Knowledge Discovery of Port Scans from Darknet
Sofiane Lagraa, Jerome Francois

To cite this version:
Sofiane Lagraa, Jerome Francois. Knowledge Discovery of Port Scans from Darknet. IFIP/IEEE Symposium on Integrated Network and Service Management (IM) - AnNet workshop, May 2017, Lisbonne, Portugal. hal-01636215

HAL Id: hal-01636215
https://hal.archives-ouvertes.fr/hal-01636215
Submitted on 16 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—Port scanning is widely used in Internet prior for attacks in order to identify accessible and potentially vulnerable hosts. In this work, we propose an approach that allows to discover port scanning behavior patterns and group properties of port scans. This approach is based on graph modelling and graph mining. It provides to security analysts relevant information of what services are jointly targeted, and the relationship of the scanned ports. This is helpful to assess the skills and strategy of the attacker. We applied our method to data collected from a large darknet data, i.e. a full /20 network where no machines or services are or have been hosted to study scanning activities.

I. INTRODUCTION

Computers connected to a network use many services by mainly relying TCP/UDP protocols. Port or IP scanning (also known as sweeping) is one of the most common techniques by attackers to discover open ports in preamble of an attacker or an intrusion through those ones. Hence, scanning methods are part of network-based discovery techniques, still for emerging threats like Advanced Persistent Threats [5]. Therefore, an in-depth understanding of scanning techniques is necessary for improving security: detection, prevention or forensics. There are three main types of scans: vertical, horizontal and block scans. Vertical scan is described as a single IP being tested on multiple ports. Horizontal scan is described as trying scan against a group of IPs for a single port. Block scan is a combination of both of them. All of them cannot.

This paper aims at making a deeper and more sophisticated analysis of vertical scans by seeking the relationship of commonly scanned TCP ports. Our approach relies on building dependencies among those latter and extract the predominant roles of certain ports within sequence of consecutive targeted ports as well as extracting relationships among them, i.e. discover groups of commonly scanned ports. This can help the security analysts to guard against attacks and improve prevention and detection tools. For example, given the scans performed on the following successive ports 80, 591, 8008, 8080 and 443, we want to discover the relationship of commonly scanned ports with their dependencies, then the significant service ecosystem is that these ports are used for "HTTP" traffic. Thus, the analysts can update the web server security. We have validated the method on real data collected from a darknet in order to observe and understand (1) the behavior of different scans and (2) extract the "modi operandi" of targeted ports.

In this paper, our contributions are as follows:
1) a graph-based model of port scans relationships,
2) a knowledge discovery methodology using graph mining techniques based on the proposed model,
3) an application of our methodology on real darknet data.

The rest of the paper is organized as follows. Section II briefly presents the background on darknet and related work of analysis techniques of ports scans. Section III provides an overview of dataset and highlights research questions. Section IV and Section V present the first and the second part of our approach concerning the port-scan graph model and port-scan graph analytic, respectively. Finally, Section VI draws conclusion and discusses the future work.

II. BACKGROUND AND RELATED WORK

A vertical port scanner sends a message to each port of predefined list and waits for a certain replies. For example, a common manner (TCP SYN-scan) to discover TCP open ports is to try to initiate new connections with SYN-flagged packets (TCP SYN scan). If the port is open, a SYN/ACK packet is sent back while a RST packet is representative of a closed port.

From darknet data we have collected, we conduct the study over all TCP-SYN packets which represent 93.79% of TCP packets. The darknet (network telescopes) refers to the servers configured to trap adversaries and collect suspicious data. Since these servers run without interacting with no legitimate users (attackers) in passive mode and correspond to unused hosts or devices, any observed traffic destined to them raises suspicion and hence necessitates investigation. They do not offer Internet services, nor does it use any. Darknet has been used in the past to extract different insights on probes or scanning activities [3], [10], bots, DDoS (Distributed Denial-of-Service) attacks due to victims reply (backscatter) packets to spoofed IP addresses [14]. For instance, DDoS attacks and scanning activities can be detected using the existing techniques and tools such as IDS Suricata¹, topological analysis technique [7], and statistical techniques [9]. However, these tools and techniques provide scanning profiles of ports such as the number of countries are responsible for the majority of large scans, the number of performed scans, and the targeted services.

