Data analytics for Cyber security

-Cybersecurity through Network and Graph Data-

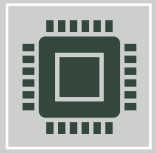
Vandana P. Janeja

©2022 Janeja. All rights reserved.

Outline

Graph Properties

Graphs for Cybersecurity: Understanding Evolving Network Communication



Graphs for Cybersecurity: Similarity-Based Redundancy in Router Connectivity

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graphs in cybersecurity

- Graph (as discussed in chapter 4) depict the relationships between various entities
- A graph is comprised of nodes and links (or vertices and edges)
- The node or vertex represents an entity such as an IP address and the edge or link represents the relationship between the two entities, such as message sent from one IP address to another
- Network traffic data lends itself to graph representation which opens up a vast array of graph-based analytics that can be applied to studying cybersecurity issues

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graph examples in Cybersecurity: Communication graphs

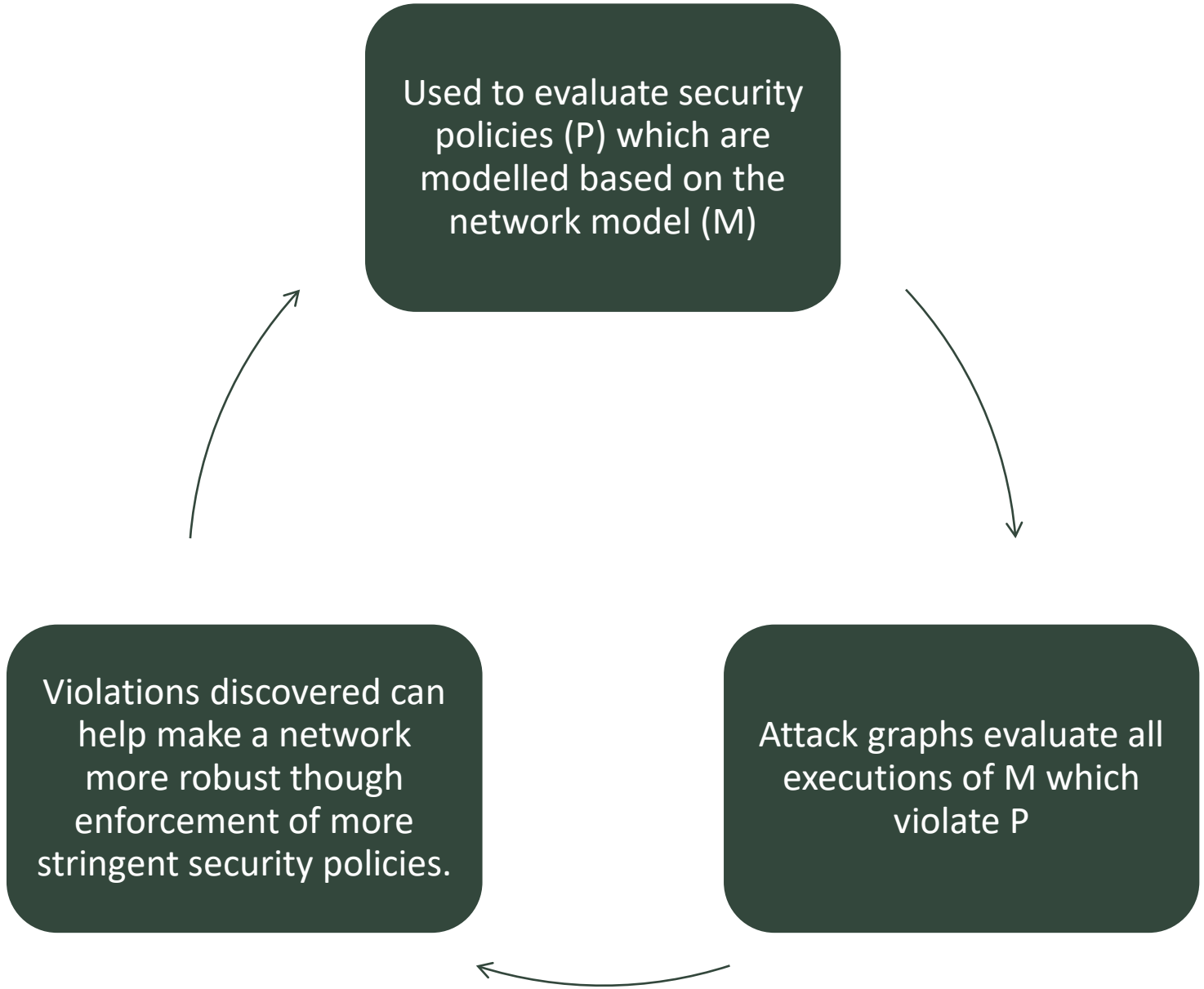
Network traffic has been extensively studied as a graph,

source and destination IP addresses are vertices and the communication between them is the edge

Various properties such as centrality, density and diameter have been used to evaluate the state of the graph and studying communication patterns over time

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graph examples in Cybersecurity:Attack graphs



Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graph examples in Cybersecurity: **Threat Similarity based graphs**

Graphs have been used to find relationships between known attacks and vulnerabilities, Such that known attacks are the nodes and the similarity between two vulnerabilities is the edge

The weight on the edge depicts the level of similarity

These types of graphs can be augmented with semantic and contextual information

These graphs can be queried when a new unknown threat is seen, such as a zero day attack

By computing the similarities between the zero day attack and the known attacks a score can be provided to indicate the level of threat from this zero day.

