# Social Network Analysis

*15*

Learning Objectives

* Understand the concepts of Social Networks and Social Network Analysis
* Learn the many applications of SNA
* Know the different network topologies
* Identify the techniques and algorithms for discovering subnetworks
* Understand the concept of a web page
* Learn about PageRank and how it is recursively computed
* Summarize the many differences between SNA and data mining

### INTRODUCTION

Social Networks are a graphical representation of relationships among people and/ or entities. Social Network Analysis (SNA) is the art and science of discovering patterns of interaction and influence within the participants in a network. These participants could be people, organizations, machines, concepts, or any other kinds of entities. An ideal application of social network analysis will discover essential characteristics of a network including its central nodes and its subnetwork structure. Subnetworks are clusters of nodes, where the within-subnetwork connections are stronger than the connections with nodes outside the subnetwork. SNA is accomplished by graphically representing social relationships into a network of nodes and links and applying iterative computational techniques to measure the strengths of relationships. The social network analysis ultimately helps relate the totality of the network to the Unified Field which is the ultimate entity with infinite relationships among everything.

#### Caselet: The Social Life of Books—Visualizing Communities of Interest via Purchase Patterns on the WWW

*Amazon lists the top 6 books that were bought by individuals who also bought the book currently being browsed. This is one of Amazon’s value-added services. Let’s choose Tom Petzinger’s ‘The New Pioneers’. What themes would we see in the recommended books? What other topics are Tom’s readers interested in? Will Tom’s book ends up in the center of one large, massively interconnected cluster – a single*

*the community of interest? Or, will it end up linking together otherwise disconnected clusters – diverse communities of interest?*

*Figure C15.1 Shows the network surrounding ‘The New Pioneers’. Each node represents a book. A red line links books that were purchased together. The buying pattern of the books has been self-organized into emergent clusters that are named for the content of each cluster. Tom’s book does span a diversity of interests!*

*The most common measure in social networks is the network centrality. To assess the positional advantage, we measure each node’s network centrality. We have two parts of the network – the dense complexity science cluster and the dense Internet economy cluster, and also the other 2 interconnected clusters forming a large network component.*

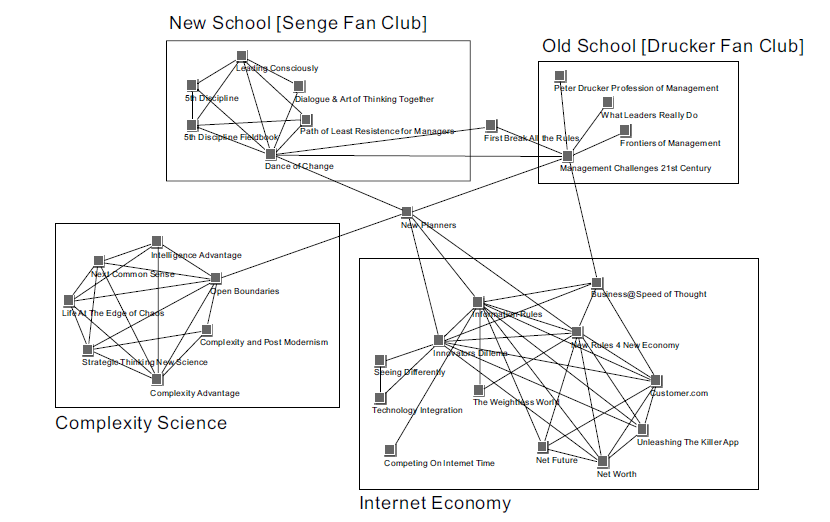


FIGURE C15.1

(*Source: *orgnet.com/booknet.html)

* 1. *Can these maps provide an opportunity to the businesses (to authors and publicists) that Amazon competes against?*
  2. *Could a similar map of services be developed for a healthcare situation? How?*

### Applications of SNA

*Self-awareness* Visualizing his/her social network can help a person organize their relationships and support network.

*Communities* Social Network Analysis can help identify, construct, and strengthening networks within communities to build wellness, comfort, and resilience. Analysis of joint authoring relationships and citations help identify subnetworks of specializations of knowledge in an academic field. Researchers at Northwestern University found that the most determinant factor in the success of a Broadway play was the strength of relationships amongst the crew and cast.

