International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7009(P)
Issue 02, Volume 04 (February 2017) www.ijiris.com ISSN: 2349-7017(O)

Utilizing Graph Theory to Model Forensic Examination

Chuck Easttom
Collin College Professional Development
chuck@chuckeasttom.com

Abstract *— one of the areas of discrete mathematics is graph theory. From a pure mathematics viewpoint, graph theory studies the pairwise relationships between objects. Those objects are vertices. Graph theory is frequently applied to analysing relationships between objects. It is a natural extension of graph theory to apply that mathematical tool to the evaluation of forensic evidence. In fact the literature reveals several, limited, forensic applications of graph theory. The current paper describes a more broad based application of graph theory to the problem of evaluation relationships in forensic investigation. The process takes standard graph theory and identifies entities in the investigation as vertices with the connections between the various entities as edges. Those entities can be suspects, victims, computer system, or any entity relevant to the investigation. Regardless of the nature of the entity, all entities are represented as vertices, and the relationship between them is represented as edges connecting the vertices. This allows the mathematical modelling of the events in question and facilitates analysis of the data.*

Keywords *— Graph Theory, Mathematical Modelling, Digital Forensics, Forensics*

I. INTRODUCTION

The purpose of this paper is to fully describe how graph theory can be applied to analyzing evidence in a digital forensics investigation. The goal is to use the well-established principles of graph theory to mathematically model the elements in a computer based crime. This paper is an expansion of my paper Applying Graph Theory to Evidence Evaluation (Easttom 2016). This paper significantly expands the mathematics introduced in the earlier paper. The review of literature is also expanded. While the principles in this paper could be applied to any forensic investigation, the primary focus of this paper is digital forensics. Therefore, the investigations contemplated involve cyber-crimes. Graph theory has been applied to several limited forensic applications, and it has been used in non-forensic contexts, that are nevertheless related to computer systems.

II. REVIEW OF LITERATURE

The functionality of the RSA algorithm is based on aspects of number theory involving prime Graph theory is a robust tool for examining relationships between any set of objects. The essentials of graph theory are relatively easy to grasp. This this paper will begin with a review of the essential elements of graph theory, with particular attention to those elements that will later be applied to forensic evidence. Graph theory provides a methodology to mathematically examine objects and the relationship between those objects. The first step is to define precisely what a graph is. Put formally: A finite graph G (V, E) is a pair (V, E), where V is a finite set and E is a binary relation on V (Deo, 2016). Now let us examine that definition in more reader-friendly terms. A graph begins with a set of nodes which are referred to as vertices. The specific objects that these vertices represent are irrelevant to the mathematics of graph theory. The connections between vertices are called edges (Chartrand, 1985).

These edges are ordered pairs and not necessarily symmetrical (Balakrishnan, 2010). This means that the connection between vertices may be directional (Clark & Holton, 1991). The directionality of an edge indicates that the connection is coming from one vertex to another. Any edge that connects to a vertex is said to be incident to that edge. If the edge proceeds from a vertex, it is described as being incident from that vertex (Balakrishnan, 2010). If the connection proceeds to that vertex, then it is described as being incident to that vertex. In pure mathematical terms, the incidence is simply showing directionality, the nature of that directionality is irrelevant. When a graph is directional it is referred to as a digraph and the edges are referred to as arcs (Chartrand, 1985). For the purpose of modelling digital investigations, digraphs will normally be used. When a graph as no multiple edges or loops, then it is said to be a simple graph.

If two vertices are connected via more than one edge, then this graph is considered a multiple graph or multigraph (Bollobás, 2013). For the purposes of evaluating forensic evidence, many graphs will be both digraphs and multigraphs. The edges and vertices are described based on the connections. For example, the degree of a vertex is simply the number of edges incident to (connected to) that vertex. An edge or an arc is said to be incident to a vertex, if it connects to that vertex. A vertex can connect to itself, forming a loop (Bondy & Murty, 2008). If a digraph's edges have a specific cost or weight associated with each edge, then the graph is considered to be a weighted digraph (Trudeau, 1994). For the purposes of evaluating relationships between objects, weighted digraphs are very effective.. This allows the graph to represent that some relationship is initiated by one vertex and is directed to another vertex, and the significance or weight of that relationship.

