09-17-2015

# DATA SCIENCE – MINI-PROJECT – 1

## 1. INTRODUCTION

Network traffic between nodes generally gives a lot of insight on the behavior between the network nodes. The flow of request packets usually follows a certain pattern. Each protocol specific request packet serves as a means for transferring some information to other nodes in the networks and keep the network alive (making sure nodes are reachable and active). Brief introduction on various protocols seen in the dataset.

A. ICMP

i. Expanded as Internet Control Message Protocol
ii. Used by nodes to determine the status of the destination node i.e. could the destination be reached or not
iii. Typical control messages are
    1. *Requesting a ping*
    2. *Replying with a ping*
    3. *Destination host unreachable*

B. DNS

i. Expanded as Domain Name System
ii. Responsible for resolving domain name to IP address, understandable by other nodes

C. ARP

i. Expanded as Address Resolution Protocol
ii. Used for converting network address to physical address
iii. Typically works by broadcasting to all nodes and wait for response

D. NTP

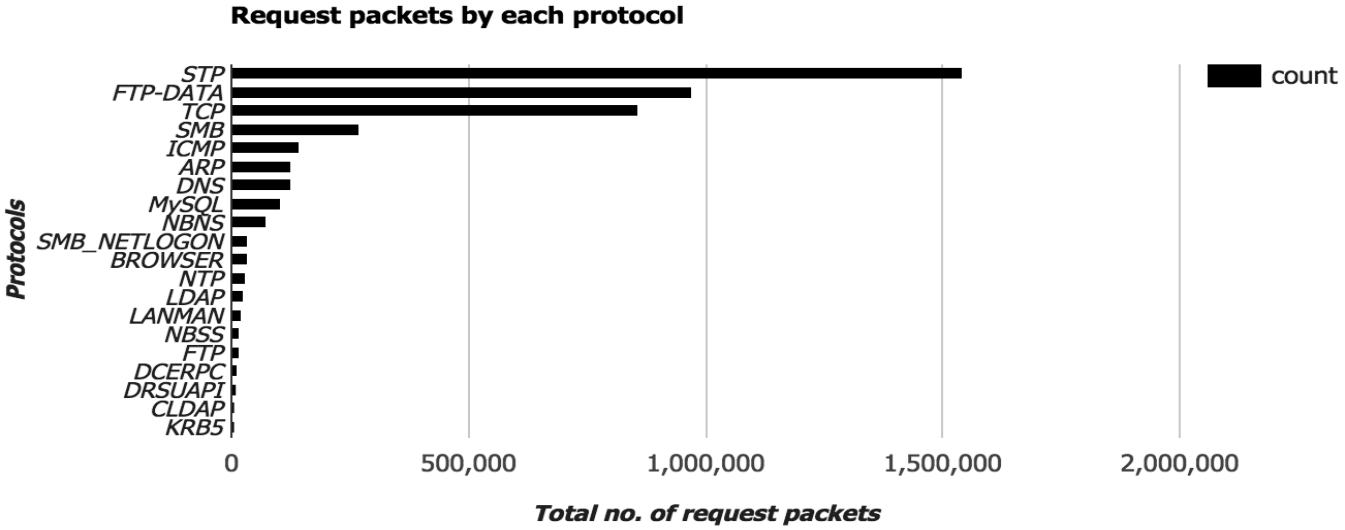
i. Expanded as Network Time Protocol
ii. Responsible for synchronizing clock between nodes in the network

## 2. PROBLEM/APPROACH

Approach here is to find some anomalies or patterns with interaction between nodes on different protocols. Start with different protocols seen in the dataset and observe patterns or abnormalities in top few protocols. Also, finding the driving node for heavy traffic in each protocol.

By parsing through the info columns of each protocol's packet transfer, few observations were made. Info column data was cleaned using Python to determine useful information.

## 3. RESULTS




Over the entire period of time, the above figure represents the number of packets sent/received by the nodes grouped by protocol used. Looks like STP packets are the highest, which is used to keep track of paths to the nodes in the network. Leaving the top 4 protocols for obviously they are most widely used.

A. ICMP

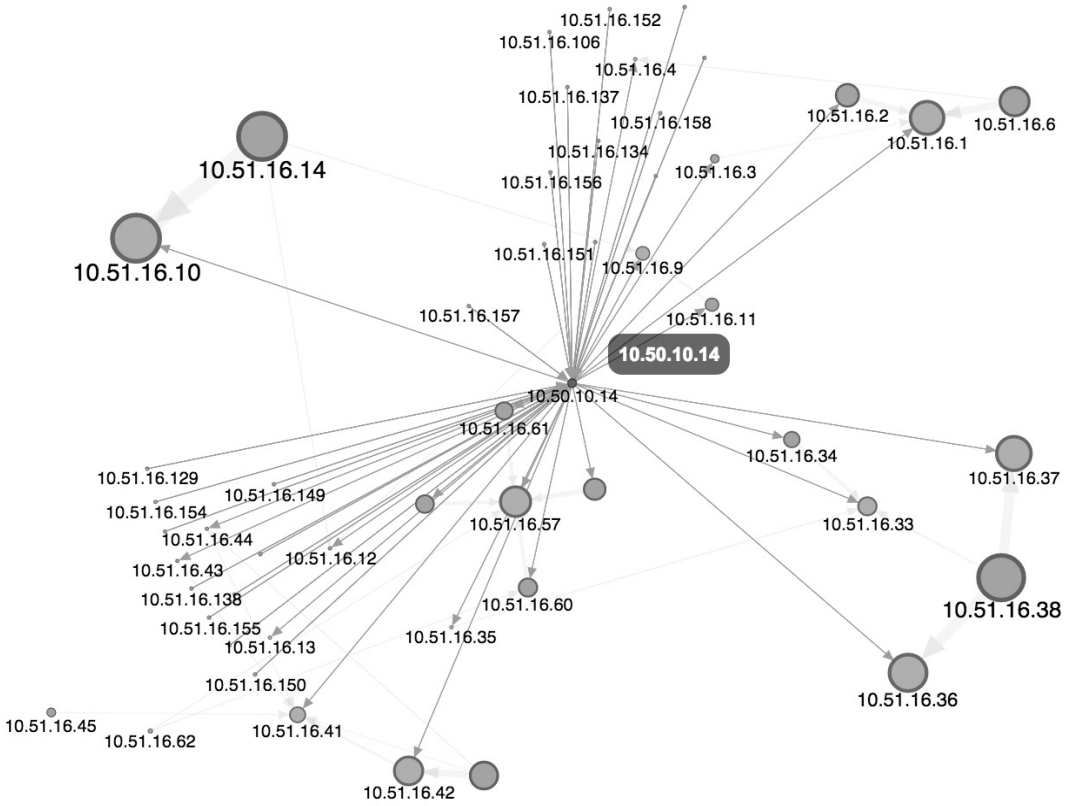



*Figure 1. Lots of traffic to and from a node whose IP is different*

*a. Observation 1*

In Figure 1., it clearly shows that one particular node is driving most the network traffic for ICMP protocol and the IP address of that particular node is different from the other nodes. It can be inferred that this node is extremely important for the network to be healthy but on the other hand, it could be a bogus node which is attracting traffic from other nodes for no useful purpose.

*b. Observation 2*

There are lots of packet transfers between nodes 10.51.16.14 and 10.51.16.10, upon further investigating, 10.51.16.10 node is unreachable for most of the time. But later on it starts sending out packets to node 10.50.10.14 for a very short period time, which is odd.

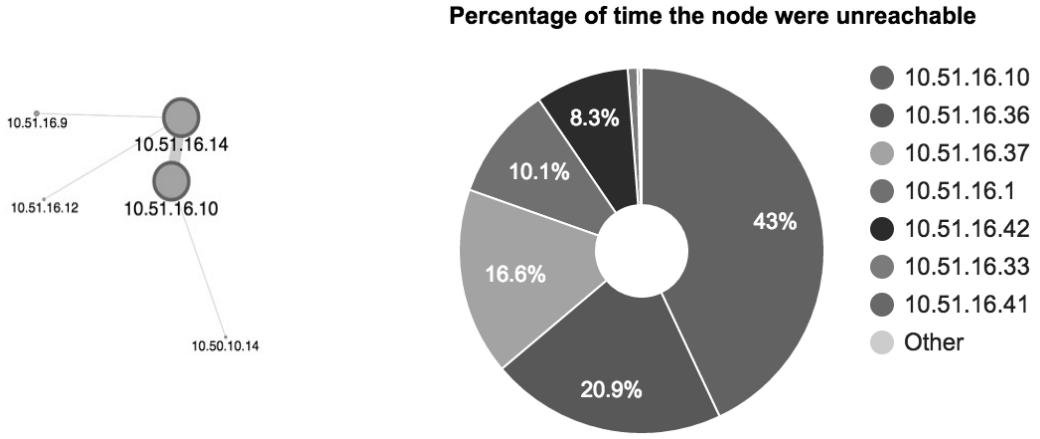



*Figure 2. Percentage of time the nodes were down*

B. DNS

*a. Observation 1*

In Figure 2., there is a huge traffic to the node 10.50.5.100 and all of them are incoming traffic It looks like the node may be down, but that information is not communicated back to other nodes. These traffic packet to this node are completely useless and it clogs the network.

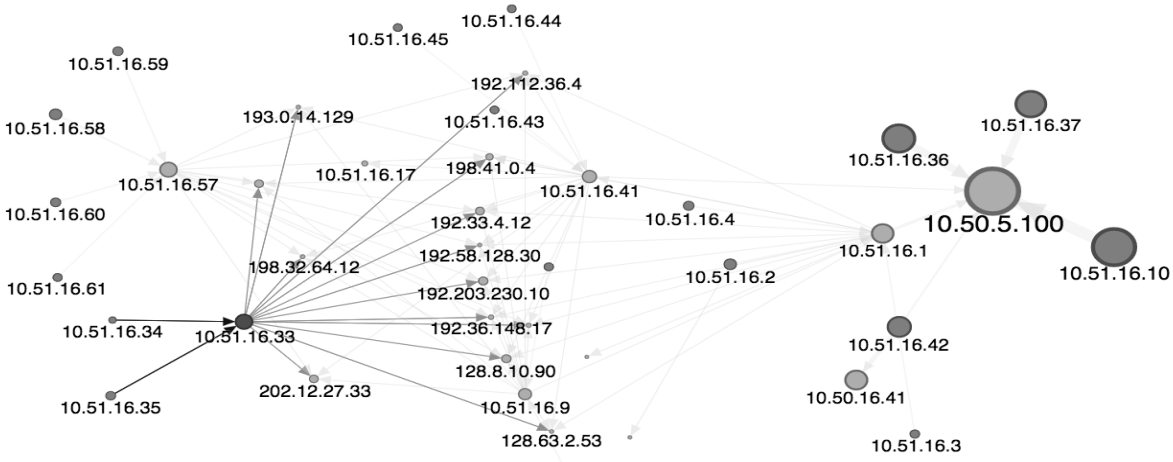



*Figure 3. DNS Protocol Network Graph*

*b. Observation 2*

There is a general pattern of packet transfers where the nodes whose last number is in their 50s range sets their destination to a node in whose last number is close to the range. But there are again some anomalies to this, where nodes like 36, 37 communicate to 10.50.5.100 directly.

C. ARP

*a. Observation 1*

In Figure 3., by looking at the traffic pattern, we can clearly identify that Vmware_42:c2:3a is completely isolated and only Broadcasting to all the nodes unnecessarily asking information about nodes (10.51.16.180 && 10.51.16.33) not part of the network. This could be a faulty hardware or someone is clogging the network with useless information to bring down the network.

