An Enhanced View of Incidence Functions for Applying Graph Theory to Modeling Network Intrusions

CHUCK EASTTOM, MO ADDA,
School of Computing
University of Portsmouth
Portsmouth, Hampshire, UK

Abstract: - Graph theory provides a robust tool for modeling a diverse range of subjects. It has been widely applied to computer networks and even network attacks. However, the incidence function in graph theory is often given a cursory treatment. This current research involves applying a range of information theory equations to describe the incidence function in a graph of a computer network. This improves modeling of computer network attacks and intrusions. Specifically attacks that involve substantial changes in network traffic can be more accurately modeled, if the incidence function of the graph is expanded.

Key-Words: - Graph theory, Shannon entropy, Rényi entropy, incidence functions, information theory


1 Introduction

Graph theory is widely used to model a variety of activities. It has been a core part of engineering and network modeling for many years [1] [2] [3]. Graph theory has also been applied to such disparate topics as neurological degenerative diseases [4] and analysis of stocks [5].

Various aspects of graph theory, including algebraic graph theory have been explored for modeling a diverse range of subjects [6]. Graph theory has been utilized with linear algebra to understand systems based on the graphs incidence matrix [7] [8] and degree matrix [9].

Many researchers have found that understanding the Laplacian of a graph is also informative for understanding the system being modeled. The Laplacian matrix is usually defined as the degree matrix minus the adjacency matrix [11]. The Laplacian matrix is also referred to as the admittance matrix, Kirchhoff matrix, or discrete Laplacian. The Laplacian is sometimes normalized as shown in equation 1.

\[
\mathcal{L}_{ij}(G) = \begin{cases} 
1 & \text{if } i = j \text{ and } d_j \neq 0 \\
-\frac{1}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent} \\
0 & \text{otherwise}
\end{cases}
\]

Calculating the graphs Laplacian leads to other information from algebraic graph theory. As one example, Kirchhoff’s theorem states that the determinant of the Laplacian of a graph indicates number of spanning trees in a graph [12]. A spanning tree T of an undirected graph G is a subgraph that is a tree which includes all of the vertices of G, with minimum possible number of edges [13]. By first calculating the graphs Laplacian, additional data immediately follows, such as using the Laplacians determinant to gain information about spanning trees in that graph. This is one example of the breadth and depth of research in applying algebraic graph theory to various modelling problems. The existing body of graph theory research provides a rich set of mathematical tools for analysing many phenomena, including network intrusions. This current work does not refute the existing body of research, but rather seeks to add to it.

What is missing from the literature is an exploration of the incidence function. An incidence function is normally defined as a function that maps every edge to a pair of vertices, either ordered or unordered [14] [15]. While other aspects of graph theory have an extensive corpus of research exploring them, incidence functions have sparse research. This is particularly interesting when using graph theory for modelling network traffic, or computer network intrusions. Merely having two or more vertices connected by an incidence function is interesting but knowing the specific nature of that connection would be more useful. Thus, understanding the incidence function is quite important.

In network traffic, it is not accurate to simply state two network nodes, or vertices, are connected. Understanding the nature of that connection is
essential. For example, it can be useful to understand the nature of the data flowing between the two vertices, as well as the amount of data. Therefore, expanding the incidence function to include more rigorous descriptions would improve the ability to model networks and network attacks utilizing graph theory.

The current literature and research does not adequately explore incidence functions. Incidence functions are mentioned as simply the connection between two vertices. The specific nature of the connection is not explored. This current study explores the actual function that connects two or more edges.

2 Problem Formulation
There is a gap in the literature regarding the exploration of incidence functions as they relate to specific modeling problems with graph theory. While other aspects of graph theory have an extensive amount of research, incidence functions is an overlooked area. A better understanding of incidence functions will lead to more information being derived from modeling real-world systems with graph theory.

There is a need to expand the examination of the incidence function. A better understanding of the function that connects the vertices, will improve the accuracy and efficacy of graph theory modeling of systems. The particular focus in this current study is the application of graph theory to modeling computer network intrusions and attacks. There is existing research on modeling networks and network intrusions with graph theory [16] [17]. However, those attempts have not expanded the examination of the incidence function. This current study will expand the examination of the incidence function when applying graph theory to modeling network intrusions.

3 Problem Solution
Given the current study is directed towards application of graph theory to computer intrusions it is an important first step to understand the nature of network traffic. Ultimately, such traffic is information. Whether that data is transmitted via radio waves, electrical signal in unshielded twisted pair cable, or some other media, the traffic is information. Even malicious traffic is information. Therefore, a logical place to begin expanding the incidence function would be to integrate information theory into incidence functions. The solution posited in this paper is to include the concepts of information theory to describe the incidence function between vertices in a computer network graph, in general, emphasizing in particular on intrusions.