*Marketing* There is a popular network insight that any two people are related to each other through at most seven degrees of links. Organizations can use this insight to reach out with their message to a large number of people and also to listen actively to opinion leaders as ways to understand their customers’ needs and behaviors. Politicians can reach out to opinion leaders to get their message out.

*Public Health* Awareness of networks can help identify the paths that certain diseases take to spread. Public health professionals can isolate and contain diseases before they expand to other networks.

### Network Topologies

There are two primary types of network topologies – ring-type and hub-spoke topologies. Each of the topologies has different characteristics and benefits.

In the ring network, nodes typically connect to their adjacent nodes in the network. Potentially all nodes can be connected. A ring network could be dense where every node has a direct connection with practically every node. Or it could be sparse where every node connects to a small subset of the nodes. The number of connections needed to reach from one node to another could vary tremendously. A dense network, with more connections, will allow many direct connections between pairs of nodes. In a sparse network, one may need to traverse many connections to reach the desired destination. A peer-to-peer email (or messaging) network is an example of the ring model, as anyone can potentially directly connect with anyone else. In practice, the email networks are sparse as people connect directly with only a subset of people.

In the hub-spoke model, there is one central hub node to which all the other nodes are connected. There are no direct relationships between the nodes. Nodes connect through the hub node. This is a hierarchical network structure since the hub node is central to the network. The hub node is structurally more

important as it is central to all communications between other peripheral nodes. The good thing about a hub-spoke network is that one could predictably reach from any node to any other node through traversing exactly just two connections. As an example, modern airlines operate on this model to maintain hub networks from which they operate flights to a large number of airports.

The density of a network can be defined as the average number of connections per node. The cohesiveness of the network is a related concept, which is the average number of connections needed to reach from one node to the other.

Another way of analyzing networks is to define the centrality (or importance) of every node. The number of links associated with a node is a sign of the centrality of the node. In the ring network in the figure below, each node has exactly 2 links. Thus, there is no central node. However, in the hub-spoke network, the hub-node *N* has 8 links while all other nodes have only 1 link each. Thus, node *N* has a high centrality.

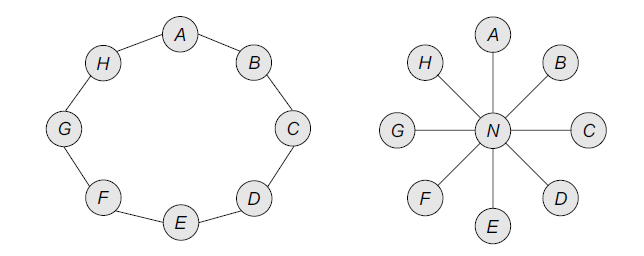


FIGURE 15.1 Network Topologies: Ring (left) and Hub-spoke (right)