Darknets are also known as darkspace, blackhole monitors, unused IP addresses, or network telescopes. To harmonize the terminology, we use the word darknet throughout this paper. A darknet is a whole subnetwork, which is announced over Internet such that packets sent to the IP addresses are properly routed over. This subnetwork is defined by the prefix length. It does not host any services and so no legitimate traffic is

¹https://suricata-ids.org/
supposed to reach it. The entity hosting the darknet is then
discreetly collecting all incoming packets, i.e. without replying
or notifying any of them. They have been proved to be appropriate
monitoring techniques for security, e.g. to observe botnets, scans
or distributed denial-of-service attacks [3], [10], [1]. Although
there can be primarily seen as a naïve collector, especially
when compared to more advanced Internet security monitoring
approaches like honeypots, they have been demonstrated to be
complementary [11]. In [1], the authors discovered patterns of
darknet traffic and how it changes over time.

In [8], the authors presented the measurement and analysis of
a 12-day world-wide cyber scanning campaign targeting
VoIP (SIP) servers using a darknet. Similarly, in [9], behavior
of horizontal scanning scans have been researched on. In [18], the authors presented an animated 3-D scatter plot
visualization of port scanning on darknet data. Finally, in [7],
a topological data analysis technique is able to analyze and
visualize a large number of IP packets in order to make malicious activities patterns easily observable by security
analysts. Among these patterns, they found scanning activities
observed by a darknet.

Compared to previous approaches, it is worth to mention
that our work focuses on vertical scan. Besides, most of current
studies either rely on evaluating trends over targeted ports or
behavior of a scan towards a single port. Our approach tends
to automatically discover dependencies among targeted ports
rather than deducing those relations thanks to observation like
in [8]. In that sense it is closer to work related about automated
discovery of network service dependencies [17]. However, the
goal of the latter is to discover normal dependencies that
exist in benign applications whereas ours establish relations
in attacker behaviors when performing TCP scans. To achieve
that, we provide behavior pattern of port-scan through graph
mining methods which provide intrinsic analysis of scanned
ports where the classical statistical tools do not allow to
discover.

III. DATASET OVERVIEW AND RESEARCH QUESTIONS

Darknet data used in this paper are collected from a /20
darknet network (i.e., 4096 addresses) for 2 years (Nov.2014
~ Nov.2016). During this period 2,884,539,435 packets were
captured representing 500 GB of data with an increasing trend
over the months as shown in Figure 1. As highlighted in
Table I, the number of TCP packets represents more than 78%
of packets. This is a challenging task to analyze port-scan
strategies from a huge amount of traffics.

Since, there are 65,536 possible TCP ports, we segment the
scanned port numbers on three ranges according to Internet
Assigned Numbers Authority (IANA): system or well-known
ports (0-1023), user or registered ports/or vendors use for
applications (1024-49151), and the dynamic and/or private
or ephemeral ports (49152-65535). The difference uses of
these ranges is described in RFC6335 [6]. Figure 2 shows the
segmentation of the different scanned Port numbers during the
24-months. The system ports are frequently targeted compared
to the user ports and dynamic ports. Indeed they contain
the majority of well used standardized services, especially
web services (HTTP and HTTPS) which are now a de facto
standard for majority of applications. However, as attackers
may also target specific and specialized ports, we conduct
the study over all 2,125,080,102 TCP-SYN packets which
represent 93,79% of TCP packets. Indeed, through priori
analysis with the Suricata IDS

2[https://suricata-ids.org/]

and our own tool [7], we
discovered that most of TCP-SYN packets are related to port
scanning even if other types of traffic like DDoS are also
present. Discarding those few ones is risky as they are being
mixed with scans. Keeping their related packets for analysis
has zero impact since they are targeting a unique TCP port
although our approach aims to find correlation over multiple
TCP ports.