There have been some attempts to incorporate small portions of information theory into narrowly defined specific applications of computer network traffic analysis [18] [19]. However, these attempts have been very narrowly focused and do not include graph theory. The goal of this work is to expand upon existing research that utilizes graph theory to understand computer networks, by incorporating a spectrum of information theory techniques into graph theory. The information theory functions serve as the incidence functions for the graph of the intrusion. This will provide a more robust tool for understanding computer networks that currently exists in the literature.

The current study posits that the incidence function is one or more information theory equations. Rather than simply state that the incidence function connects two or more vertices via their edges or arcs, the connection between vertices is a measurement of the information flow between the vertices. The vertices are connected via information. That information flow is quantified either via the amount of information (Shannon Entropy) or the diversity of information (various diversity indices).

3.1 Shannon Information
A logical place to begin is with integrating Shannon information into graph theory. This metric was first proposed by Claude Shannon and was intended to quantify the information content present in a particular text string. The value is referred to as information entropy. Essentially, the informational content of a message is that message’s entropy $p_i$, denotes the proportion of symbols which belong to the to the $i$th class of character in the string being examined [20]. information entropy is expressed as shown in equation 2.

$$H' = - \sum_{i=1}^{n} p_i * \ln(p_i) \quad (2)$$

This value can be readily integrated into incidence functions. An incidence function maps edges to pairs of vertices. In the case of a computer network, the vertices would be nodes on the network. What flows between them is information. Quantifying how much information is flowing between to vertices provides a better understanding of the network. This is true for analyzing network...
traffic or for analyzing intrusions. A Denial of Service attack, as one example, would yield a sudden and substantial increase in information flowing between nodes. This fact will be explored later in this paper with a specific case study and a laboratory experiment.

Understanding the change in flow of information would provide a better understanding of the attack. The volume of traffic varies with different DoS attacks. Thus, the Shannon Information or Shannon Entropy can provide insight into the nature of the DoS attack.

3.2 Simpson Index

The Simpson index was introduced in 1949 by Edward H. Simpson to measure the degree of concentration of individuals of a specific category. When this is applied in economics, it is referred to as Herfindahl index or the Herfindahl–Hirschman index. The Simpson index equals the probability that two entities taken at random from the dataset of interest represent the same class or type of entities [21], as shown in equation 3.

$$\lambda = \sum_{i=1}^{R} p_i^2,$$  \hspace{5cm} (3)

In equation 3, $R$ is the total number of types or classes in the dataset. This is referred to as the richness of the dataset. This equation is also equal to the weighted arithmetic mean of the proportional abundances $p_i$ of the types of interest, with the proportional abundances themselves being used as the weights. One could use either the Simpson index, or the Shannon entropy separately to understand what connects two vertices in a computer network. However, it may be advantageous to consider both metrics. This would make the incidence function between the two vertices, a system of two equations.

In the case of a Denial of Service attack, the information flow should increase dramatically, but the diversity of the data should reduce. This is due to many DoS attacks being predicated on a flood of a specific type of packet [22] [23].

Thus, when graphing a network, a sudden increase in both Shannon entropy and a decrease in the Simpson index would be indicative of a DoS attack. This is only one example. A large data exfiltration would have similar characteristics but would be a) less volume and b) fewer vertices (network nodes) involved. Berba’s study in 2019 clearly demonstrated the substantial spike in traffic from a single IP and of a single protocol that indicated data exfiltration [24].

3.3 Shannon-Weaver Index

A metric related to Simpson index is the Shannon-Weaver index. This metric quantifies the concentration of individuals that are classified into groups [25]. The formula is shown in equation 4.

$$H = - \sum_{i=1}^{S} p_i \ln p_i$$  \hspace{5cm} (4)

S is the quantity of individual species found in the community being studied. The $p_i$ is still the proportion of symbols that are part of the $i$th category of symbol in the text string of interest, similar to Shannon entropy formulation.

There are several variants of the Simpson index. Two examples of such variations are the Berger–Parker index, the Inverse Simpson index, the Dominance index, and the Gini-Simpson index. All of these variations are directed towards studying how one particular class of items compare to the entirety of items. As with the Simpson index, when applied to computer networks, this can be indicative of DoS attacks, data exfiltration, or other types of network attacks [26], [27], [28].

An example of how DoS attacks affect network traffic is found in 2018. On November 13, 2018 several internet service providers in Cambodia were attacked with a large-scale distributed Denial of Service attack. The average attack size per IP prefix was 2.48Gbps [29]. In a worst-case scenario, an attack of this size spread across 38 IP prefixes is potent enough to overwhelm a 10Gbps ISP line. The attack was a UDP (User Datagram Packet) flood on random ports [30]. The attack targeted more than 80 subnets for six hours. At its peak the bandwidth utilized was 150 Gbps [31].

