0 do not contain actionable IP addresses.

Known False Positives:
The IP 192.169.55.45 is the internal IP address of our US Server (74.115.208.105). These syslog entries consists of known "Logon Failure event" that is the "expected behavior", while the IP 192.0.2.43 is the IP address of our firewall itself that generate heartbeat packets. These are known to be "false positives". The same is true with both 202.91.161.143 and 202.91.161.153 which are company owned IP addresses that are subject to Access control list that were triggered.

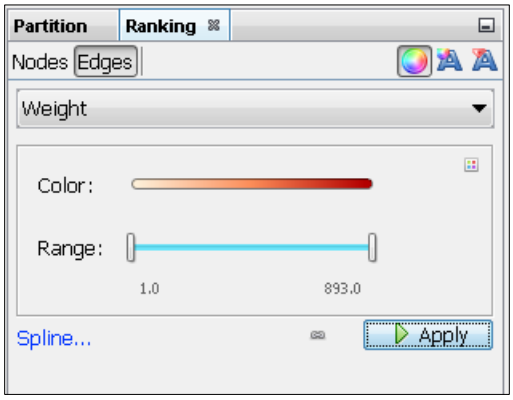
I then use **Partition/Type** to color code the sourceIP (Red) and destinationIP (Blue). This helps to clearly show us which nodes are the source (of attacks) and which nodes are the destination (targets) of attacks.



Then I went to **Ranking/Nodes** and selected **Degree** to apply different Node Sizes (5 to 200) based on it. I wanted to be able to visually identify nodes that are the source of most attacks or destination of the most attacks.



Then I applied Ranking/edges/**Weight. I wante**d to be able to identify the occurrence of each combination of source and destination nodes. The thicker the lines, the more the occurrence of both the source and destination in events.



I then computed for the following statistics: Average Degree, Modularity, EigenVector Centrality, and Average Path Length. This will give us the following values:

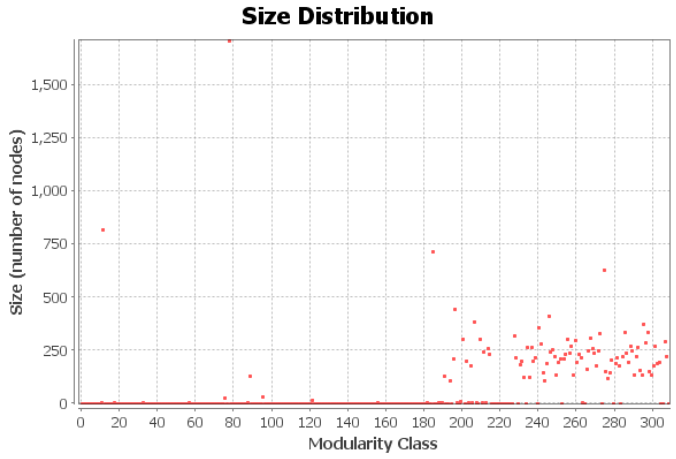
Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0.818
Modularity with resolution: 0.818
Number of Communities: 309

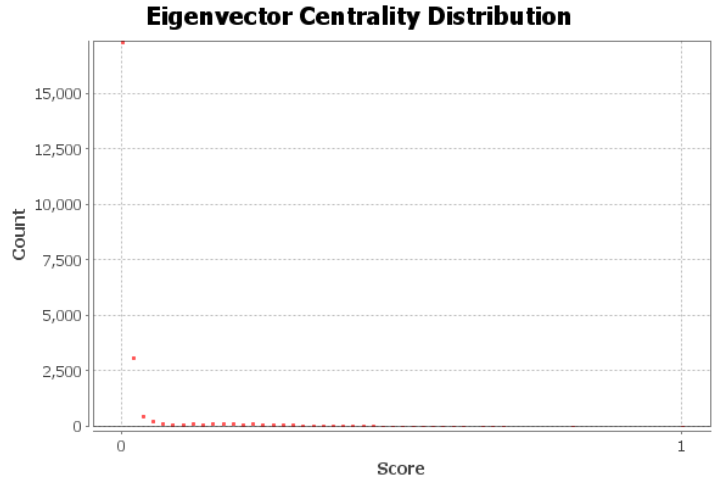


Eigenvector Centrality Report

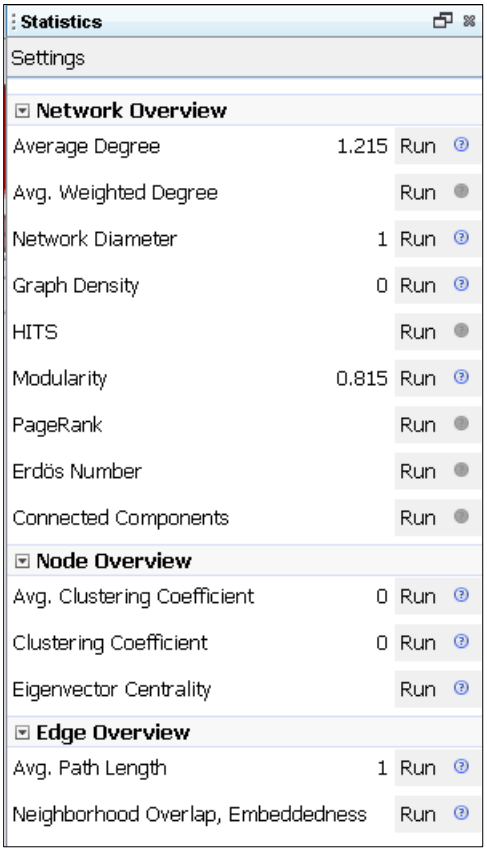
Parameters:

Network Interpretation: directed
Number of iterations: 100
Sum change: 0.0

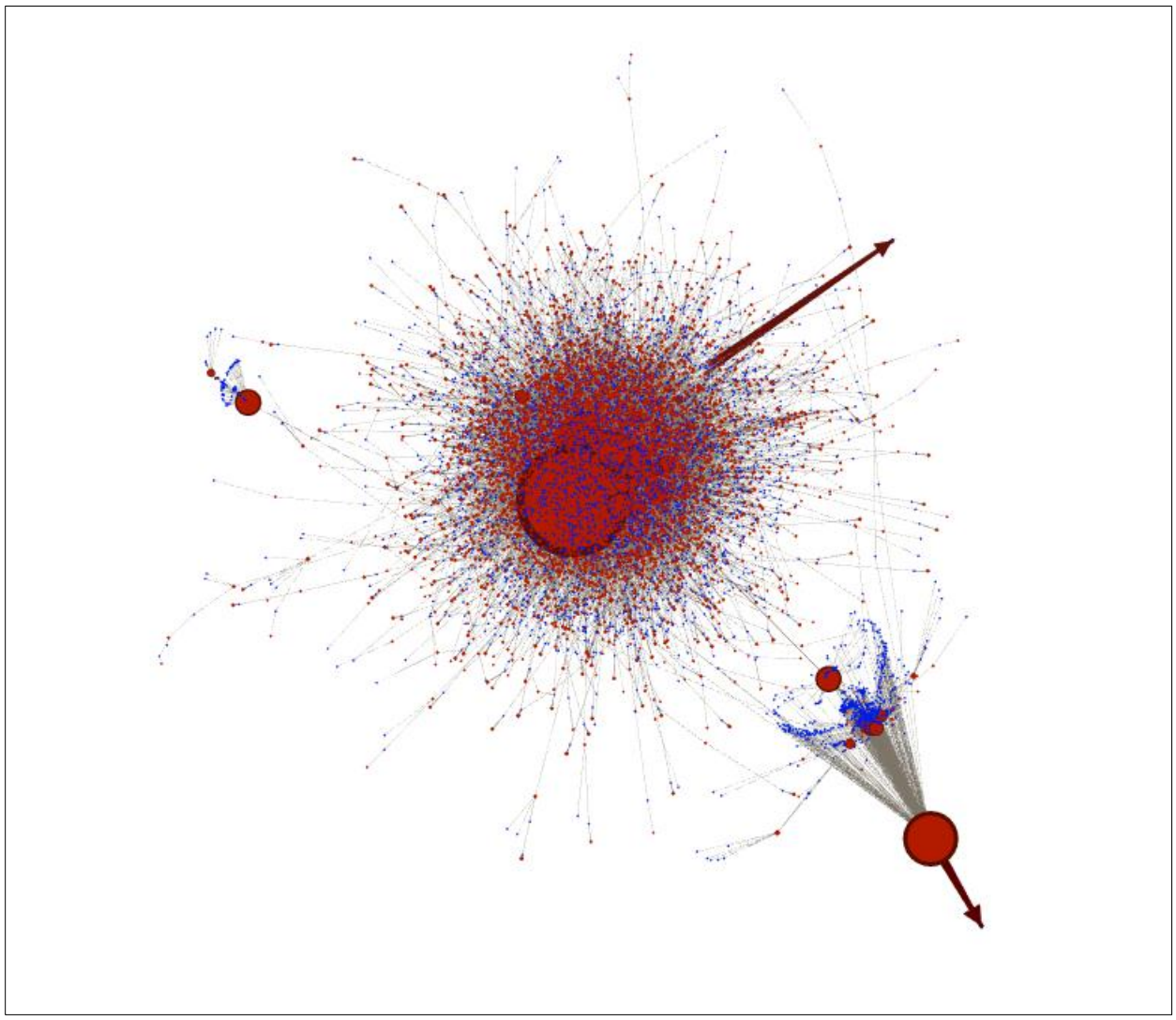
Results:



Here are the statistics AFTER computation is finished:

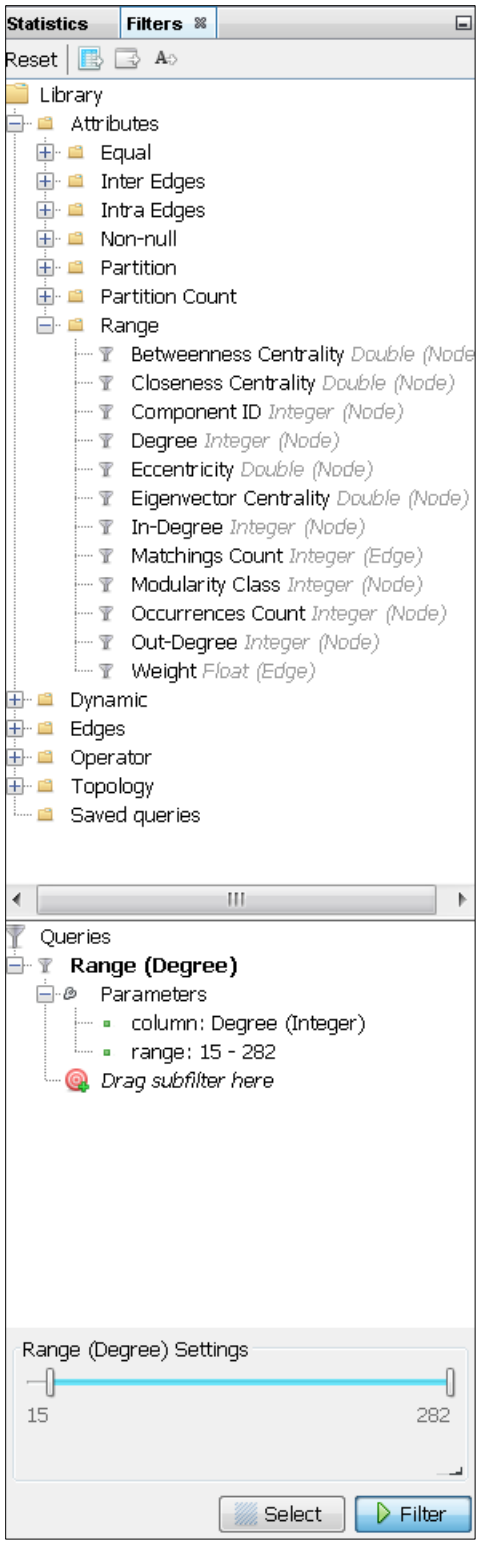


I applied **ForceAltas2** layout then hit Run. I then get an output like this:

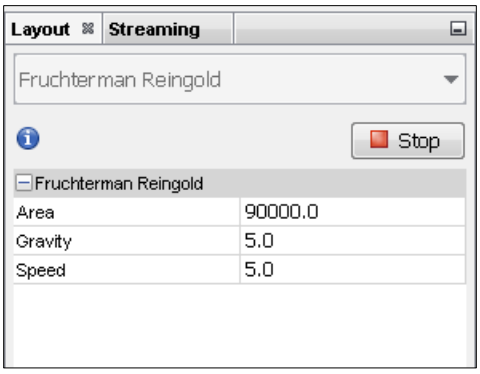


This graph has too many data points and it takes a long time to process. Next I proceeded to filter it to reduce the data points to smaller but still significant data population for us to analyze. The use of Giant component did not reduce the node population by a significant degree.

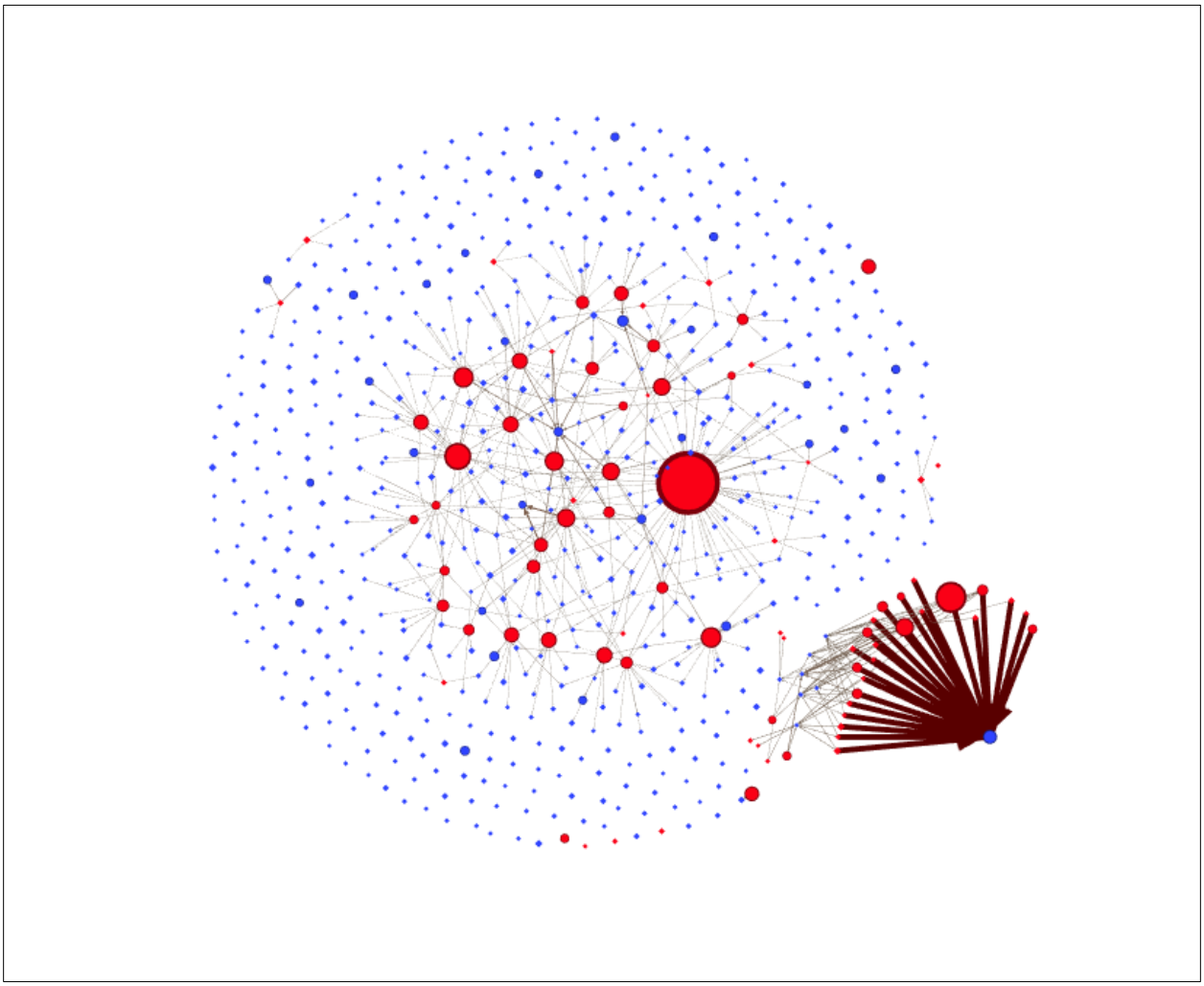
So I then filter based on the range of Degrees that each NODE has. It reduces the number of nodes to only those that have more than 15 incidents (sum of either sourceip or destinationIP occurrence). Items with less than 15 degrees will be filtered out. The nodes with less than 15 incidents are deemed to be **'uninteresting'**. This filter helps to **focus the analysis** on the larger events.