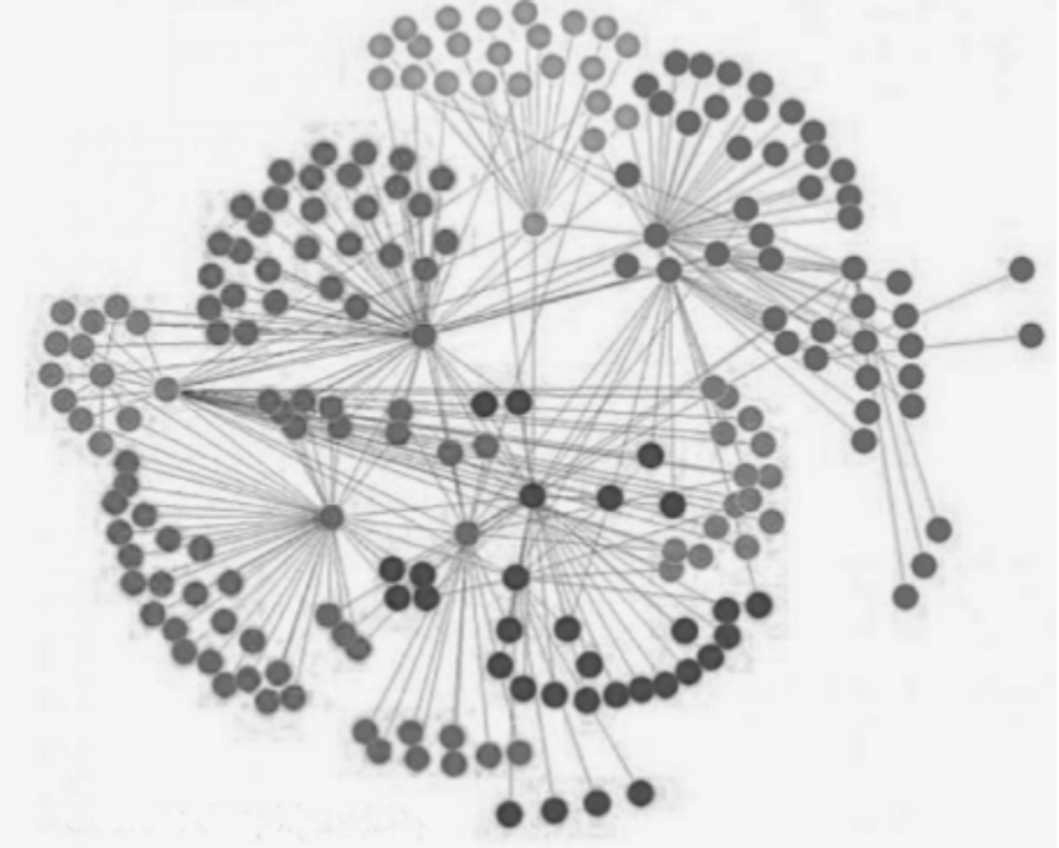
A variant that combines both the above types is the network of networks in which each participating network will connect with other networks at selected points of contact. For example, the internet is a network of networks. Here, all the commercial, university, government, and similar computer networks connect to one other at certain designated nodes (called gateways) to exchange information and conduct business.

### TECHNIQUES AND ALGORITHMS

There are two major levels of social network analysis – discovering subnetworks within the network and ranking the nodes to find more important nodes or hubs.

*Finding Subnetworks*

A large network could be better analyzed and managed if it can be seen as an interconnected set of distinct subnetworks each with its own distinct identity and unique characteristics. This is like doing a cluster analysis of nodes. Nodes with strong ties between them would belong to the same subnetwork, while those with weak or no ties would belong to separate subnetworks. This is an unsupervised learning technique, as in Apriori there is no correct number of subnetworks in a network. The usefulness of the subnetwork structure for decision-making is the main criterion for adopting a particular structure.



Source: GrapLab Inc.

FIGURE 15.2 A Network with Distinct Subnetworks

Visual representations of networks can help identify subnetworks. The use of color can help differentiate the types of nodes. Representing strong ties with thicker or bolder lines could help visually identify the stronger relationships. A subnetwork could be a collection of strong relationships around a hub node. In that case, the

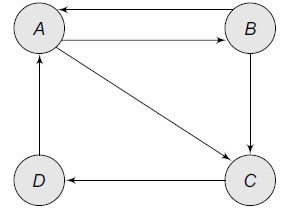
hub node could represent a distinct subnetwork. A subnetwork could also be a subset of nodes with dense relationships between them. In that case, one or more nodes will act as gateways to the rest of the network.

*Computing Importance of Nodes*

When the connections between nodes in the network have a direction to them, then the nodes can be compared for their relative influence or rank. (The terms importance and influence and rank of a node will be used interchangeably for this discussion.) This is done using the ‘Influence Flow model’. Every outbound link from a node can be considered an outflow of influence. Every incoming link is similarly an inflow of influence. More in-links to a node means greater importance. Thus there will be many direct and indirect flows of influence between any two nodes in the network.

Computing the relative influence of each node is done based on an input-output matrix of flows of influence among the nodes. Assume each node has an influence value. The computational task is to identify a set of rank values that satisfies the set of links between the nodes. It is an iterative task where we begin with some initial values and continue to iterate till the rank values stabilize.

Consider the following simple network with 4 nodes (A, B, C, D) and 6 directed links between them as shown in Figure 15.3. Note that there is a bidirectional link. Here are the links

Node A links into B

Node B links into C

Node C links into D

Node D links into A

Node A links into C

Node B links into A.

FIGURE 15.3

The goal is to find the relative importance, or rank, or every node in the network. This will help identify the most important node(s) in the network.