In 2017 the mean bandwidth in Cambodia was 13 mbps [32]. Figure 1 demonstrates the actual impact on bandwidth of this attack.
The top portion of figure 1 shows packet latency and the bottom shows traceroutes. During the actual attack, latency increased substantially and the number of traceroutes that did not time out decreased substantially. These are tell-tale signs of a Denial of Service attack.

The bandwidth utilization for the Cambodia Network Exchange fluctuates between 2 Gbps and 20 Gbps, throughout the day [33]. Getting precise data for Cambodia internet usage is simply not possible as such data is not published.

Clearly the percentage of UDP traffic dramatically increased during the attack, thus altering the Shannon Entropy. Furthermore, rather than the normal distribution of protocols in network traffic, the traffic during the attack was overwhelmingly UDP which would also affect all of the diversity indices. A specific laboratory experiment demonstrating the details of these effects is provided later in this paper.

3.5 Hartley Entropy

Hartley entropy was introduced by Ralph Hartley in 1928, and thus predates Shannon’s work on information theory. This is often simply called the Hartley function. If a sample from a finite set $A$ uniformly at random is picked, the information revealed after the outcome is known is given by the Hartley function, shown in equation 5 [34].

$$H_0(A) := \log_b |A| \quad (5)$$

In equation 4, $|A|$ denotes the cardinality of A. If the base of the logarithm is 2, then the unit of uncertainty is now referred to as the Shannon. If it is the natural logarithm, then the unit is the nat. Hartley used a base-ten logarithm, and with this base, the unit of information is called the Hartley in his honour.

Like the Shannon-Weaver and Simpson indices, the Hartley entropy provides insight regarding the diversity in a given dataset. Substantial and sudden changes in the diversity of network traffic can be indicative of several different network attacks including DoS attacks and data exfiltration.

3.6 Rényi entropy

Related to some of the previous entropy values discussed so far is the Rényi entropy. The Rényi entropy generalizes the Hartley entropy, the Shannon entropy, the collision entropy and the min-entropy [35], as shown in equation 6.

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left( \sum_{i=1}^{n} p_i^\alpha \right) \quad (6)$$

In equation 5, $X$ is a discrete random variable with possible outcomes $(1, 2, ..., n)$ and corresponding probabilities $p_i$ for $i = 1$ to $n$. Since this metric is generally used in information theory, the logarithm is typically base 2, and the order, $\alpha$, is $0 < \alpha < 1$.

The Rényi entropy is a measure of the diversity in a dataset. The Rényi entropy is also important in quantum information, where it can be used as a measure of entanglement. As with other diversity metrics, this can be useful in fully understanding the data flow between two or more network nodes.

3.7 Synthesizing the metrics

The goal of incorporating information theory with graph theory is to provide a more robust modeling tool. In the current study the focus is on applications in computer networks including computer network intrusions. However, this process can be applied to any system that involves information.

The issue addressed in this current study is to expand on what is currently considered regarding incidence functions. All too often the incidence function is simply discussed as a connecting two vertices via an edge or arc [36], [37], [38]. It is commonplace in the literature to provide no further details on the incidence function. Expanding the definition of the incidence function to include the information flowing between two vertices adds detail to graphs that model any information related network.

When considering two network nodes, the connection between them is an information flow. Examining that flow is critical to fully understanding the network itself. The methods
posited in this section are not mutually exclusive. Rather, they are a set of tools that can be utilized as needed in the specific situation.

3.8 A Case Study

Detailed specifics are not available on most actual DoS attacks. Published case studies generally provide information regarding peak bandwidth utilization, but not details such as the distribution of protocols in network traffic, or a granular view of the changes in network traffic [39], [40], [41], [42]. Therefore, a laboratory network was established with typical network services and simulated user traffic, then that network was subjected to a typical Denial of Service attack. By using a laboratory environment, all the details of the traffic could be captured, and the equations posited in this study could be fully tested.

The lab was setup with a Linux server running Ubuntu 19.4 and common services [43], [44], [45], [46]. There were four client virtual machines setup. Two where running Windows 10 and 2 running Ubuntu 19.4. The clients were used for typical network activities such as visiting web sites, DNS lookups, and email. Traffic was measured with Wireshark. The arithmetic mean of traffic by protocol is shown in table 1.