In the current context we can consider elements in a case to be objects or vertices. For example, a target web server in a cyber breach would be one vertex. The suspect would be another. Various sites from which the attack(s) were launched would be still other vertices. Then any evidence showing a connection between two vertices, would be an edge. This would lead to a graph similar to what is shown in figure 1.

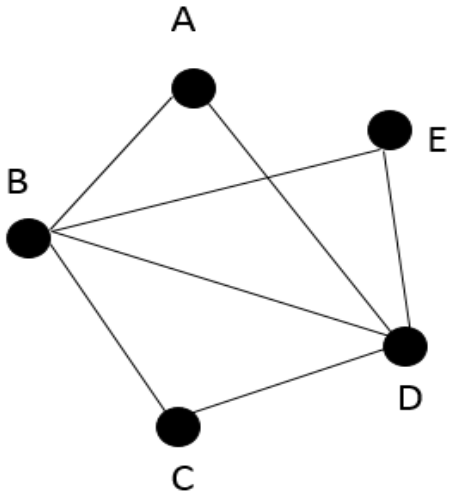


Figure 1: A basic graph

A more robust model would be accomplished with a digraph using arcs. This is shown in figure 2. Note that figure 2 also demonstrates a loop, on vertex D.

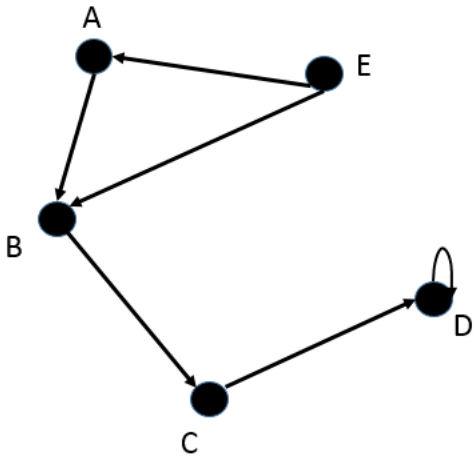


Figure 2: A basic digraph

Graph theory has been previously applied to evaluating network traffic (Wang, 2010; Ahlswede, Cai, Li, & Yeung, 2000; Holme, 2003; Amaral and Ottino, 2004). In that context, graph theory was used to evaluate network traffic. Wang's approach in his 2010 dissertation was to utilize graph theory to categorize and aggregate network evidence in order to present a cohesive and comprehensible map of the network traffic. This approach can be very useful in describing network evidence. Given that networks consist of a set of nodes that are connected, it is natural to represent the nodes as vertices, and the connections as edges.

In the case of network traffic, the nodes are servers, routers, switches, client computers, and other network devices. The level of traffic between two nodes (or vertices) can be represented by assigning a weight to the edge. In one study, graph theory was applied to the study of heroin seized in drug arrests (Zufferey, et.al, 2006). In this study, graph theory was used to evaluate heroin and the cutting agents used in producing heroin that is sold to consumers. The authors of the study state "An application of graph theoretic methods has been performed, in order to highlight the possible relationships between the location of seizures and co-occurrences of particular heroin cutting agents. An analysis of the co-occurrences to establish several main combinations has been done." Graph theory was used in this instance, to recognize patterns in the applications of cutting agents used in heroin.

Another approach is to combine graph theory with statistical methods to evaluate causal relationships (Peterson & Shenoi, 2011). The primary focus in this study was the combination of Bayes theorem with graph theory to provide a means to characterize causal relationships among variables. This approach could be used in evaluating various hypotheses.