Then I applied the **Fruchterman Reingold** to the Layout, enable Text labeling, apply **Noverlap** and finally applied the **Label Adjust** to the layouts.

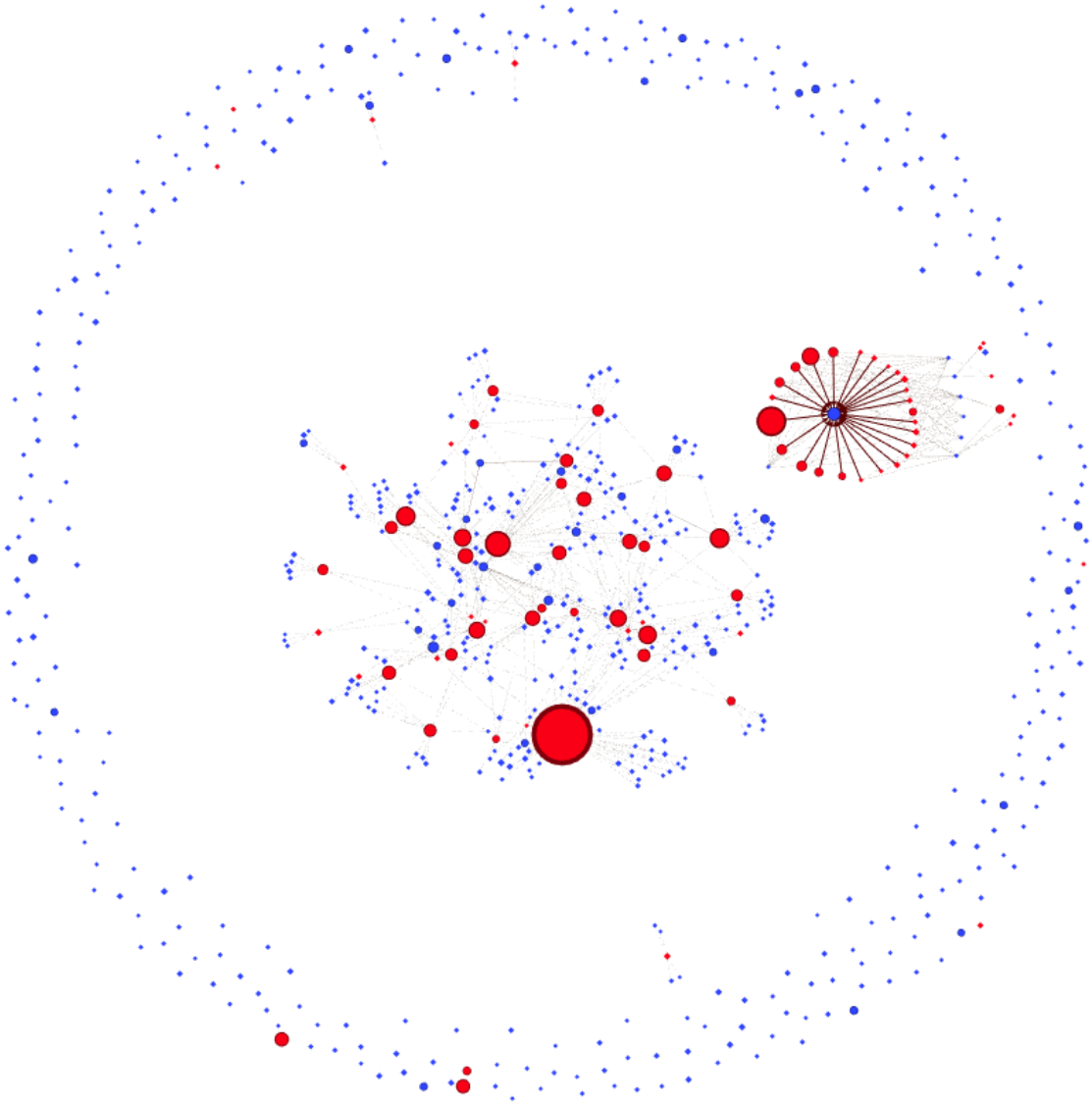


I got a graphics like this:

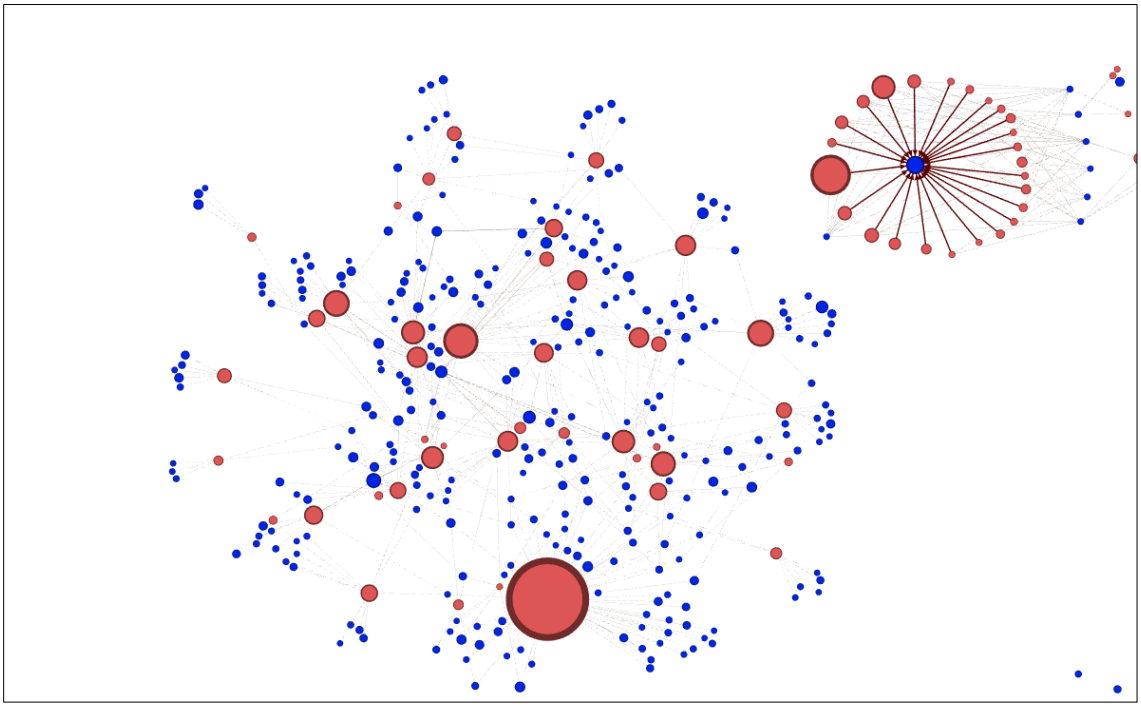


From here, I could now easily see the number of 'Attackers' (in red) and whose node size indicates the larger number of nodes it 'attacks' (the larger the size of the red circles, the more nodes it attacks). Then there are 'thick red edges' that denote the weight (intensity) of the attacks against node/s

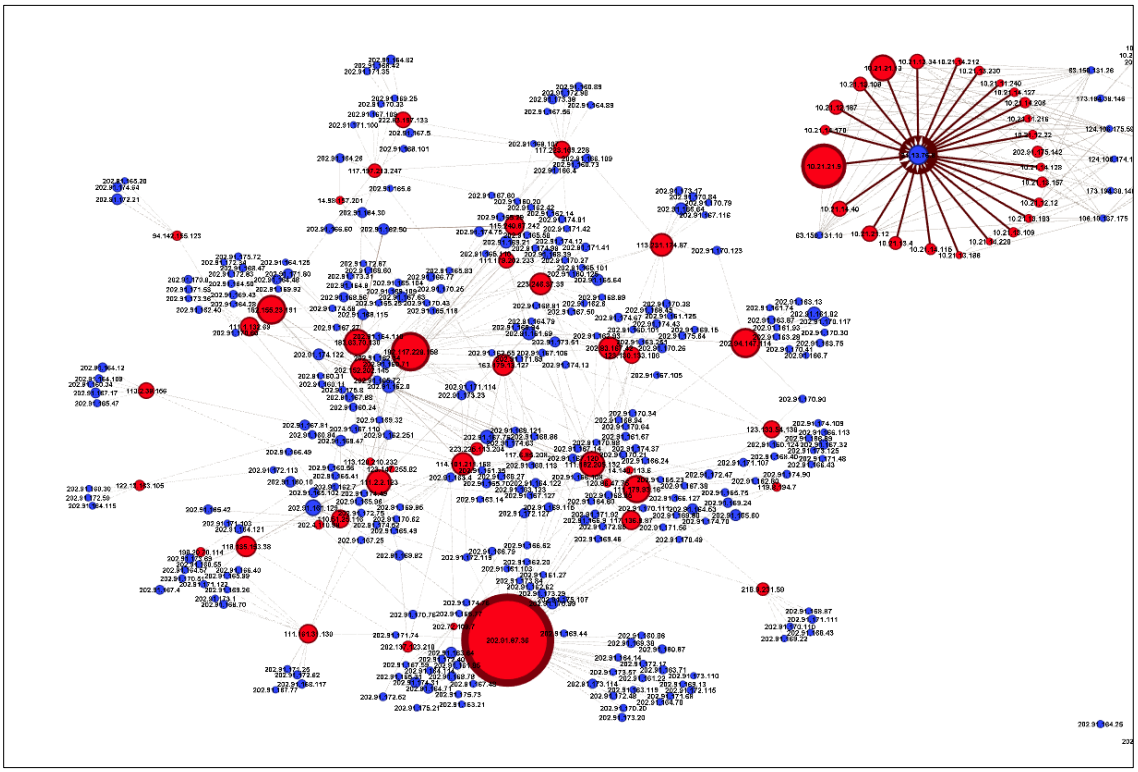
(destinationIPs). I also experimented with another layout by using **Yifan HU** and elected to use this Layout:



If I discount the nodes at the outer edges as ‘uninteresting’ and zoom-in I get this:



With TEXT label enabled (IP addresses shown in black):



Interpretation:

Limitation of data

As the data from the Syslogs only had Sourceip and DestinationIP addresses pairings, I had to set this up as a Bi-Partite Network as opposed to choosing Normal or the Citation type network. Thus the average clustering coefficient is 0. It is not surprising that we get the average path length and Network Diameter values to be 1.