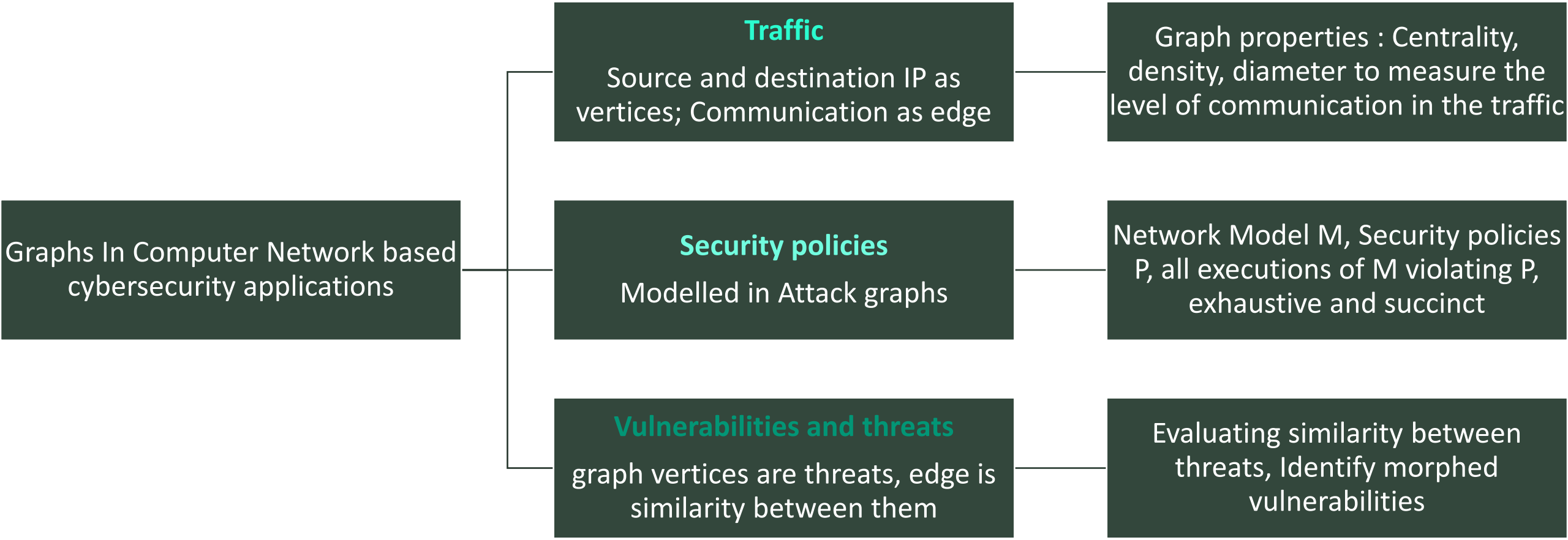
Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graph examples in Cybersecurity: Threat Propagation

- A cyber threat not only impacts an individual entity (one single computer or IP address) but can easily propagate to other connected entities
- Studies have been conducted to identify the spread and prevention of threats such as in high-risk groups of IP addresses
- Some of these studies and approaches are restricted to individual IP behavior, such as black and white-listing IP addresses, which limit the understanding of the dynamics involved in the spread of the threat
- Apart from individual behavior, high-risk behavior can also be captured at two other levels: the micro-social level (e.g., personal network), and macro level (through analysis of the network structural factors influencing the transmission mechanism)
- Computer networks and communication across the networks provide a framework to study the spread and prevention of high-risk behavior for wide spread cyber attacks

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Cybersecurity Applications that can utilize Graphs



Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

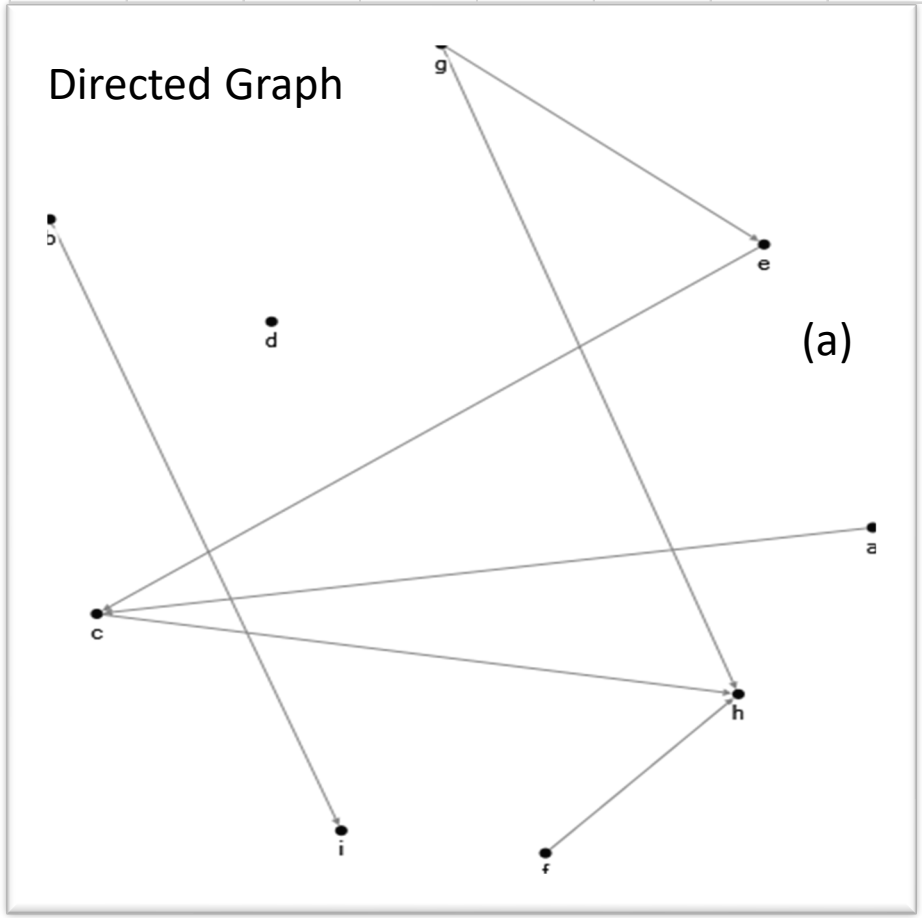
Graph Properties: **Adjacency Matrix**

- A graph can be represented as an adjacency matrix
- Let us consider a graph G with E edges and V vertices
- The adjacency matrix of G is a V X V matrix where each cell in the matrix has a 0 for no edge between the pair of vertices or 1 for an edge between the pair of vertices
- If there are multiple edges between two vertices then the value of the matrix cell is the number of edges. In an undirected graph the matrix is symmetric
- The sum of the row or column is the same, which is the degree of a vertex
- A degree is the number of edges incident on a node or vertex
- In an undirected graph the matrix is asymmetric
- The sum of the row or column in the asymmetric matrix is the out degree and in degree respectively

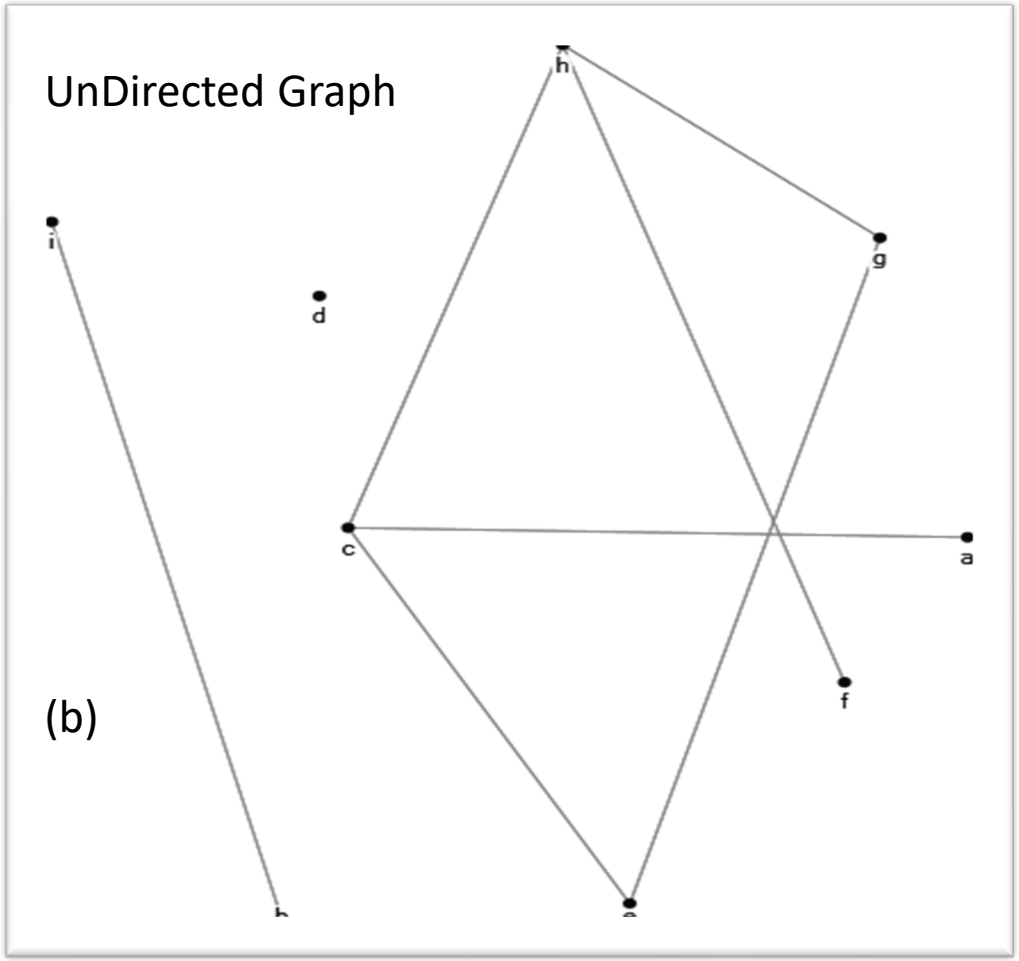
Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Example: Adjacency Matrix

	a	b	c	d	e	f	g	h	i	Out-Degree
a	0	0	1	0	0	0	0	0	0	1
b	0	0	0	0	0	0	0	0	1	1
c	0	0	0	0	0	0	0	1	0	1
d	0	0	0	0	0	0	0	0	0	0
e	0	0	1	0	0	0	0	0	0	1
f	0	0	0	0	0	0	0	1	0	1
g	0	0	0	0	1	0	0	1	0	2
h	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0
In-Degree	0	0	2	0	1	0	0	3	1	



	a	b	c	d	e	f	g	h	i	Degree
a	0	0	1	0	0	0	0	0	0	1
b	0	0	0	0	0	0	0	0	1	1
c	1	0	0	0	1	0	0	1	0	3
d	0	0	0	0	0	0	0	0	0	0
e	0	0	1	0	0	0	1	0	0	2
f	0	0	0	0	0	0	0	1	0	1
g	0	0	0	0	1	0	0	1	0	2
h	0	0	1	0	0	1	1	0	0	3
i	0	1	0	0	0	0	0	0	0	1
Degree	1	1	3	0	2	1	2	3	1	



Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

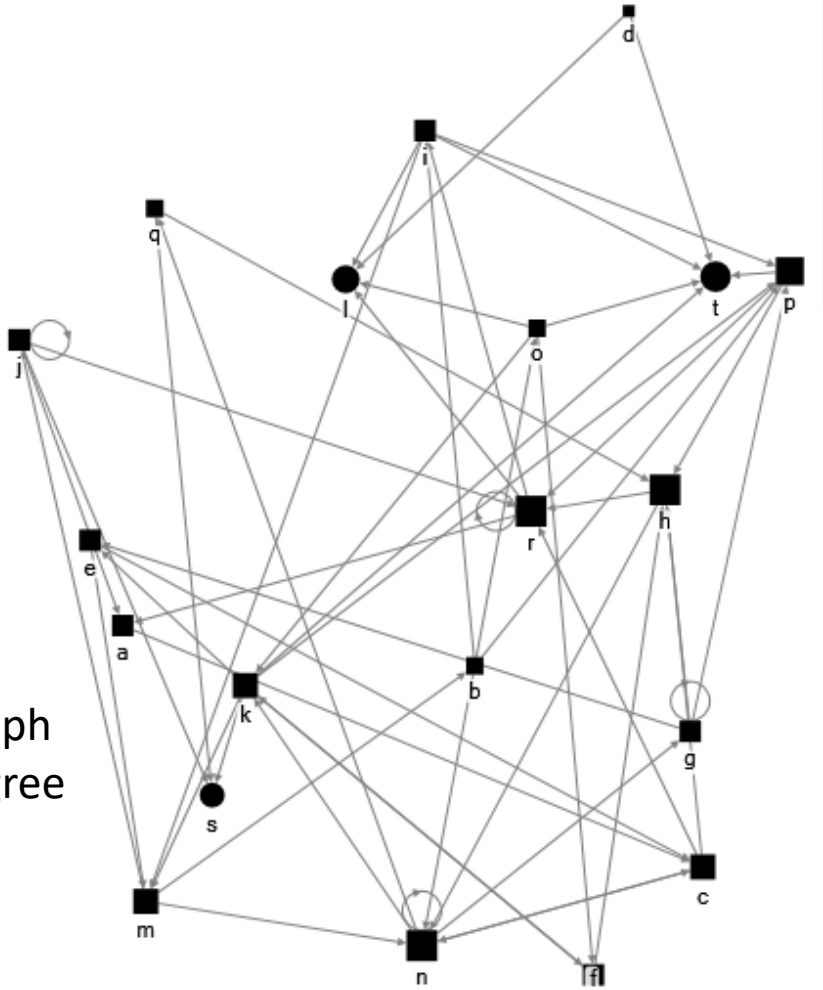
Graph Properties

- **Walk**
  - In a graph the sequence of edges leading from one vertex to another is called a walk
  - A walk where no vertex appears more than once is a path
  - The distance between two vertices is the shortest path between the two vertices, also referred to as Geodesic path
  - A graph can be disconnected when some parts of the graph are not connected by a path
  - For example in the graphs in figure, there are three disconnected components.
- **Centrality**
  - Centrality in a graph structure indicates prominence of a node in a graph
  - A node with high centrality will potentially be more influential
  - A simple type of centrality is the degree centrality which indicates the total number of edges incident on a node in an undirected graph
  - This can be divided into in-degree and out-degree for directed graphs
  - Betweenness centrality of a vertex *v*, measures the number of geodesic paths connecting vertices *ij* that are passing through *v*, divided by the total number of shortest paths connecting *ij*
  - This is summed across all *ij* to get the overall betweenness centrality of *v*
- **Cliques:**
  - A clique is defined as any complete sub graph such that each pair of the vertices in the sub graph are connected by an edge; Cliques may be mapped to coteries (such as high risk groups in the context of spreading a virus or DoS attack)
  - The nodes among different cliques may overlap, which indicate a higher relevance of those nodes in high-risk groups.

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t
a	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0
c	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0
d	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
e	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
g	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
i	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1
j	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
k	0	0	0	0																
l	0	0	0	0																
m	0	1	0	0																
n	0	0	1	0																
o	0	0	0	0																
p	0	0	0	0																
q	0	0	0	0																
r	1	0	0	0																
s	0	0	0	0																
t	0	0	0	0																

(A)



As Directed Graph
Shape: Out Degree
Size: In Degree

	Degree	In	Out
a	3	2	1
b	5	1	4
c	5	3	3
d	2	0	2
e	4	2	2
f	3	2	2
g	6	2	4
h	7	5	2
i	6	2	4
j	6	2	4
k	8	3	6
l	4	4	0
m	6	3	3
n	9	5	5
o	5	1	4
p	7	4	3
q	3	1	2
r	9	5	4
s	3	3	0
t	5	5	0