Graph theory has been suggested as a methodology for the study of data in the examination of unstructured data in emails (Haggerty, Karran, Lamb, & Taylor, 2011). In their paper, the authors describe analyzing the strength of relationships in the unstructured data, via the application of graph theory. The data elements are represented as nodes, and the vertices are used to describe both the presence of a relationship as well as the strength of that relationship.Other studies have not embraced the robust application of graph theory to forensic analysis, but have frequently used graphs as a visual aid for displaying data (Zadora & Ramos, 2010; Catanese& Fiumara, 2010). The use of graphs to display data can be effective in communicating the data, but fails to utilize the modelling capability of graph theory.

III.THE METHODOLOGY

A significant challenge in collecting and categorizing digital evidence is to appropriately attribute evidence. While this can be an issue in any forensic investigation, it is a particular problem with digital forensics. It is insufficient to simply trace a given attack vector to a specific network, or even a specific computer. The investigator must be able to attribute the activity in question to a specific user (Chaski, 2005). Particularly in workplace networks, multiple parties may have potential access to a given computer. Applying graph theory to a case can aid in attribution. By creating a graph representation of the case wherein each relevant entity is a vertex and the connections between vertices are edges, the forensic analyst can then apply graph theory to evaluating the evidence. The current methodology involves applying graph theory to create a mathematical model of the investigation. This can assist in attribution, but can also provide a robust view of the entire investigation and all elements therein.

Modern graph theory can be divided into two broad sub-categories. The first involves the algebraic aspects of graph theory (Balakrishnan, 2010). In this sub-topic of graph theory the primary focus is representing the vertices and edges, along with weighting those edges. Essentially this is a descriptive application of graph theory. The second sub-category of graph theory involves optimization problems. This aspect of graph theory has been applied to network traffic analysis to optimize the path of data through a network. In this paper, the focus will be on algebraic graph theory. However, it is certainly possible to apply optimization aspects of graph theory to the analysis of forensic evidence. It is recommended that optimization aspects of graph theory be considered as future areas of research to expand upon the ideas set forth in this current paper. Graph theory can be applied to a variety of different aspects of forensic science. The current methodology is concerned with describing the evidence in question and evaluating the connections between individual evidence items, suspects, victims, and any other entities relevant to a given investigation

The issue of directionality has already been discussed; however for forensic examination how to model direction is an important issue. The mathematics of graph theory simply state that an arc can be incident from one vertex to another. However, graph theory does not indicated how one determines that the arc begins at one vertex, rather than the other. For the purpose of forensic examination direction should always be from the initiator to the target. For example, if a person visits a website, even if that person downloads files or data, the connection is from the person to the website. The issue for forensic examination is not the direction of the flow of data, but rather who initiated the flow of data.

The first step in applying graph theory to any investigation is to identify the various entities involved in the incident in question. These entities will be represented as vertices. When any two entities have any connection, that connection is represented as an edge. The edge should always be an arc that models not only the connection, but the initiator of that connection. It is possible, even likely, that some vertices will have multiple edges. For example, if the investigation involves confidential information that was stolen from company A and subsequently found in company B, one can represent either each company as a vertex, or relevant personnel as vertices. In most investigations, the more granular approach will be more effective.Continuing with the previous example, each employee at company A that had access to the data in question would need to be represented. Then any connections these employees had to company B would be represented as edges. If evidence shows that the data passed through some intermediate entity, such as a hacker external to either company or perhaps a dark web market, then each employee in either company A or company B would have any relationships to that third-party entity represented as edges.