<table>
<thead>
<tr>
<th>Port</th>
<th>Protocol</th>
<th>Mean % of Traffic</th>
<th>TCP/UDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>20/21</td>
<td>FTP</td>
<td>4%</td>
<td>TCP</td>
</tr>
<tr>
<td>53</td>
<td>DNS</td>
<td>32%</td>
<td>UDP</td>
</tr>
<tr>
<td>80</td>
<td>HTTP</td>
<td>16.5%</td>
<td>TCP</td>
</tr>
<tr>
<td>443</td>
<td>HTTPS</td>
<td>30.3%</td>
<td>TCP</td>
</tr>
<tr>
<td>465</td>
<td>SMTPS</td>
<td>0.5%</td>
<td>TCP</td>
</tr>
<tr>
<td>995</td>
<td>POP3S</td>
<td>9.5%</td>
<td>TCP</td>
</tr>
</tbody>
</table>

Table 1 Network Protocols

Table 1 describes the specific protocols that were in operation on the test network. The percentage of traffic used by each protocol is provided. The mean total bandwidth was 2.5 mbps. After establishing a lab environment and running typical network traffic, the next step was to evaluate the content of this traffic utilizing the equations posited in this study. The results of the calculations are shown in table 2.

<table>
<thead>
<tr>
<th>Equation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shannon Entropy</td>
<td>0.322</td>
</tr>
<tr>
<td>Simpson Index</td>
<td>.765</td>
</tr>
<tr>
<td>Shannon Weaver Index</td>
<td>1.57</td>
</tr>
<tr>
<td>Rényi entropy</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

Table 2 Values before Attack

For the Rényi entropy a $\alpha$ value of .32 was utilized. This is due to the fact that as $\alpha$ approaches zero, the Rényi entropy increasingly weighs all possible events more equally. That is not appropriate with network traffic. Therefore, a $\alpha$ value was used equal to the largest probability for a single packet type. Once these statistics were calculated the next step was to execute a Denial of Service attack on the network. To accomplish this a UDP flood was executed on the network using a bandwidth of 10 mbps. This altered the distribution of protocols. The distribution during attack is shown in Table 3.

Table 3 Traffic Flow During Attack

<table>
<thead>
<tr>
<th>Port</th>
<th>Protocol</th>
<th>Mean % of Traffic</th>
<th>TCP/UDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>20/21</td>
<td>FTP</td>
<td>0.5%</td>
<td>TCP</td>
</tr>
<tr>
<td>53</td>
<td>DNS</td>
<td>5%</td>
<td>UDP</td>
</tr>
<tr>
<td>80</td>
<td>HTTP</td>
<td>2.5%</td>
<td>TCP</td>
</tr>
<tr>
<td>443</td>
<td>HTTPS</td>
<td>4.5%</td>
<td>TCP</td>
</tr>
<tr>
<td>465</td>
<td>SMTPS</td>
<td>1%</td>
<td>TCP</td>
</tr>
<tr>
<td>995</td>
<td>POP3S</td>
<td>1.5%</td>
<td>TCP</td>
</tr>
<tr>
<td>80</td>
<td>UDP FLOOD</td>
<td>85.5%</td>
<td>UDP</td>
</tr>
</tbody>
</table>

The change in traffic flow also changed the entropy and diversity calculations. The new values are shown in table 4.

Table 4 Values During Attack

Given that during the attack the probability of the traffic being UDP flood was dramatically increased an $\alpha$ value equal to the probability of UDP flood traffic was used, .855. This is also consistent with the values used for the pre-attack calculations.

The results of the experiment demonstrate that the Shannon Entropy increased by 64.55%. The Simpson Index decreased by 63.31%. The Shannon Weaver Index decreased by 58.6%. The Rényi entropy decreased by 350%. The information content increased significantly, while all diversity indices decreased significantly. This indicates that using these equations provides an understanding of a flood-based Denial of Service attack.

4 Conclusion

This paper provides an outline to apply information theory to more fully understanding the incidence functions in graph theory. As graph theory is

Chuck Easttom, Mo Adda
routinely used in a wide range of modeling applications, this work can enhance such applications. Equations from information theory quantify the connection between vertices, and thus are the incidence functions.

A range of information theory metrics including the Simpson index, Shannon entropy, and Rényi entropy were proposed as possible incidence functions when applying graph theory to computer networks. It must also be noted that these metrics can be used in combination to understand not only the flow of information, but also the diversity in that flow.

This work can be expanded by including rates of change for both information and information diversity. Partial differential equations are already applied with thermodynamic entropy. Shannon himself created differential entropy but was incorrect in his equation. Limiting density of discrete points is a technique created by Edwin Jaynes to correct issues with Claude Shannon's differential entropy formula. It would be of interest to explore utilizing this function, or some similar equation, for the incidence function in graph theory.

In addition to expanding the incidence function, other aspects of algebraic graph theory should be explored for applications in analyzing network traffic. Spectral graph theory and fractional graph theory are obvious areas to explore. Specialized matrices such as Jacobian matrices are also areas that should be explored in future studies.

While the current study is directed to studying computer networks, information theory can be integrated into graph theory in other modeling contexts. For example, consider the previously mentioned neurological applications of graph theory. Viewing signals between neurons as information would provide additional data on the neurological structures being studied. In the case of degenerative neurological disorders, one would expect to see a reduction in the information flow.

References:


