

SOCIAL NETWORK ANALYSIS

BAD714D

Module-1

Networks and Society - What is Social Network Analysis, why do We Study Social Networks, Applications of Social Network Analysis, Preliminaries, Three Levels of Social Network Analysis.

Network Measures - Network Basics, Node Centrality, Assortativity, Transitivity and Reciprocity, Similarity, Degeneracy.

T1 – Chapter 1 (1.1. – 1.5), Chapter 2 (2.1 – 2.6)

1.1 WHAT IS SOCIAL NETWORK ANALYSIS?

Network is an abstract representation of relations among entities. Entities in a network are represented as nodes, with links denoting interactions between the entities. For example, railway routes can be described as a network where nodes represent the railroad stations, and links between nodes specify a pathway connecting the stations. Social structure, at the macro-level, is the organisation of social relations and patterns existing in a particular society. A social network is a simplified representation of the social structure characterised by actors and ties, where actors represent individuals, groups, and organisations, and ties illustrate the interactions among actors that can range from the relationship among friends, family bond to trade relations between countries. Social networks, in the mathematical and network theory terminology, are represented as graphs where nodes refer to actors and links symbolise the ties. Figure 1.1 presents a social network of a small-scale organisation. Social network representation enables us to focus on relationships between individuals than just individuals who form the society.

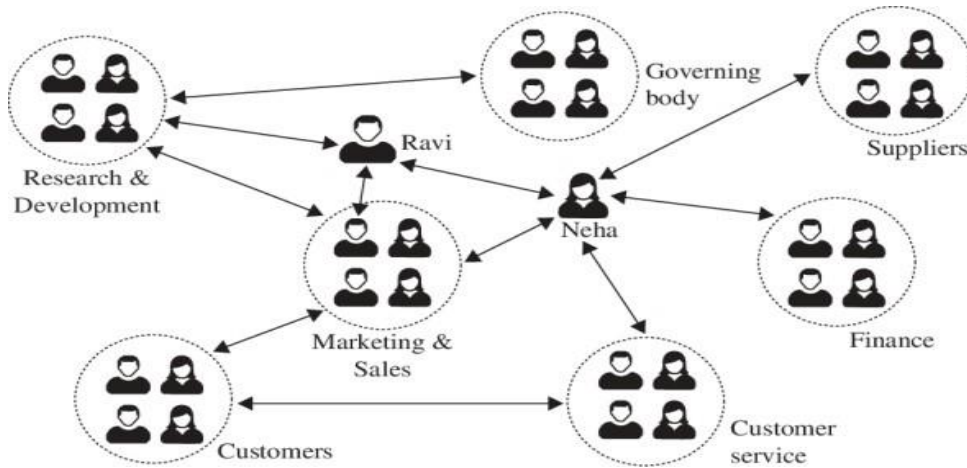


FIGURE 1.1. A small-scale organisation social network where entities in the organisation are represented as actors/nodes, and links between them denote interactions/collaborations between these entities.

The structure of a social network is dependent on individuals and the relationships between individuals, and this differentiates the study of social networks from other paradigms that consider individuals as separate identities.

SNA is the application of networks and graph theory to analyse the relations present in a society (online or offline). Similarities among individuals, social relations, interactions, and information flow within individuals or groups are some of the subjects that are studied and explored as part of SNA.

Facebook friendship network, scientist collaboration network, and employee interaction network of an organisation are a few examples of the social structures that are studied and analysed as a part of SNA.

Figure 1.2 presents a famous instance of Florentine families' network in the early 15th century (Breiger and Pattison 1986).¹ Although a small-scale network, the Florentine family network provides some significance of representing social structure as a social network, and simple analysis of the network, such as the degree of connectivity, offers some straightforward but interesting insights.