The representation of entities and connections as a graph is a relatively simple process. The key issue will be to ensure that all entities and all of the connections are represented. As with any modelling tool, the model can only be as accurate as input allows. The next issue of interest in an investigation is to analyse the connections, represented as arcs and to weigh those arcs. How to properly weight the arcs (connections) is a key element in appropriately analysing the data. For example, how strong is the connection between employee A2 in company A and the third-party intermediary C? While it may seem preferable to weight edges based on a ratio measurement that is usually not only impossible but inaccurate. For example, a ratio measurement might be the total number of network packets sent between entity A and entity B, divided by the total number of packets both entities sent to any destination in a given time frame. This number would give you a ratio value for the traffic, but such a measurement would give no insight into the nature or value of the traffic. In such a situation, casual visits to a website would count equally with deliberate hacking of the website.

For the purposes of evaluating evidence, an ordinal or interval measurement will be the most accurate. Ordinal measurements merely require a ranking, without specific interval spacing (Gibilisco, 2004). An example ranking for a connection between a given computer user and a particular web server could be 1) casual and infrequent visits; 2) casual and frequent visits; 3) significant interaction with the web server; and 4) either administrative access to the web server or deliberate hacking of the web server. This type of ordinal ranking actually provides the investigator with a much clearer understanding of the nature of the user's interaction with the web server. An ordinal approach, with a small number of ordinals, also simplifies the weighting process. For the purposes of analysing digital evidence, ordinal approaches with 5 or fewer ordinals are recommended. To illustrate the application of graph theory to forensic investigations, let us begin with a scenario involving a simple cybercrime. In this scenario a specific company has been infected with spyware that was used to steal sensitive data. A former employee is suspected of creating and deploying the spyware. However, the investigation has also shown that at least one current employee visited a website that is known to have been infected with spyware. The suspect, the infected website, and the victim network all form vertices. The next question becomes how are these elements connected? If there is a connection between the former employee and the victim, post-employment, then that would be an edge connecting those to vertices. At this point we have a very simple graph. The suspect is vertex A, the infected website is vertex B, and the victim company is vertex C. This simple graph is shown in figure 3.

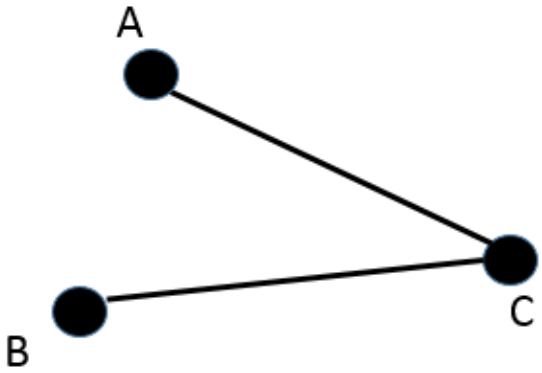


Figure 3: Graph of hypothetical crime

At this point no weighting or directionality has been included in the graph. That would be an additional step the investigator would then take. While this can be illustrated with a traditional graph, graph theory does not actually require the imagery of a graph. One can represent data in a variety of matrices. The adjacency matrix is a simply matrix that displays which vertices are adjacent. The adjacency matrix is shown below.

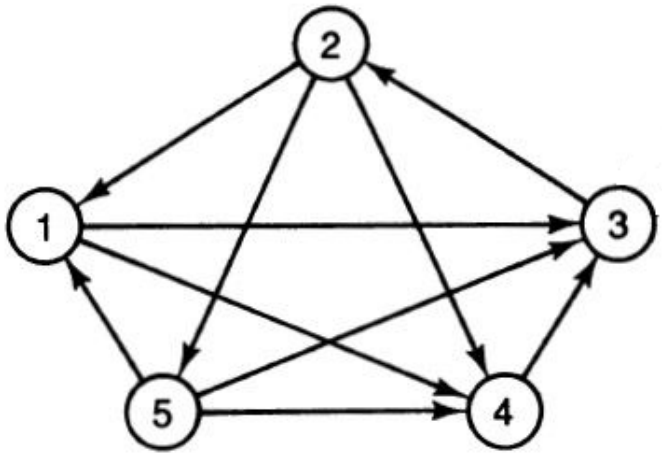
	A	B	C
A			1
B			1
C	1	1	

Adjacency Matrix

If the situation is this simple, then the investigator utilizes traditional forensics techniques such as ascertaining the dates and times when the current employee visited the website (B), and when the former employee (A) accessed the company network, and what specifically he or she did. But in some investigations there can be complexities beyond this.