This meant that our virtualization is limited to showing a single network's perspective, so it does not show nodes BEHIND the SourceIP. A hacker might conceivably be controlling several SourceIP in launching an attack against a network, and our syslog data will only show the different SourceIPs but NOT the IP of the hacker controlling the different SourceIPs (attacking Nodes). There would be no links from the attacking Nodes back to the hacker node.

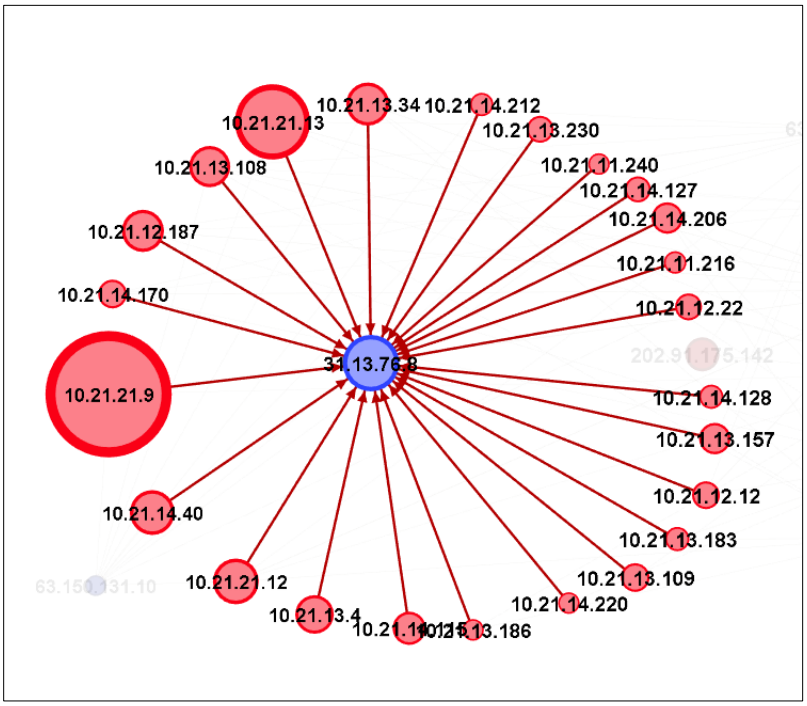
We therefore lack the capability to detect communities outside of our own network. We are equally unable to use any of the centrality measures in a significant way.

New Insights from Gephi Graphic:

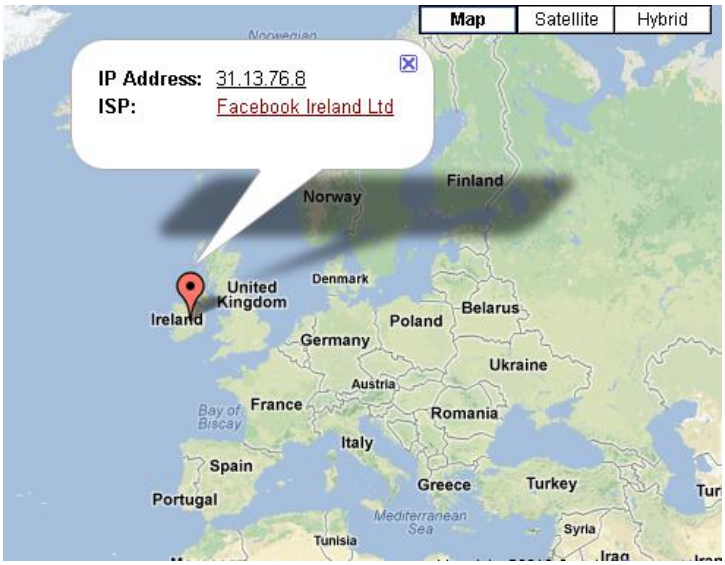
We have a new found ability to visualize the network incidents to easily show Attackers (sourceIP in **red circles**) and Victims (DestinationIP **in blue circles**). We also gain *additional insights* into the magnitude of each attack from the **thickness of the edges**, and the number of nodes targeted from the **size of the attacking Nodes** itself.

In particular, we find among several items of interest, the following:

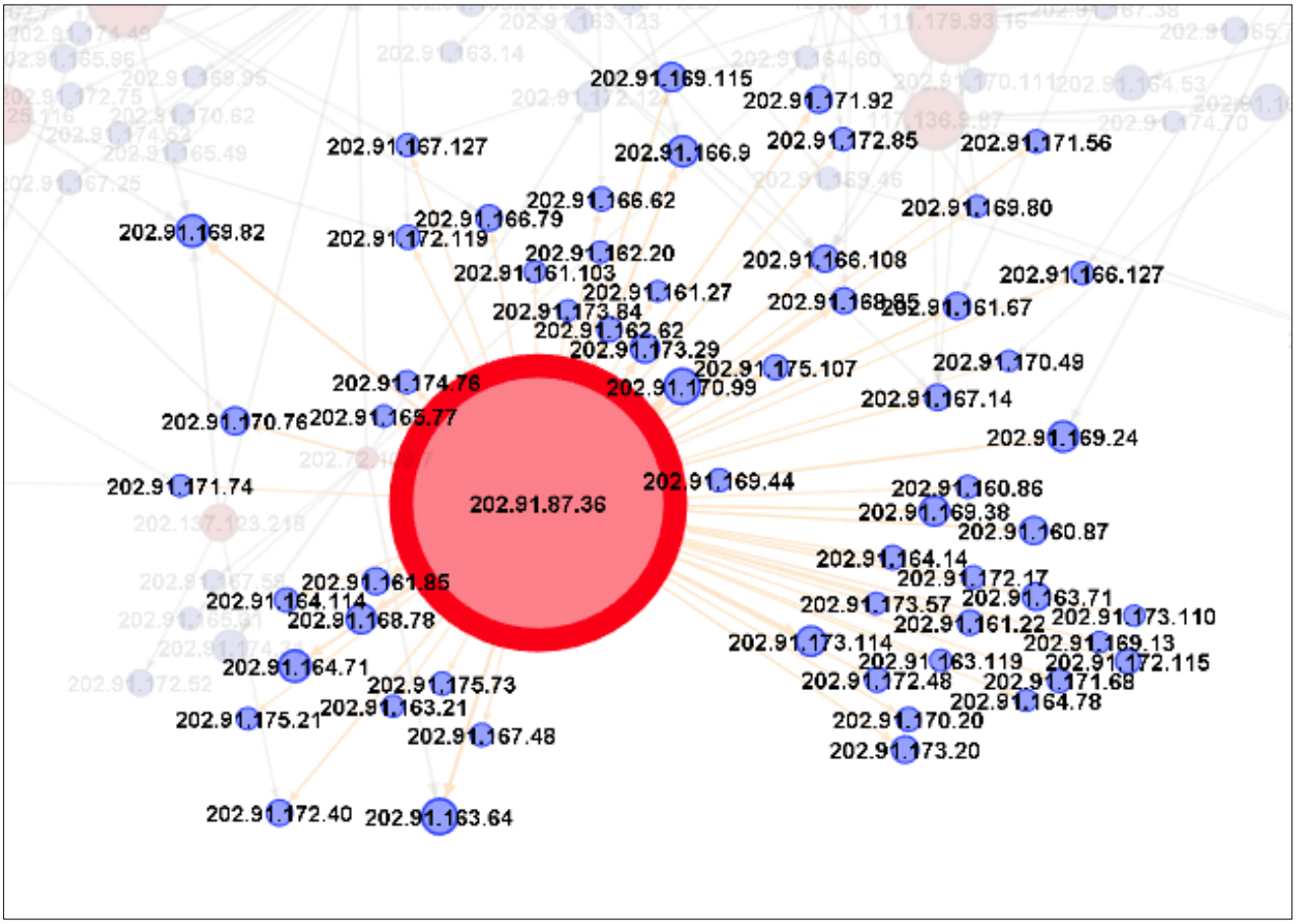
1. Several in house machines (Nodes) are accessing forbidden FACEBOOK website.
   I zoomed in on the lower right corner and reproduced it below:



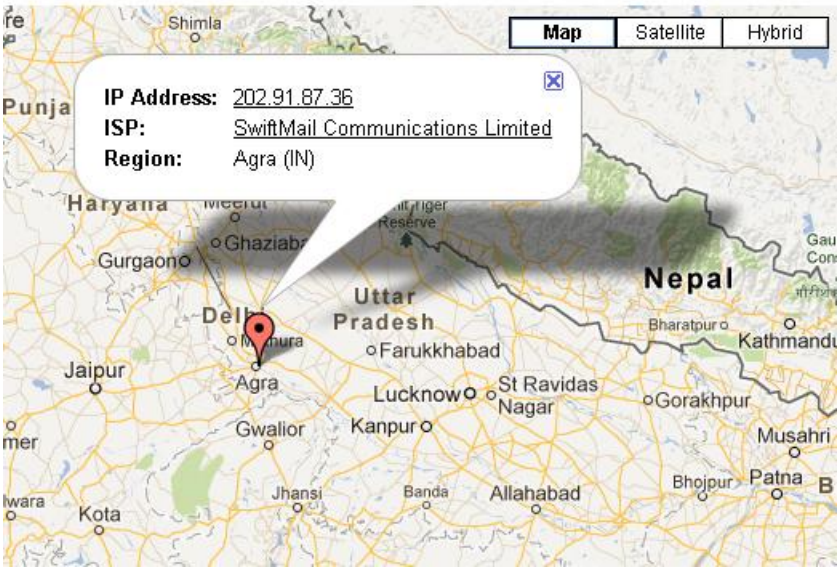
This shows a significant number of NODEs (sourceIP) that are connecting ('attacking') to the same destination node: **31.13.76.8**. A research shows that this IP address belongs to Facebook. Note: Facebook is banned in the internal network, these syslog events quickly shows up the number of machines attempting to connect to Facebook.