In this paper, we focus on the last four months from 08-
2016 to 11-2016 which represent a higher number of TCP-
SYN packets and more significant results. Table II represents
the top-10 of probed ports in each month. We see that the
services relating to traffic, servers and database are most
targeted services in scans. However, these statistical results
do provide the relationships between the probed ports.

Our aim is to discover, analyze the relationship groups of
ports probed together, and so understand the attacker scanning
strategies, especially by answering the following questions:

- Does it exist particular ports which are frequently probed
  in vertical scanning ?
- Does it target successive ports of the same segment or
  different ports of different segments ?
- Are there a strongly relationship between set of ports?
  Do they belong to the same targeted ecosystem services
  such as set of messaging or database services ?

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Number of packets</th>
<th>Weight (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>2,265,756,934</td>
<td>78.54%</td>
</tr>
<tr>
<td>UDP</td>
<td>352,449,519</td>
<td>12.21%</td>
</tr>
<tr>
<td>ICMP</td>
<td>33,235,987</td>
<td>1.15%</td>
</tr>
</tbody>
</table>

TABLE I: Number of packets per protocol

Fig. 1: Number of packets according to the months
In order to answer these questions, we propose a graph-based modelling approach for scanned ports through the time.

IV. PORT-SCAN GRAPH MODEL

Different steps are needed for transforming TCP-SYN packets into a graph describing TCP ports and their relationships from a scanning perspective. Source and destination IP addresses and ports have been widely used in the intrusion detection domain [13], [2]. They are good indicators of attack traffic flows and targeted services. In our context, we are interested in the scanning process to successive ports. Relying on a graph representation will allow to condense information and thus to model global knowledge from multiple instances, i.e. multiple scans involving different IP addresses.

A. Port numbers transformation and representation

Definition 1 (Port sequence): Let P be the set of all TCP ports. Assuming, a source and destination IP address, respectively IP_{src} and IP_{dst}, we denote $S = ((IP_{src}, IP_{dst}), T_s, T_e)$, the sequence of targeted ports by $IP_{src}$ to $IP_{dst}$ between the starting time $T_s$ and the ending time $T_e$, where $T_s < T_e$. $S$ is thus a list of TCP destination ports ordered by time: $S = < (p_1, t_1), \ldots, (p_n, t_n), \ldots >$, where $p_i \in P$, $T_s \geq t_i \leq T_e$ such that $t_{i+1} > t_i$.

Figure 3 shows an example of a port sequence with 5 targeted ports $(25, 26, 27, 443, 8080)$ at time 50, 60, 70, 87 and 98 respectively. In our particular case of the darknet, no replies are sent back and no service are running. As a result, the sequences of ports thus only represents relations among attackers attempts rather than dependencies among multiple ports used by the applications [17].

B. Constructing a port-scan graph

In practice, the length of port sequences can be very long, and can reach up to a million of TCP port instances over one month. In addition, the ports can be redundant in a sequence in most of the time. To characterize causal relations between two successive ports in a port sequence, we introduce the notion of the port-scan graph model as an intuitive graph representation for successive scans in all port sequences.

A port-scan graph is constructed from multiple successive scans. In other way, it is a directed graph that represents successive relationships between targeted ports in port sequences. Specifically, each vertex represents a port number $p_i$ and each edge $(p_i, p_j)$ indicates that a scan after to port $p_j$ occurs after a scan to the port $p_i$, where $p_i \neq p_j$. $p_i \neq p_j$ means that we neglect successive probes to identical port. As introduced in section III, the advantage of such an approach is automatically discard DDoS attacks from the analysis which are the main type of mixed traffic with scans in darknet data. Formerly, port-scan graph over a set of type of ports $V$ is a labelled directed graph $G = (V, E, \beta)$:

- $V$ is the set of observed port numbers from from $P$.
- $E$ is a set of edges in $G$. Let $p_u$ and $p_v$ be be two port numbers in $V$. There is an edge $(p_u, p_v) \in E$ if and only if there exists a dependency rule $p_u \xrightarrow{t_{uv}} p_v$.
- \( \beta \) is a function that assigns for each edge $(p_u, p_v)$ the number of dependency occurrence $I_{p_u,p_v}$.