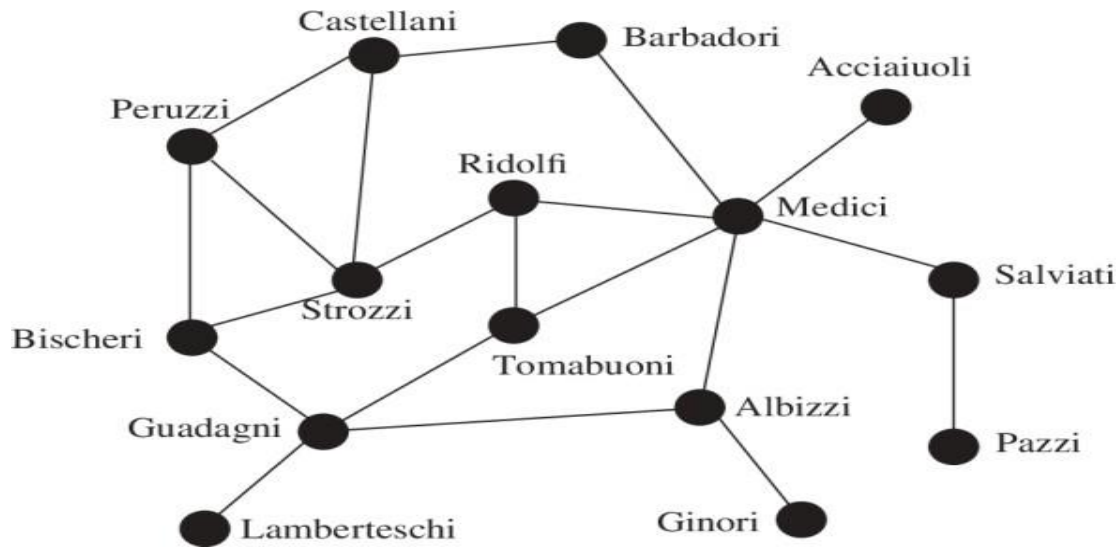


FIGURE 1.2. Florentine families network in the early 15th century – when the Medici family consolidated various powers in the society. Importance of family relations, calculated for each family as the count of connections/marriages with other families, provides an initial analysis for the rise of power of the Medici family.

1.2 WHY DO WE STUDY SOCIAL NETWORKS?

The idea of characterising societies as networks and analysing different properties is not new and goes back to the time when sociology methods were used to investigate individuals' properties to describe social structure.

Attributes limited to an individual, such as age, gender, occupation, and family name, could be quantitatively analysed and categorised to understand the household income and its effects on individuals relations. Traditional studies, however, lack consideration of social relations for analysing the networks, which is integral to the formation of structures in society. The study of interactions in society is vital for exploring underlying patterns and properties of networks, for example, how social interactions influence the network, flow of information in the network, individuals' roles, the formation of communities, and the evolution of networks over time. Modern SNA methods thus take individuals as well as their interaction information into account to model and study the networks.

For example, SNA methods are used to study the influence structure in small and large organisations where hierarchies in the organisation are examined to analyse information flow. Network dynamics is another critical area where modern SNA methods are

leveraged to understand evolution of various social networks over time. Graph clustering or community detection is another primary application of SNA, such as identifying cultures in the network and its effects on politics. The emergence of new data sources, such as social media sites, and recent advancements in computing have enabled researchers to collect and process network data at a large scale and apply SNA to a range of problems.

Summarising research topics by analysing citation networks, transportation networks for urban planning, and friendship recommendations in social media platforms are some examples of practical applications of SNA. Figure 1.3 presents the rise of social media networks over the last two decades. Consequently, the study of SNA is essential now more than ever to understand the patterns and evolution of extensive scale networks and their effect on society.

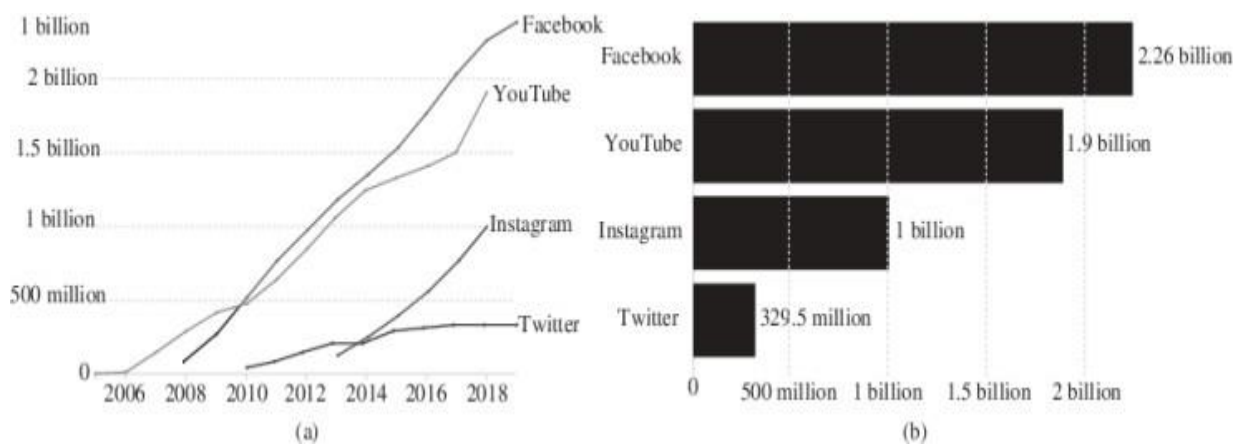


FIGURE 1.3. (a) Number of people using social media platforms, 2004 to 2019. (b) Number of people using social media platforms till 2018. Figure source: <http://ourworldindata.org/rise-of-social-media>. Statistics source: Statista (www.statista.com) and TNW (<http://thenextweb.com/tech/2019/06/11/most-popular-social-media-networks-year-animated>).