We begin by assigning the variables for influence (or rank) value for each node, as , , , and . The goal is to find the relative values of these variables.

There are two outbound links from node *A* to nodes *B* and *C*. Thus, both *B* and *C* receive half of node *A*’s influence. Similarly, there are two outbound links from node *B* to nodes *C* and *A*, so both *C* and *A* receive half of node *B*’s influence.

There is the only outbound link from node *D* into node *A*. Thus, node *A* gets all the influence of node *D*. There is the only outbound link from node *C* into node *D* and hence, node *D* gets all the influence of node *C*.

Node *A* gets all of the influence of node *D* and half the influence of node *B*. Thus,

Node *B* gets half the influence of node *A*. Thus,

Node *C* gets half the influence of node *A* and half the influence of node *B*. Thus,

Node *D* gets all of the influence of node *C* and half the influence of node *B*.

Thus,

We have 4 equations using 4 variables. These can be solved mathematically.

We can represent the coefficients of these 4 equations in a matrix form as shown in Dataset 15.1 given below. This is the Influence Matrix. The zero values represent that the term is not represented in an equation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset 15.1 |  | | | |
|  | *Ra* | *Rb* | *Rc* | *Rd* |
| *Ra* | 0 | 0.50 | 0 | 1.00 |
| *Rb* | 0.50 | 0 | 0 | 0 |
| *Rc* | 0.50 | 0.50 | 0 | 0 |
| *Rd* | 0 | 0 | 1.00 | 0 |

For simplification, let us also state that all the rank values add up to 1. Thus, each node has a fraction as the rank value. Let us start with an initial set of rank values and then iteratively compute new rank values till they stabilize. One can start with any initial rank values, such as 1/*n* or ¼ for each of the nodes.

|  |  |
| --- | --- |
| Variable | Initial Value |
| *Ra* | 0.250 |
| *Rb* | 0.250 |
| *Rc* | 0.250 |
| *Rd* | 0.250 |

Computing the revised values using the equations stated earlier, we get a revised set of values shown as Iteration1. (This can be computed easily by creating formulae using the influence matrix in a spreadsheet such as Excel.)

|  |  |  |
| --- | --- | --- |
| Variable | Initial Value | Iteration1 |
| *Ra* | 0.250 | 0.375 |
| *Rb* | 0.250 | 0.125 |
| *Rc* | 0.250 | 0.250 |
| *Rd* | 0.250 | 0.250 |

Using the rank values from Iteration1 as the new starting values, we can compute new values for these variables, shown as Iteration2. Rank values will continue to change.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Initial Value | Iteration1 | Iteration2 |
| *Ra* | 0.250 | 0.375 | 0.3125 |
| *Rb* | 0.250 | 0.125 | 0.1875 |
| *Rc* | 0.250 | 0.250 | 0.250 |
| *Rd* | 0.250 | 0.250 | 0.250 |

Working from values of Iteration2 and so, we can do a few more iterations till the values stabilize. Dataset 15.2 shows the final values after the 8th iteration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset 15.2 |  | | | | |
| Variable | Initial Value | Iteration1 | Iteration2 | … | Iteration8 |
| *Ra* | 0.250 | 0.375 | 0.313 | … | 0.333 |
| *Rb* | 0.250 | 0.125 | 0.188 | … | 0.167 |
| *Rc* | 0.250 | 0.250 | 0.250 | … | 0.250 |
| *Rd* | 0.250 | 0.250 | 0.250 | … | 0.250 |

The final rank shows that the rank of node *A* is the highest at 0.333. Thus, the most important node is *A*. The lowest rank is 0.167 of *Rb*. Thus, *B* is the least important node. Nodes *C* and *D* are in the middle. In this case, their ranks did not change at all.

The relative scores of the nodes in this network would have been the same irrespective of the initial values chosen for the computations. It may take a longer or shorter number of iterations for the results to stabilize for different sets of initial values.

### PAGERANK

PageRank is a particular application of the social network analysis techniques above to compute the relative importance of websites in the overall World Wide Web. The data on websites and their links are gathered through web crawler bots that traverse through the webpages at frequent intervals. Every webpage is a node in a social network and all the hyperlinks from that page become directed links to other webpages. Every outbound link from a webpage is considered an outflow of influence of that webpage. An iterative computational technique is applied to compute the relative importance of each page. That score is called PageRank, according to an eponymous algorithm invented by the founders of Google, the web search company.