For example, if the former employee did access the company network, but for a benign purpose such as assisting trouble shooting some issue for a former colleague, or completing human resources paperwork, that changes the situation. However, if there is a connection between the former employee and the website that the current employee visited that would add further complexity. Also how frequently the infected website was visited from within the company will impact how likely that is to be the source of the spyware. If the suspect former employee had access and the capability to create spyware that would also impact how likely the suspect was to be the source of the spyware. These issues can be introduced to the graph via weighting the various edges. For example, when weighting the connection between an employee at the victim company and the infected website can be expressed as an ordinal value, such as 1) visited the web site very infrequently and is not known to have downloaded anything; 2) visited the website with some frequency; 3) routinely visited the website; and 4) is known to have downloaded files from the infected website.

Similar weighting can be applied to the suspect who is believed to have created the malware found on the infected website. An ordinal system can again be used to illustrate how strong the connection is between the suspect and the infected website. The strongest connection would be if the suspect is known to have uploaded something to that website proximate to the time the website became infected. In this scenario, which is growing more complex, graph theory can be a valuable tool for evaluating the evidence. Furthermore, it may be more useful to use the adjacency matrix rather than a pictorial description of the graph. The adjacency matrix, as well as other algebraic forms of a graph (Godsil & Royle, 2013), are more readily introduced into computer algorithms or spreadsheets, making analyses of the data easier. An incidence matrix can also be useful in evaluating forensic evidence. An incidence matrix records edges that are incident from a given vertex. Consider the following graph:



Note that the edges have direction, so that vertex 2 is incident to vertices 1, 4, and 5. To create an incidence matrix, the vertices are the rows and the edges are the columns. For directed graph, an edge is only considered if it is incident from a given vertex.

The incidence matrix is shown below.

0	0	1	1	0
1	0	0	1	1
0	1	0	0	0
0	0	1	0	0
1	0	1	1	0

Incidence Matrix

Since vertex 1 is incident to vertices 3 and 4, there is a 1 in each of those columns. However, vertices 2 and 5 are incident to vertex 1, so even though there is an edge connection vertex 1 to vertices 2 and 5, those entries are 0. This modified incident matrix can give the investigator a very easy to analyze view of the connections between the various entities in a case (suspects, victims, systems, etc.). The standard incidence matrix of graph theory could be further modified to allow for weighting the edges. One method for accomplishing this would be to simply put the weight in parentheses in the incidence matrix. Consider the example where the investigator is using an ordinal weighting system number from 1 to 4, with 1 being minimal connection and 4 being a very strong connection. The incidence matrix from the previous graph could be re-written as follows.

0	0	1 (2)	1 (3)	0
1 (4)	0	0	1 (1)	1 (1)
0	1(3)	0	0	0
0	0	1(1)	0	0
1(1)	0	1(3)	1(1)	0

Weighted Incidence Matrix

This modified incidence matrix provides the investigator with a clearer view of the relationships between entities in the case. This provides a mathematical model of the case that can be readily analysed.

IV. CONCLUSIONS

Graph theory is a well-established, mathematical method for evaluating relationships. By applying the principles of graph theory to a forensic investigation, the relationships between entities can be weighted and mathematically studied. This provides a rigorous mechanism for evaluating complex forensic cases. In most investigations the graph will be a multi-graph and a digraph. Weighting the arcs in a digraph is useful in evaluating forensic evidence. It is recommended that small ordinal scales be used to weight the arcs in the graph. The goal is to provide a mathematical model of the investigation, including all the entities in the investigation and the relationship between those entities.

