Community detection based on DNS querying patterns

Aruna Chakkirala
Solutions Architect

Infoblox

Bangalore, India
achakkir@gmail.com , achakkirala@infoblox.com

***Abstract* — DNS servers capture all the domain querying activity initiated by various source IP addresses. A single user querying a domain has no meaning beyond the purpose it serves. A collective of such calls from the user may reveal something about the user while an analysis of multiple users and their querying patterns can provide useful insights. This data is represented as a weighted undirected graph between the source IP addresses. The nodes of the graphs have edges if and only if they have queried for a common domain. The edge weight is another important factor which quantifies the number of mutually queried domains and hence the affinity between the two nodes. The available data from the DNS logs was transformed into a weighted undirected graph and fed into a community detection algorithm. The community detection was conducted using the Louvain modularity premise, the output looks very promising as it provides a grouping of nodes based on the input graph. This leads to an understanding of similarly behaving source IP addresses.**

Keywords: DNS, querying pattern, community detection, modularity

I. INTRODUCTION

Applications generate voluminous log files and aim to capture every action encountered. These log files are generally verbose and require domain knowledge to create interpretations. Transforming this large volume of log information into meaning insights implies mining the data and applying appropriate algorithms.

Similar to any application, DNS also creates log files for every query it processes. Each log line captures the identity of a source which is raising a query for a particular domain. In response it expects an appropriate resolver response which is most often an IP address. Considering the large volume of internet traffic that a typical organization handles, these log files tend to grow very quickly. But the logs contain the vital combination of source IP along with its queried domain which could indicate a behavioral aspect of the particular source. Analyzing a large dataset which constitutes of many source IP addresses which have individually queried on domains can further reveal patterns of similarly behaving source IP addresses.

For instance, all members of a team would likely be querying for domains which host a commonly used service in the team. The team members along with the rest of organization will also query for public domains like google.com or yahoo.com. While the querying of public domains creates a relationship between most of the members, the stronger bond is signified by the commonly queried domains within the team. This clusters determined from similarly querying domains serve of interest here as they constitute a similarity in behavior within the cluster.

II. METHOD

A. A typical DNS query log line is of the following format

01-Jan-2018 15:01:21.997 client 172.205.42.215#51341: query: eng-lab.beta.datasta.com IN A + (172.205.3.10)

B. Many such log lines exist and contain the vital information pair of the source IP and the queried domain.

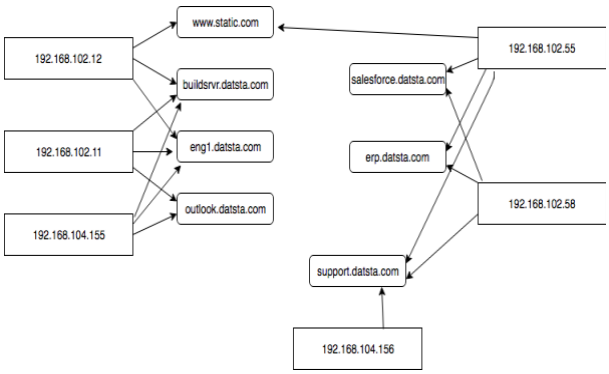
A sample dataset is created by extracting the source IP address, queried domain and timestamp. This dataset is further mined to pull out all the source IP addresses which have queried for a particular domain.

C. The extracted dataset is best represented as a graph.

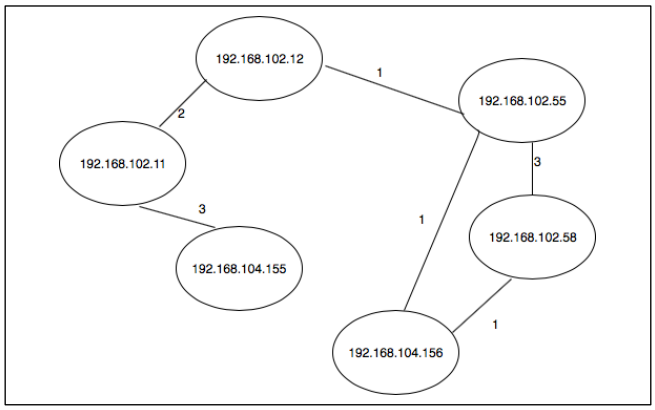
Each of the source IP addresses form a node in the graph and each node has an edge with its neighbor provided they have exhibited some similar behavior. The similar behavior in this context is the queried domains and identifying all source IP addresses which have queried for the same domain. Plotting the graph with the nodes and adding edges when there are mutually queried domains results in an almost complete graph. The graph edges also have weights computed from the number of mutually queried domains. Thus the log data is represented as ann undirected weighted graph.

D. Visualization of DNS Querying pattern

Each source IP address queries a set of domains and any overlaps are clearly visible in the figure below.



E. The user querying scenario above indicates that all the source IP addresses which belong to the same group, have some similarities in their querying patterns. The similarities and relationship between source IP addresses can be represented as an undirected weighted graph. Graph nodes with mutually queried domains have an edge of weight one. And when there is more than one mutually queried domain, the edge weight is higher. Th figure below illustrates the resulting graph.