As Undirected Graph
Size: Degree

Example: Centrality

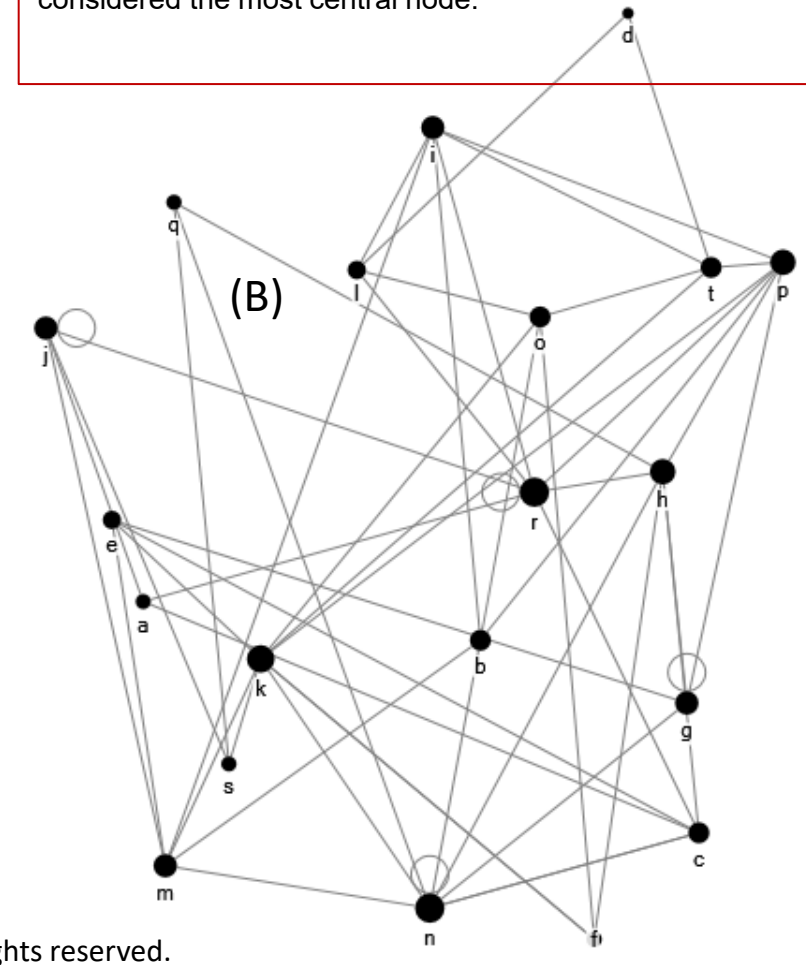
(A) presents a directed graph view and (B) shows an undirected view of the graph

In (A) it is evident that node K has high out degree and nodes H, N, R and T have high out degree

Over all vertex N has the highest degree

In terms of centrality in an undirected graph vertex n will be considered the most central node.

(B)



Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graph Properties: Centrality Over Time

- Graph properties can be used to evaluate consistency of nodes over time
- Example depicts three time periods *T1, T2 and T3* with their corresponding graphs
- Each time period has a set of nodes *'a'* through *'f'* which communicate with each other
- The corresponding graphs and degrees are shown in the figure
- A simple degree threshold can be used such that anything less than or equal to the threshold is not central
- In this example dataset threshold is two, then the central nodes are shown which are equal or greater than this threshold
- Node *'c'* which is central in time period *T1* does not show up in the other time periods
- Association rules for each time period are also shown which provide a complementary view of the communication in this example
- For example '*f'* which is not a central node is also not frequent as support is low
- However, in all time periods the rule *f*=> *e* is always present
- Similarly, the nodes '*a'* and'd' are consistently central and the rule *a*=>*d* is consistently a strong rule in terms of support and confidence
- On the other hand *a*=>*e* has high confidence and both nodes are central, however the support is low as compared to other rules with central nodes
- Multiple types of analysis can be used to analyze the network traffic data to provide complimentary knowledge.