This paper focused on the application of algebraic graph theory in order to provide a mathematical model of a given event. That model can then be useful for an investigator who is attempting to analysed the event in question. While this paper did not explore optimization aspects of graph theory, such applications would be useful in many forensic investigations. Further research is recommended in the applications of graph theory to forensic investigations, particularly the use of optimization aspects of graph theory in order to evaluate connections between entities in an investigation. It also would be beneficial to examine isomorphism's in graph theory and how those might allow diverse crimes committed by the same actor(s) to be linked.

REFERENCES

[1]. Ahlswede, R., Cai, N., Li, S. Y., & Yeung, R. W. (2000). Network information flow. *IEEE Transactions on information theory, 46*(4), 1204-1216
[2]. Amaral, L. A., & Ottino, J. M. (2004). Complex networks. *The European Physical Journal B-Condensed Matter and Complex Systems, 38*(2), 147-162.
[3]. Balakrishnan, V.K. (2010). *Introductory Discrete Mathematics*. Mineola, New York: Dover Publications
[4]. Bollobás, B. (2013). *Modern graph theory (Vol. 184)*. Springer Science & Business Media
[5]. Bondy, A., Murty, U. (2008). *Graph Theory*. New York City, NY: Springer Publishing
[6]. Catanese, S. A., Fiumara, G. (2010, October). A visual tool for forensic analysis of mobile phone traffic. *In Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence* (pp. 71-76). ACM.
[7]. Chartrand, C. (1985). *Introductory Graph Theory*. New York City, NY: Dover Publication.
[8]. Chaski, C. (2005). Who's at The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence, 4*(1).
[9]. Clark, J., & Holton, D. A. (1991). *A first look at graph theory (Vol. 1)*. Teaneck, NJ: World Scientific.
[10]. Deo, N. (2016). *Graph Theory with Applications to Engineering and Computer Science*. Mineola, NY: Dover Publications
[11]. Easttom, C. (2016). Applying Graph Theory to Evidence Evaluation. *Research Gate*. DOI: 10.13140/RG.2.2.23391.0528
[12]. Gibilisco, S. (2004). *Statistics Demystified*. New York City, NY: McGraw-Hill.
[13]. Godsil, C., & Royle, G. F. (2013). *Algebraic graph theory (Vol. 207)*. Springer Science & Business Media.
[14]. Haggerty, J., Karran, A., Lamb, D., & Taylor, M. (2011). A Framework for the Forensic Investigation of Unstructured Email Relationship Data. *International Journal of Digital Crime and Forensics, 3*(3), 1-18.
[15]. Holme, P. (2003). Congestion and Centrality in Traffic Flow on Complex Networks. *Advances in Complex Systems, 6*(02), 163-176
[16]. Peterson, G., Shenoi, S. (2011). Advances in Digital Forensics VII: 7th IFIP WG 11.9 International Conference.
[17]. Trudeau, R. (1994). *Introduction to Graph Theory*. Mineola, New York: Dover Publications.
[18]. Wang, Wei, (2010). A Graph Oriented Approach for Network Forensic Analysis. Graduate Theses and Dissertations. Paper 11736
[19]. Wang, W., & Daniels, T. E. (2006, September). Diffusion and Graph Spectral Methods for Network Forensic Analysis. In Proceedings of the 2006 workshop on New security paradigms (pp. 99-106). ACM.
[20]. Zadora, G., & Ramos, D. (2010). Evaluation of Glass Samples for Forensic Purposes—An Application of Likelihood Ratios And An Information–Theoretical Approach. *Chemometrics and Intelligent Laboratory Systems, 102*(2), 63-83.
[21]. Zufferey, A., Ratle, F., Ribaud, O., Esseiva, P, & Kanevski, M. (2006). Pattern Detection in Forensic Case Data Using Graph Theory: Application to Heroin Cutting Agents. *Forensic Science International 167* (2-3), pp 242–246.