The represented associations between all port numbers represent the signature features of all port sequences, i.e. the most and common behavior of IP source addresses targeting IP destination addresses.

TABLE II: Top-10 of scan probes

<table>
<thead>
<tr>
<th>Month-Year</th>
<th>Port</th>
<th>Number</th>
<th>Month-Year</th>
<th>Port</th>
<th>Number</th>
<th>Month-Year</th>
<th>Port</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>08-2016</td>
<td>telnet</td>
<td>106,891,476</td>
<td>09-2016</td>
<td>telnet</td>
<td>122,793,736</td>
<td>10-2016</td>
<td>telnet</td>
<td>183,601,219</td>
</tr>
<tr>
<td></td>
<td>ms-sql-s</td>
<td>7,788,167</td>
<td></td>
<td>3d-nsd</td>
<td>9,615,076</td>
<td></td>
<td>3d-nsd</td>
<td>19,755,823</td>
</tr>
<tr>
<td></td>
<td>ssh</td>
<td>3,921,299</td>
<td></td>
<td>microsoft-ds</td>
<td>8,285,083</td>
<td></td>
<td>microsoft-ds</td>
<td>7,975,819</td>
</tr>
<tr>
<td></td>
<td>ms-wbt-server</td>
<td>2,241,857</td>
<td></td>
<td>ms-sql-s</td>
<td>7,885,440</td>
<td></td>
<td>ms-sql-s</td>
<td>7,096,882</td>
</tr>
<tr>
<td></td>
<td>http</td>
<td>2,123,664</td>
<td></td>
<td>ssh</td>
<td>3,148,455</td>
<td></td>
<td>ssh</td>
<td>3,589,657</td>
</tr>
<tr>
<td></td>
<td>http-alt</td>
<td>1,257,436</td>
<td></td>
<td>ms-wbt-server</td>
<td>2,376,006</td>
<td></td>
<td>ms-wbt-server</td>
<td>2,479,072</td>
</tr>
<tr>
<td></td>
<td>https</td>
<td>1,238,662</td>
<td></td>
<td>http</td>
<td>1,729,630</td>
<td></td>
<td>http</td>
<td>2,019,860</td>
</tr>
<tr>
<td></td>
<td>mysql</td>
<td>1,199,221</td>
<td></td>
<td>http-alt</td>
<td>1,052,039</td>
<td></td>
<td>http-alt</td>
<td>1,121,846</td>
</tr>
<tr>
<td></td>
<td>dsrv-port</td>
<td>847,182</td>
<td></td>
<td>https</td>
<td>1,035,851</td>
<td></td>
<td>https</td>
<td>873,097</td>
</tr>
<tr>
<td></td>
<td>smtp</td>
<td>464,722</td>
<td></td>
<td>mysql</td>
<td>950,356</td>
<td></td>
<td>mysql</td>
<td>1,058,366</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>http-alt</td>
<td>1,087,097</td>
<td></td>
<td>http-alt</td>
<td>1,111,872</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>https</td>
<td>1,090,220</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2: Scanned Port numbers segmented on three ranges: System Ports (0-1023), User Ports (1024-49151), Dynamic and/or Private Ports (49152-65535)

Fig. 3: Port sequence
V. PORT-SCAN GRAPH ANALYTICS

Graph algorithms or graph analytic are analytic tools used to assess structural properties of the graph and, as a final objective, to determine strength and direction of relationships between objects represented as vertex in a graph. For example, in a graph representing relationships (such as "interacting" or "probing" another port or IP address) between machines, graph analytic can to extract the following knowledge:

- the centrality of a node (TCP port), i.e. its predominant presence in shortest paths between other nodes (shortest scanning sequences)
- the clusters of ports being well connected in the graph, i.e. detecting dense partitions.

Such knowledge allows to answer the different questions raised in section III.

To evaluate how relations among targeted ports evolved in time, our analysis is monthly-based. Hence, 24 unique graphs have been created from the two years of collected data. The built graphs contains of up to 6284 vertices and 589117 edges and very low density.