III. COMMUNITY DETECTION

Many systems of scientific interest can be represented as networks, sets of nodes or vertices joined in pairs by lines or edges[1]. One issue that has received a considerable amount of attention is the detection and characterization of community structure in networks[3,4], meaning the appearance of densely connected groups of vertices, with only sparser connections between groups[1].

A promising approach consists in decomposing the networks into sub-units or communities, which are sets of highly inter-connected nodes. The identification of these communities is of crucial importance as they may help to uncover a-priori unknown functional modules such as topics in information networks or cyber-communities in social networks. Moreover, the resulting meta-network, whose nodes are the communities, may then be used to visualize the original network structure [2].

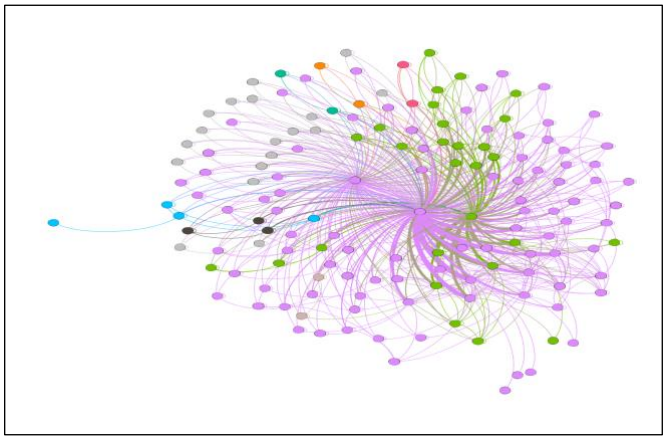
The problem of community detection requires the partition of a network into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected. Precise formulations of this optimization problem are known to be computationally intractable. Several algorithms have therefore been proposed to find reasonably good partitions in a reasonably fast way. This search for fast algorithms has attracted much interest in recent years due to the increasing availability of large network data sets and the impact of networks on every day life. One can distinguish several types of community detection algorithms: divisive algorithms detect inter-community links and remove them from the network [5, 6, 7], agglomerative algorithms merge similar nodes/communities recursively [8] and optimization methods are based on the maximisation of an objective function [9, 10,11]. The quality of the partitions resulting from these methods is often measured by the so-called modularity of the partition. The modularity of a partition is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities [5, 12]. In the case of weighted networks (weighted networks are networks that have weights on their links, such as the number of communications between two mobile phone users), it is defined as [13]

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j), \quad (1)$$

where A_{ij} represents the weight of the edge between i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i, c_i is the community to which vertex i is assigned, the δ-function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and $m = \frac{1}{2}\sum_{ij} A_{ij}$.

IV. COMMUNITY DETECTION FINDINGS

The resulting graph created from the sample DNS logs had 154 nodes and 751 edges. Only edges within a minimum weight of 2 were considered to eliminate nodes with only a single common queried domain. The community detection and visualization was conducted in Gephi using its inbuilt modularity algorithm. The graph created from the DNS logs served as the input dataset for Gephi and the modularity algorithm was applied on the dataset. The Gephi run resulted in 17 detected communities through the Louvain modularity algorithm. From this list of detected communities, those communities with a membership of less than 1.3% was given the color grey. The rest of the communities were given distinctive colors other than grey to provide a visual representation of the identified communities as indicated in the figure below.



V. CONCLUSION

The results revealed detected communities from the input data of a weighted undirected graph. Each of the detected communities consisted of source IP addresses which exhibited similar behaviors and hence belonged to either a team in the organization or were bound together by similar activities.

ACKNOWLEDGMENT

The author would like to thank Roger Barlow, Senior Product Manager at Infoblox for identifying and visualizing the need to create DNS log data insights. The author would like to thank Peter Rizk, Sr Director, Technical Marketing at Infoblox for providing the means to conduct this analysis. The author is also highly grateful to Infoblox for the opportunity provided.

REFERENCES

[1] Newman MEJ (2006b) Modularity and community structure in networks. Proc Natl Acad Sci 103(23): 8577–8582.

[2] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre - Fast unfolding of communities in large networks (2008) arXiv:0803.0476v2 [physics.soc-ph].

[3] Newman M. E. J. (2004) *Eur. Phys. J. B* **38**:321–330.

[4] Danon L. , Duch J. , Diaz-Guilera A. , Arenas A. (2005) *J. Stat. Mech*, P09008.

[5] Girvan M and Newman M E J, 2002 Proc. Natl. Acad. Sci. USA 99 7821

[6] Newman M E J and Girvan M, 2004 Phys. Rev. E 69 026113.

[7] Radicchi F, Castellano C, Cecconi F, Loreto V and Parisi D, 2004 Proc. Natl. Acad. Sci. USA 101 2658.

[8] Pons P and Latapy M, 2006 Journal of Graph Algorithms and Applications 10 191.

[9] [Clauset A, Newman M E J and Moore C, 2004 Phys. Rev. E 70 066111.

[10] Wu F and Huberman B A, 2004 Eur. Phys. J. B 38 331.

[11] Newman M E J, 2006 Phys. Rev. E 74 036104.

[12] Newman M E J, 2006 Proc. Natl. Acad. Sci. USA 103 8577

[13] Newman M E J, 2004 Phys. Rev. E 70 056131.