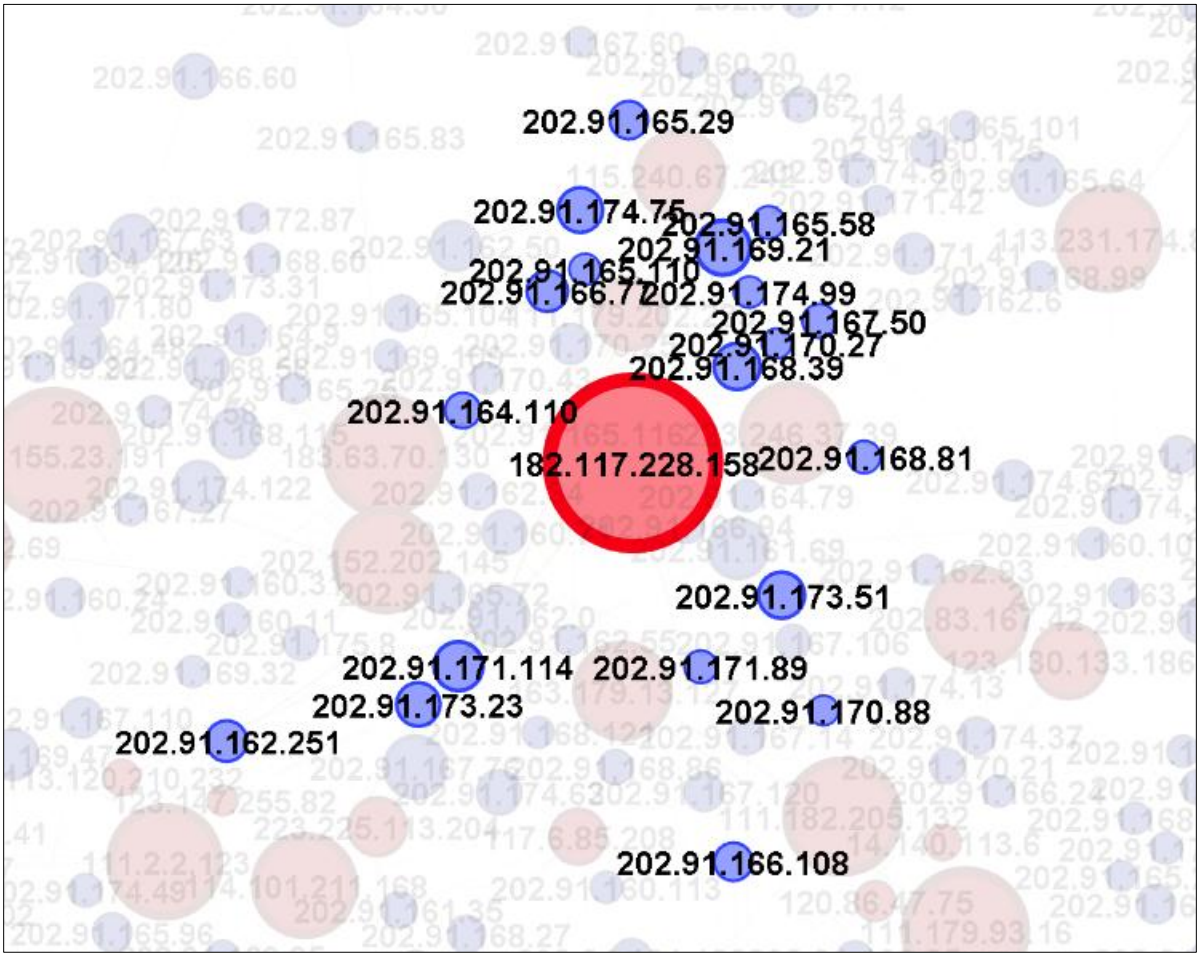
2. Largest attacks come from an Indian IP address.
   For this I looked at the mid-lower section of the graph and zoom in to take a closer look at the largest RED circle node in the graph:



The IP belongs to SwiftMail in India and it is apparently targeting a lot of Nodes.



3. The second largest Red Circled Node is 182.117.228.158 (Chinese IP)



And so with the rest of the largest attackers:

Data Table

Nodes | Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More action

Nodes	..	..	Type	Occurrences Count	In-Degree	Out-Degree	Degree
202.91.87.36	so	20	SourceIP	299	0	282	282
10.21.21.9	so	10	SourceIP	372	0	134	134
182.117.228.158	so	18	SourceIP	116	0	116	116
202.94.147.114	so	20	SourceIP	106	0	87	87
182.155.23.191	so	18	SourceIP	87	0	86	86
111.179.93.16	so	11	SourceIP	81	0	81	81
10.21.21.13	so	10	SourceIP	289	0	77	77
183.63.70.130	so	18	SourceIP	78	0	77	77
111.182.205.132	so	11	SourceIP	75	0	75	75
111.2.2.123	so	11	SourceIP	74	0	73	73
113.231.174.87	so	11	SourceIP	67	0	67	67
202.152.202.145	so	20	SourceIP	67	0	67	67
114.101.211.168	so	11	SourceIP	66	0	66	66
202.83.167.42	so	20	SourceIP	69	0	65	65
223.246.37.39	so	22	SourceIP	66	0	64	64
163.179.13.127	so	16	SourceIP	62	0	62	62
192.168.2.89	so	19	SourceIP	123	0	61	61
10.21.13.19	so	10	SourceIP	130	0	60	60
118.135.153.38	so	11	SourceIP	60	0	60	60
115.240.67.242	so	11	SourceIP	57	0	57	57

So merely investigating and responding appropriately to the top 1% of the attackers, we are effectively able to significantly reduce the number of attacks. Of the total 17359 attackers that accounted for 37,242 incidents, resolving the top 1% of the attackers (173 nodes) accounted for 17% of total incidents (6510 incidents).