A. Centrality measures

Centrality score measure the communication importance of a vertex in terms of how central it is.

In this paper, we study two popular types of centrality: degree (baseline), and betweenness (flow-based).

1) Degree centrality: The degree centrality is the simplest and most popular centrality measure and gives a highly local view of the graph around each node.

Indeed it is the number of neighbors of a node. In a directed graph, each vertex has an indegree and an outdegree. Let \( G = (V, E) \) and \( \forall v \in V \). Indegree of vertex \( v \) is the number of edges which are coming into the vertex \( v \). The indegree of \( v \) is denoted \( \text{deg}^{-}(v) \). Outdegree of vertex \( v \) is the number of edges which are going out from the vertex \( v \). The outdegree of \( v \) is denoted \( \text{deg}^{+}(v) \).

2) Betweenness measure: The betweenness centrality measures are flow-related by considering the information flowing through edges. The most well-known centrality in this group is the Freeman’s betweenness centrality [12]. It measures how much a given vertex lies in the weighted shortest paths of other vertices. Let \( \delta_{st} = \delta_{ts} \) denote the number of shortest paths from \( s \in V \) to \( t \in V \), where by convention \( \delta_{ss} = 1 \). Let \( \delta_{st}(v) \) denote the number of shortest paths from \( s \) to \( t \) that some \( v \in V \) lies on.

\[
\text{betweenness} = \sum_{s \neq v \neq t} \frac{\delta_{st}(v)}{\delta_{st}} \tag{1}
\]

The betweenness centrality allows us to measure the probability that a sequence from the port \( s \) to the port \( t \) goes through the port \( v \), \( s \) and \( t \) have the route of the ports sequences. Betweenness centrality is a measure of the influence or importance of a vertex over the flow of scanned ports between every pair scanned ports under the assumption that information primarily flows over the shortest path between them.

3) Results: We list top 10 port-scan patterns in Table III based on both centrality measures. The betweenness provides us the ports commonly used during the scanning, but not necessary most frequent ports, which thus differs our work from the usual study. We discover, for a months, there exists at least not-assigned targeted ports.

We discovered that the degree centrality and betweenness highlight some period of times where probes targets certain kind of services (close in their use):

- In 08-2016: the centrality concerns different services and protocols such as \textit{http-alt: 8008} (HTTP Alternate), \textit{irdmi: 8000} (Intel Remote Desktop Management Interface) or \textit{memcache: 11211} (Memory cache service).
- In 09-2016: the centrality is around the computer network authentication protocols and distributed control systems such as \textit{Kerberos: 88} and \textit{d-s-n: 8086} (Distributed SCADA Networking), respectively.
- In 10-2016, the same as the last month, with additional printing protocol.
- In 11-2016, the centrality is around the transfer protocols such as \textit{http: 80}, \textit{https: 443}, \textit{http-alt: 591}, \textit{telnet: 23} and servers protocols.

There exists a large flow to network services, database service, computer network authentication protocols, and transfer protocol, in 08-2016, 09-2016, 10-2016, and 11-2016, respectively.

B. Community analytics measures

Community analytic algorithms aims to find dense subgraphs (also called clusters or communities) in a graph, in which its vertices are more connected within the cluster than with the nodes outside the cluster.

1) Modularity: Modularity measures the density of the partition of a graph into subgraphs called \textit{modules}. It measures the density of links inside communities as compared to links between communities [15]. The modularity measures essentially compares the number of links inside a given module with the expected value for a randomized graph of the same size and same degree sequence.

Figure 4 shows the densely connected clusters of vertices, with sparser connections between clusters. The ability to detect such clusters of vertices is significantly important to discover the dense group of the common targeted ports at the same time whereas centrality measures are focused on individual ports. Groups within scanned ports graphs might correspond to consecutive probes communities, or service communities.

For computing modularity class, we use the algorithm developed in [4] for community structure discovery in large graphs.

The modularity of a partition of a graph [16] can be written as

\[
Q = \sum_{m=1}^{M} \frac{l_{m}}{L} - \left( \frac{d_{m}}{2L} \right)^{2} \tag{2}
\]

where the sum is over the \( M \) modules of the partition, \( l_{m} \) is the number of links inside module \( m \), \( L \) is the total number of...
of links in the graph, and $d_m$ is the total degree of the nodes in module $m$.