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

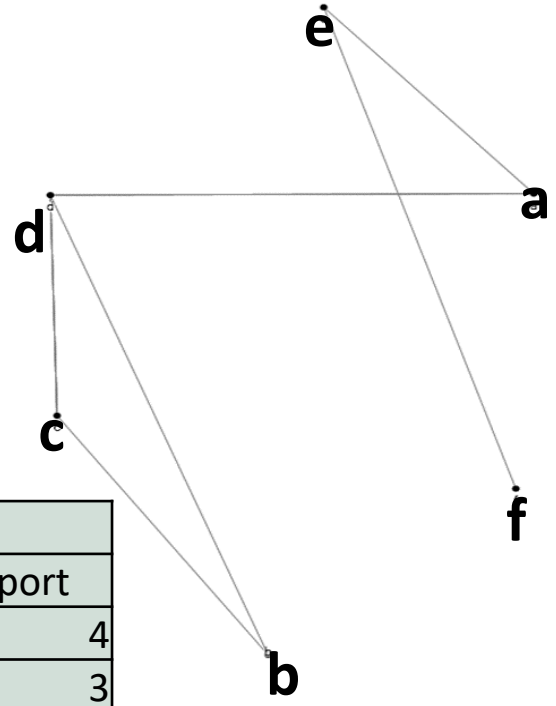
Example: Centrality Over Time

	Time T1
a	2
b	2
c	2
d	3
e	2
f	1

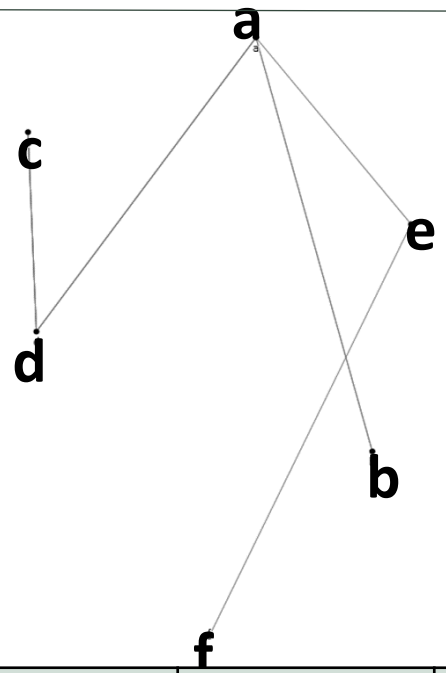
	Time T2
a	3
b	1
c	1
d	2
e	2
f	1

	Time T3
a	3
b	1
c	1
d	2
e	2
f	1

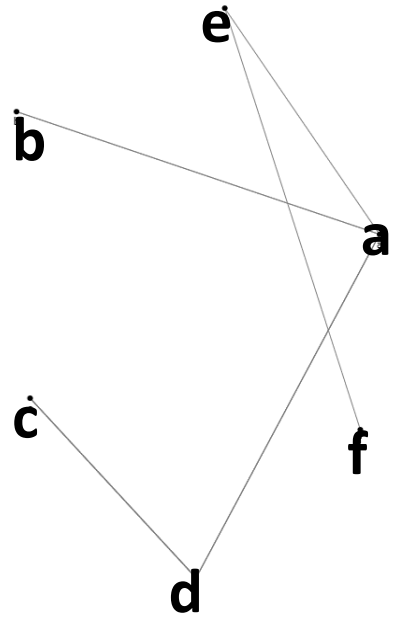
central nodes over time (degree >=2): a,b,c,d,e | a,d,e | a,d,e



	T1	
Rules	Confidence	Support
c=>d	1	4
a=>d	1	3
c=>b	1	1
a=>e	1	1
f=>e	1	1



T2		
Rules	Confidence	Support
c=>d	1	4
b=>a	1	2
a=>e	1	1
f=>e	1	1



T3		
Rules	Confidence	Support
c=>d	1	4
b=>a	1	2
a=>e	1	1
f=>e	1	1

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Understanding Evolving Network Communication: Centrality over time

- The level of connectivity to or from a given node can be used to measure the centrality of the node
- This level of centrality can be used to determine the importance of a node and how critical this node is on the network
- The degree centrality is the number of edges incident to a given node
- Page rank provides the number of highly central nodes that are adjacent to a given node
- The number of nodes that are connected to a given node and the quality of the nodes that connect to it can both be identified
- Intuitively, one can expect to find certain nodes to be consistently central over time periods
- Alternatively, a node which is central only at one or few time periods is interesting to study as it may be a source or destination of an attack or alternatively, due to the period in time, may have taken an important role in the network

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Understanding Evolving Network Communication: Densification over time

- Studies have indicated that an increase in the number of nodes on the network over a period of time, also leads to an increase in the number of edges
- This process is referred to as densification of a network
- However, these studies have looked at densification of all nodes and edges on the network without looking at evolving densification
- Densification can also be evaluated in in terms of the highly central nodes
- Determining if the change in the behavior of a critical node such as its unexpected absence from the network can identify if this affects the densification of the network
- If densification is tracked over a period of time, it is possible to identify time points where the densification is anomalous as compared to the overall trend in the network
- Such time periods may indicate unusual changes in the network
- Now if such changes are consistent over multiple large time periods (depicting periodicity, say for example every year in December) then this can be normal, however if this happens in only one sample year then it may indicate an anomalous time period

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Understanding Evolving Network Communication: Diameter over time

- Densification draws nodes closer to each other such that the distance between those nodes that are far apart, becomes smaller
- As a result, the diameter of the network reduces, a concept commonly described as the shrinking diameter
- Similar to densification, changing diameter of the network over time can provide insights about the health of the network traffic
- This analysis can be performed for the central nodes
- The change in the behavior of a critical node such as its unexpected absence from the network can impact the diameter of the network
- Drastic changes in the diameter such as expanding diameter can indicate malicious effects on the central nodes

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graphs for Cybersecurity: Similarity-Based Redundancy in Router Connectivity

- Router configuration can be a difficult and complex task, especially on larger networks, which can typically have hundreds of routers
- There is a clear need for automatic router configurations, however, many organizations continue to rely on manual configurations and reconfigurations leading to several network outages
- After several reconfigurations it is possible that the network topology may be considerably different from the originally planned topology
- As a result, problems such as bottlenecks or redundancy may crop up
- Graphs can be used to identify such bottlenecks and redundancy using graph properties and node similarities based on their connectivity
- The following example shows a router connectivity represented as a graph and the nodes similarity computed based on their connectivity

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Representing Router Connectivity as a Graph

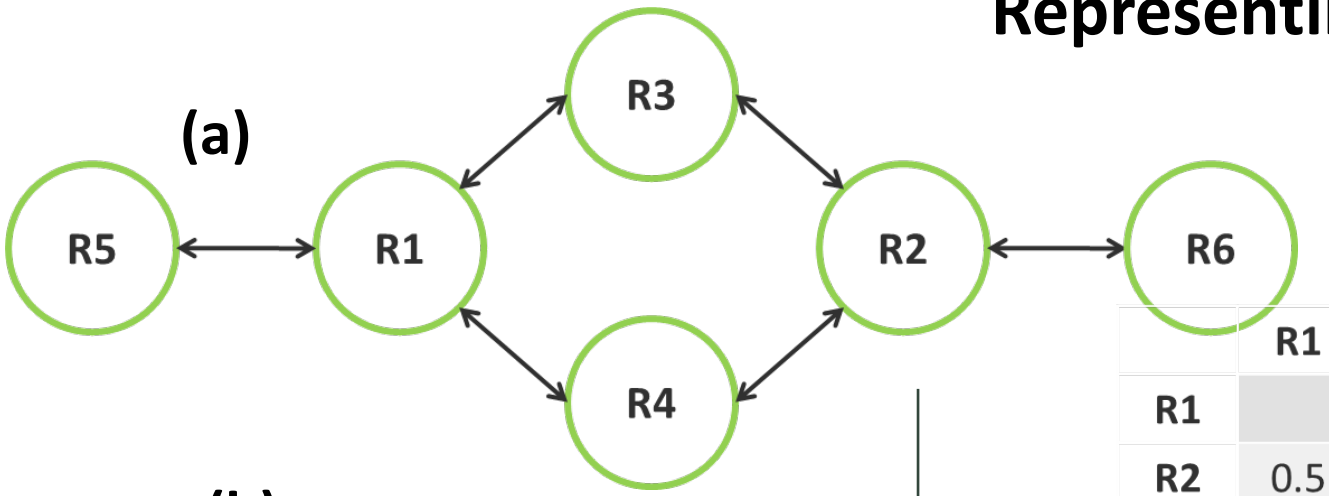
- Router connectivity can be considered as a graph and similarity coefficients can provide an intuitive method of determining whether a router is similar or unique in relation to the other routers connected on the graph representing the network
- An example use is shown using Jaccard Coefficient (JC)
- An example graph and adjacency matrix with various routers connected are shown in the example
- Pairwise similarity between each pair of the routers is computed using JC which is given as Positive matches (1-1 match) divided by the sum of Positive Matches and positive mismatches (0-1 or 1-0 mismatch)
- Thus looking at the two rows for routers R4 and R3 we can see that they have perfect set of matches (1-1) and no mismatches (0-1 or 1-0)
- The JC for similarity between R3 and R4 is 1
- They can be seen as redundant routers which are perfectly similar to replace each other's role
- On the other hand when we look at the total similarity of each node and its degree in (d) and (e) respectively
- We can see that nodes 5 and 6 have lowest degree and lowest similarity, thus they are the most vulnerable nodes in terms of connectivity
- Similarly nodes R1 and R2 are highly central but have only some similarity
- They are possible bottle neck nodes due to their unique connectivity and highly central nature.

- Such preemptive exploratory analysis using basic graph properties can help facilitate the network management for better response to threats and recovery from network impact

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Representing Router Connectivity as a Graph

(a)



(b)

	R1	R2	R3	R4	R5	R6
R1	0	0	1	1	1	0
R2	0	0	1	1	0	1
R3	1	1	0	0	0	0
R4	1	1	0	0	0	0
R5	1	0	0	0	0	0
R6	0	1	0	0	0	0

(c)

JC

	R1	R2	R3	R4	R5	R6
R1						
R2	0.5					
R3	0	0				
R4	0	0	1			
R5	0	0	0.5	0.5		
R6	0	0	0.5	0.5	0	

(d)

	Total JC
R1	0.5
R2	0.5
R3	2
R4	2
R5	1
R6	1

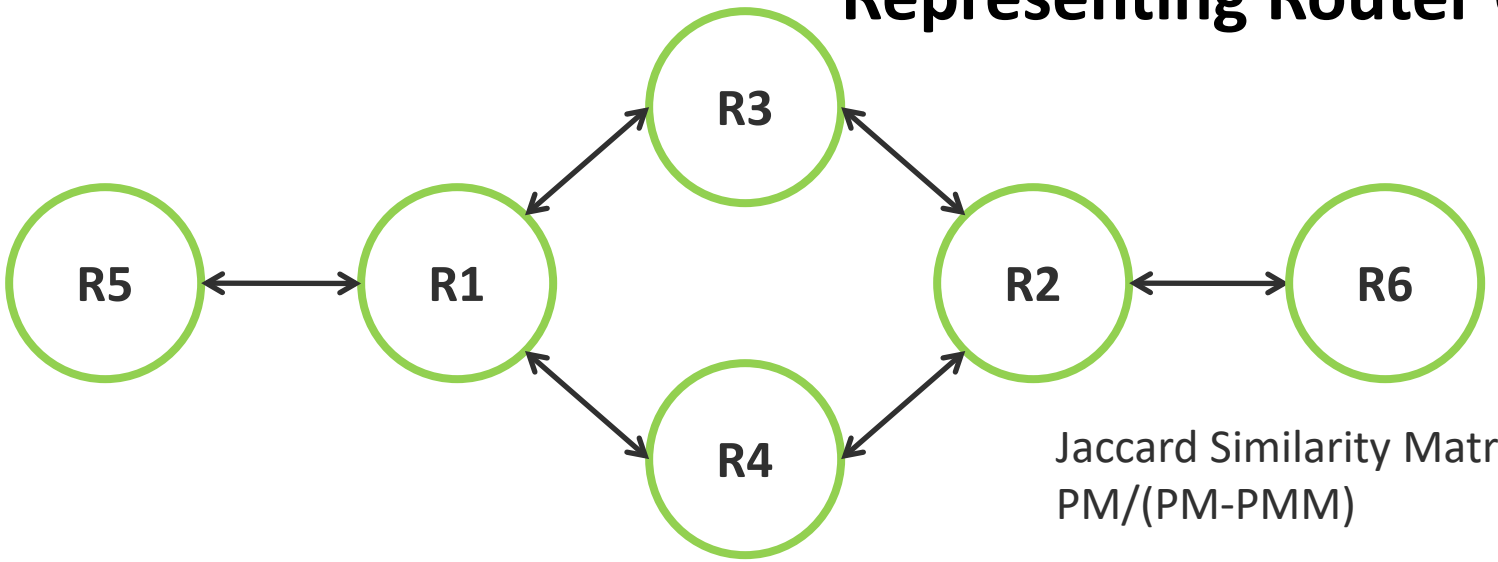
(e)