PageRank is used by Google for ordering the display of websites in response to search queries. To be shown higher in the search results, many website owners try to artificially boost their PageRank by creating many dummy websites whose ranks can be made to flow into their desired website. Also, many websites can be designed to cyclical sets of links from where the web crawler may not be able to break out. These are called spider traps.

To overcome these and other challenges, Google includes a Teleporting factor in computing the PageRank. Teleporting assumed that there is a potential link from any node to any other node, irrespective of whether it exists. Thus, the influence matrix is multiplied by a weighting factor called Beta with a typical value of 0.85 or 85 percent. The remaining weight of 0.15 or 15 percent is given to teleportation. In the teleportation matrix, each cell is given a rank of 1/*n*, where *n* is the number of nodes on the web. The two matrices are added to compute the final influence matrix. This matrix can be used to iteratively compute the PageRank of all the nodes, just as shown in the example earlier.

### PRACTICAL CONSIDERATIONS

*Network Size* Most SNA research is done using small networks. Collecting data about large networks can be very challenging. This is because the number of links is the order of the square of the number of nodes. Thus, in a network of 1000 nodes, there are potentially 1 million possible pairs of links.

*Gathering Data* Electronic communication records (emails, chats, etc.) can be harnessed to gather social network data more easily. Data on the nature and quality of relationships need to be collected using survey documents. Capturing and cleaning and organizing the data can take a lot of time and effort, just like in a typical data analytics project.

*Computation and Visualization* Modeling large networks can be computationally challenging and visualizing them also would require special skills. Big data analytical tools may be needed to compute large networks.

*Dynamic Networks* Relationships between nodes in a social network can be fluid. They can change in strength and functional nature. For example, there could be multiple relationships between two people … they could simultaneously be coworkers, coauthors, and spouses. The network should be modeled frequently to see the dynamics of the network.

Table 15.1 Social Network Analysis vs Traditional Data Analytics

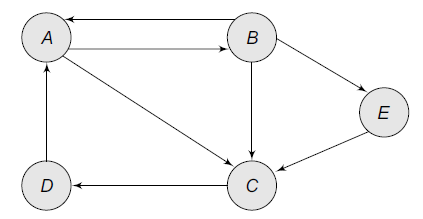
|  |  |  |
| --- | --- | --- |
| Dimension | Social Network Analysis | Traditional Data Mining |
| Nature of learning | Unsupervised learning | Supervised and unsupervised learning |
| Analysis of goals | Hub nodes, important nodes, and subnetworks | Key decision rules, cluster centroids |
| Dataset structure | A graph of nodes and (directed) links | Rectangular data of variables and in- stances |
| Analysis techniques | Visualization with statistics; iterative graphical computation | Machine learning, statistics |
| Quality measurement | Usefulness is a key criterion | Predictive accuracy for classification techniques |

Conclusion

Social network analysis is a powerful method of analyzing relationships among entities to identify strong patterns. There could be subnetworks in a network based on strong ties within the network. A computationally rigorous set of techniques can be used to rank every node in a network for its influence and importance. PageRank is the implementation of this algorithm in the context of ranking websites.

## Questions

1. What is social network analysis? How is it different from other data mining techniques such as clustering or decision trees?
2. How can SNA help with improving the gross national happiness of any country or society?
3. What kind of pitfalls should one guard against while doing SNA?
4. Data preparation takes more than 2/3 of the total time of an analytics project. Would that be true for SNA also?
5. Compute the rank values for the nodes for the following network which is a modified version of the exercise solved earlier. Which is the highest-ranked node now?



## True/False Questions

1. Social networks are a graphical representation of relationships among people and/or entities.
2. A social network represents social relationships into a network of nodes and links.
3. Social network analysis can be used to discover the subnetworks.
4. Subnetworks are clusters of nodes where the connections with nodes outside the subnetwork are stronger than within-subnetwork connections.
5. There are two essential network topologies – Ring-type and hub-spoke.
6. PageRank is a special type of network topology.
7. Teleporting is a way to enhance the PageRank of all the important websites.
8. All data mining techniques can be applied to SNA.