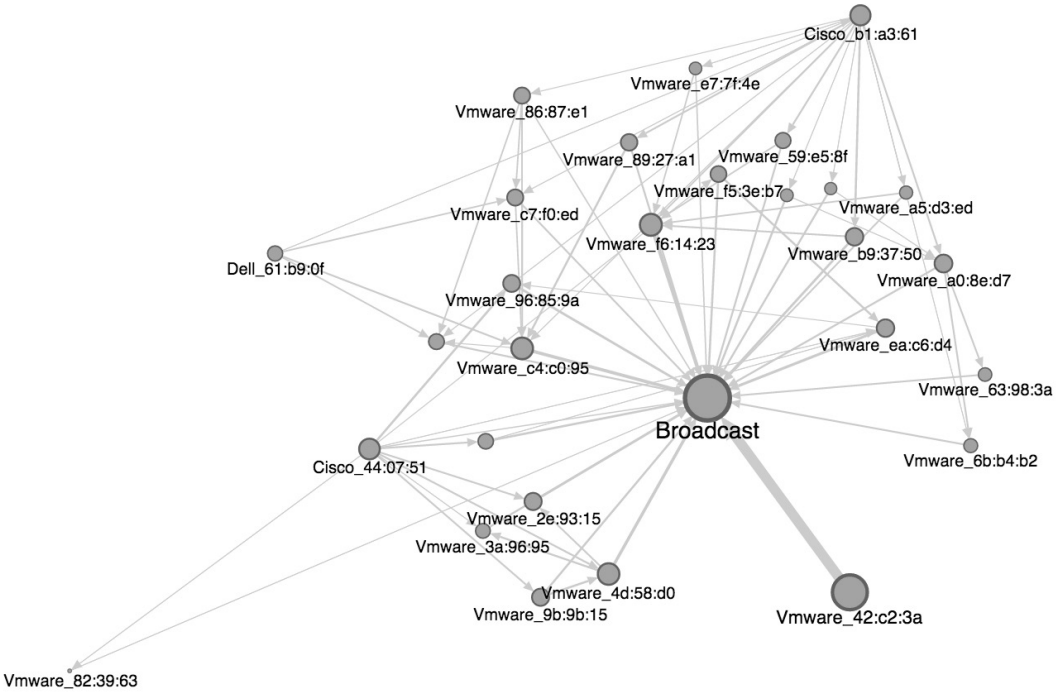



*Figure 4. ARP Protocol Network Graph*

D. NTP

*a. Observation 1*

Form Figure 4., nodes 10.51.16.122 are serving most of the requests from other nodes both directly and indirectly. Taking down that node by any means will create the network to collapse NTP protocol is responsible for synchronizing time between other nodes

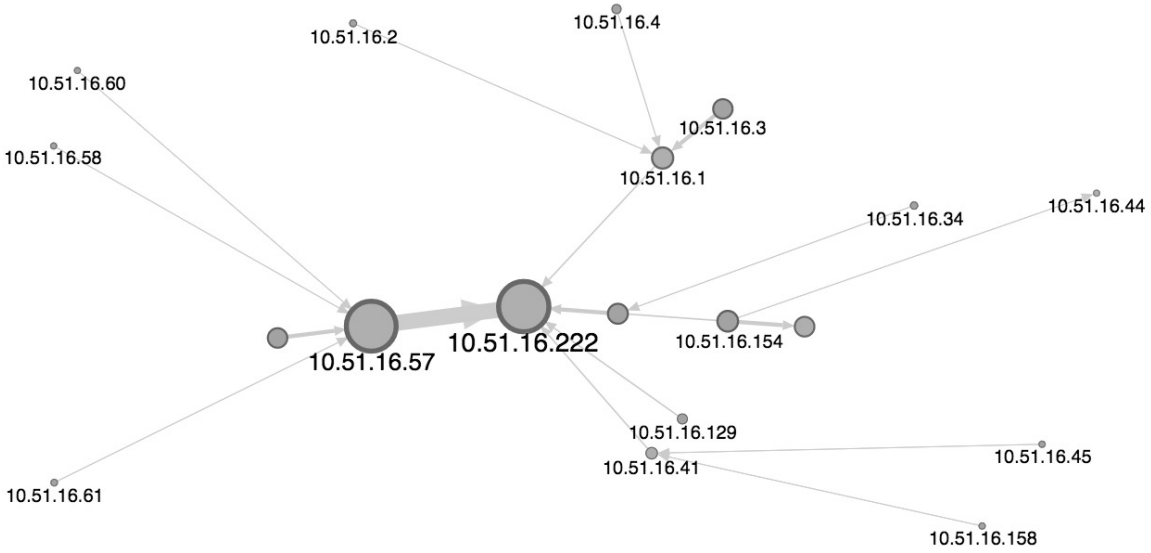



*Figure 5. NTP Protocol Network Graph*

## 4. REFERENCES

i. Google Fusion Tables for network graph and charts
ii. PostgreSQL server