Degree

	R1	R2	R3	R4	R5	R6	Total
R1	0	0	1	1	1	0	3
R2	0	0	1	1	0	1	3
R3	1	1	0	0	0	0	2
R4	1	1	0	0	0	0	2
R5	1	0	0	0	0	0	1
R6	0	1	0	0	0	0	1

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Representing Router Connectivity as a Graph



Jaccard Similarity Matrix
PM/(PM-PMM)

	R1	R2	R3	R4	R5	R6
R1	0	0	1	1	1	0
R2	0	0	1	1	0	1
R3	1	1	0	0	0	0
R4	1	1	0	0	0	0
R5	1	0	0	0	0	0
R6	0	1	0	0	0	0

	R1	R2	R3	R4	R5	R6
R1						
R2	0.5					
R3	0	0				
R4	0	0	1			
R5	0	0	0.5	0.5		
R6	0	0	0.5	0.5	0	

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Representing Router Connectivity as a Graph

	JC
R1	**0.5**
R2	**0.5**
R3	**2**
R4	**2**
R5	**1**
R6	**1**

	R1	R2	R3	R4	R5	R6	Total
R1	0	0	1	1	1	0	**3**
R2	0	0	1	1	0	1	**3**
R3	1	1	0	0	0	0	**2**
R4	1	1	0	0	0	0	**2**
R5	1	0	0	0	0	0	**1**
R6	0	1	0	0	0	0	**1**

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

Graphs for Cybersecurity: Similarity-Based Redundancy in Router Connectivity

- A bottleneck router
  - Is a router that is dissimilar to other routers on the network in terms of its connectivity, such that it has a unique function to perform through its unique set of connections
  - Other parts of the network will rely on it to connect to keep certain parts of the network connected and if this router is impacted in any way to perform its functionality then the overall network connectivity may suffer.
- A redundant router
  - Is highly similar to other routers in terms of its connectivity
  - Taking a redundant router offline would probably have little effect on the network connectivity
  - Such a router can act as a failsafe if this redundant router is similar to some critical network routers
  - On the other hand, this router can be used for load balancing or reallocating it to other useful roles

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.

References

- Wang, L., Singhal, A., & Jajodia, S. (2007, October). Toward measuring network security using attack graphs. In *Proceedings of the 2007 ACM workshop on Quality of protection* (pp. 49-54). ACM.
- Wang, L., Singhal, A., & Jajodia, S. (2007, July). Measuring the overall security of network configurations using attack graphs. In *IFIP Annual Conference on Data and Applications Security and Privacy* (pp. 98-112). Springer, Berlin, Heidelberg.
- Tartakovsky, A. G., Polunchenko, A. S., & Sokolov, G. (2013). Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, *7*(1), 4-11.
- Namayanja, J. M., & Janeja, V. P. (2014, October). Change detection in temporally evolving computer networks: A big data framework. In *Big Data (Big Data), 2014 IEEE International Conference on* (pp. 54-61). IEEE.
- Jha, S., Sheyner, O., & Wing, J. (2002). Two formal analyses of attack graphs. In *Computer Security Foundations Workshop, 2002. Proceedings. 15th IEEE* (pp. 49-63). IEEE.
- Trinius, P., Holz, T., Göbel, J., & Freiling, F. C. (2009, October). Visual analysis of malware behavior using treemaps and thread graphs. In *Visualization for Cyber Security, 2009. VizSec 2009. 6th International Workshop on* (pp. 33-38). IEEE.
- Ingols, K., Lippmann, R., & Piwowarski, K. (2006, December). Practical attack graph generation for network defense. In *Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual* (pp. 121-130). IEEE.
- AlEroud, A., & Karabatis, G. (2013). A system for cyber attack detection using contextual semantics. In *7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing* (pp. 431-442). Springer, Berlin, Heidelberg.
- Rivest, R. L., & Vuillemin, J. (1976). On recognizing graph properties from adjacency matrices. Theoretical Computer Science, 3(3), 371-384.
- Robin J Wilson. 1986. Introduction to Graph Theory. John Wiley & Sons, Inc., New York, NY, USA.
- Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery, 29(3), 626-688.
- Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215-239.
- A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the Internet, volume 2380 of Automata, Languages and Programming, page 781. Springer, 2002.
- M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In SIGCOMM, 1999.
- G. Phillips, S. Shenker, and H. Tangmunarunkit. Scaling of multicast trees: Comments on the chuang- sirbu scaling law. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication.*, 1999.
- U. Kang, C. Tsourakakis, A. Appel, C. Faloutsos, and J. Leskovec. Radius plots for mining tera-byte scale graphs: Algorithms, patterns, and observations. In *SIAM International Conference on Data Mining (SDM).*, 2010.
- J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, 2005.
- Caldwell, D., Gilbert, A., Gottlieb, J., Greenberg, A., Hjalmtysson, G., & Rexford, J. (2004). The cutting EDGE of IP router configuration. *ACM SIGCOMM Computer Communication Review*, *34*(1), 21-26.
- Lee, S., Levanti, K., & Kim, H. S. (2014). Network monitoring: Present and future. *Computer Networks*, *65*, 84-98.
- Lewis, D. M., & Janeja, V. P. (2011). An empirical evaluation of similarity coefficients for binary valued data. *IGI Global*, 44-66.

Data Analytics for Cybersecurity, ©2022 Janeja All rights reserved.