The principle of the algorithm is based on distinct community divisions process which is represented as follows:

1) Separate each vertex solely into $m$ community.
2) Calculate the increase of $Q$ for all possible community pairs.
3) Merge the greatest increase in $Q$.
4) Repeat 2 & 3 until the modularity $Q$ reaches the maximal value.

2) Results: Figure 5 represents the top-10 discovered clusters between 08/2016 and 11/2016. In each figure, the clusters are sorted in descending order of cluster size. The first line represents the number of targeted ports in different segments within the clusters. The second line represents the number of targeted ports which are assigned to specific services or not. Example: the port number 22 is assigned port number, which corresponds to the Secure Shell (SSH). In contrast, the port number 26 is an unassigned port number. In addition to extract strongly connected targeted ports, these graphs highlight the homogeneity or heterogeneity of clusters regarding the types of ports such as: user and private ports.

TABLE III: Top-10 of services discovered with the betweenness and degree measures of port-scan graph

<table>
<thead>
<tr>
<th>08-2016</th>
<th>09-2016</th>
<th>10-2016</th>
<th>11-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betweenness</td>
<td>Degree</td>
<td>Betweenness</td>
<td>Degree</td>
</tr>
<tr>
<td>ms-wbt-server</td>
<td>http-alt</td>
<td>ibdmi</td>
<td>pscsynchttp</td>
</tr>
<tr>
<td>tungsten-https</td>
<td>http-alt</td>
<td>kerberos</td>
<td>sxfer</td>
</tr>
</tbody>
</table>

Fig. 4: Clusters of vertices

ONCLUSION AND FUTURE WORK

VI. C

clusters based on the two differentiation criteria. This supposes different underlying scanning strategies. In addition, through a manual inspection of clusters, we particularly discovered two types of strategy when targeted unassigned ports: either in a random or an incremental probing of TCP ports.

Furthermore, we analyze the heterogeneous and homogeneous clusters containing assigned ports. As clusters containing not-assigned ports, it also exists consecutive probes of services such as: user and private ports.

We discover other important clusters which contain non-consecutive probes and not randomly probes. These graph clusters are completely connected and contain particular scanned ports:

- A cluster containing both two types of services: database service ports and medical service ports. Database service ports such as mysql: 3306, redis: 6379, ms-sql-s: 1443 (Microsoft-SQL-Server), radg: 6789 (GSS-API for the Oracle), ttc-ssl: 2484 (Oracle TTC SSL). Medical service ports such as ohsc: 18186 (Occupational Health SC), and biimenu: 18000 (Beckman Instruments, Inc). This cluster is shown in Figure 6.
- A cluster containing closest messaging transfer services such as: 103: Genesis Point-to-Point TransNet, 20480: emWave Message Service, 23: telnet, and 2323: 3d-nfsl. Genesis Point-to-Point TransNet: sometimes used with MS Exchange X.400 mail messaging traffic. 3d-nfsl is being used as an alternate telnet port.

Those observation clearly indicates, first, that our method is able to find proper correlation (based on the semantics of the ports) and, second, that attackers may leverage intelligent scanning to attack a dedicated type of systems.

In this paper, we have proposed a new approach based on graph analytic techniques for grouping the scanned ports. We provide a method to help the security analysts to understand what clusters of services are commonly being targeted, the importance and semantic of clusters over the scanned ports, and different strategies of the port scanning. Our experimental results, over real data collected in a darknet, highlight the
ability of our method to discover unknown specific co-targeted ports belonging to the same or different types of services. Our future plan consists on performing a semantic distance measure over TCP ports and predict the future scanned ports.

ACKNOWLEDGMENTS

This work was partially funded by HuMa, a project funded by Bpifrance and Region Lorraine under the FUI 19 framework. It is also supported by the High Security Lab hosted at Inria Nancy Grand Est (http://www.lhs.loria.fr).

REFERENCES