SNA is a multidisciplinary area involving domain knowledge from social, mathematical, statistical and computer science. With the underlying domain knowledge, over the years scientists have developed an extensive set of techniques and tools to model and analyse social relations and patterns present in the network. Some techniques help in understanding individuals' positions in the network such as identifying key individuals (e.g., influencers in a Twitter follower–followee network). Other methods explore the formation of substructures in the network to understand behavioural patterns such as the

presence of homophily subgroups in an election network. Visualisation techniques are another critical component that helps in gaining insights and explore hidden patterns. A key advantage of SNA in this emerging world of complex relations is that a similar set of techniques and tools can be leveraged for multiple domains and can impact multiple sectors.

1.3 APPLICATIONS OF SOCIAL NETWORK ANALYSIS

The domain of SNA can be viewed as a conglomeration of a wide range of fields from science and humanities, and find applications in even larger domain. Here, we discuss some of the areas where theories and models of SNA have found applications – some of them are readily perceivable, where others are somewhat non-trivial.

1.3.1 Healthcare

Combating Epidemics

Among many things that COVID-19 pandemic has taught the world include a terminology, social distance. The term “social” in the above terminology relates to the notion of social network. Application of the theory of SNA in epidemiology has a history for ages; network scientists have exploited social network models and different analysis techniques to model epidemics and to restrict their spread. With suitable model of the society, social interactions, and disease spread, it is often easier for the authority and healthcare personnel to plan combat strategies like identify superspreaders, mass quarantine of the super-spreaders, or planning partial or complete lockdown of a locality, restraining the spread of the disease even if the vaccine is not available. For example, the first recorded pandemic Black Death (Bubonic plague) claimed 75 to 200 million lives in Eurasia and parts of Africa during 1346–1353; another deadly 1918 flu pandemic (aka Spanish flu) affected ~500 millions and claimed lives between 17 and 100 million during 1918 to 1920 worldwide. However, equally deadly viral outbreaks such as 2002 SARS outbreak (8,096 cases, 774 deaths) or 2014 Ebola outbreak (28,646 cases, 11,323 deaths) have been confined to smaller geographic regions due to timely alerts from World Health Organisation (WHO) and other healthcare organisations even though no effective vaccines for these diseases could be made readily available at the moment. However, world has observed significantly less casualties now than before due to prompt actions.

Mass Vaccination Immunisation, also known as vaccination, is a popular/routine process in healthcare. Such processes have significant impact on public health and

epidemic prevention. It was possible to eradicate smallpox completely and poliomyelitis partially through proper and planned immunisation. Though mass immunisation, that is vaccination to the whole population, is a common practice, it has challenges in terms of money, manpower, and time. A better alternative would be the use of SNA to identify vulnerable part of the population and prioritise the immunisation accordingly. 2 The above strategy is effective when the supply of vaccine units is limited, and disease spread is fast like in the case of COVID-19. In addition, the same may lead faster to herd immunity and may save time, money, and lives, even when supply of vaccine is adequate.

1.3.2 Social Media and E-Commerce

Friend and Follow Recommendation

Social media expands with the gradual increase in connections between the nodes. Almost all of us have noticed automatic suggestions in our Facebook, Twitter, LinkedIn, or similar accounts about connecting a friend, joining a social group, liking or following a page, etc. We receive such personalised recommendations based on our previous activities and usage on the platform. We often find our old acquaintances or favorite pages from these suggestions. These recommendations are output of the application of the principles of network analysis to the underlying social networks.

Know Your Customers

Knowing the customers is important to any business enterprise. Large commercial chains or ecommerce platforms use network analysis principles to profile their customers based on their procurement patterns and suggest products and services accordingly. Shopping suggestions observed in the e-commerce apps in your mobile (Amazon, Walmart, or other) with mentions such as People like you buy, Frequently bought with this, or Frequently browsed, Trending are typical examples of this.

Recommendation and Viral Marketing

Word of mouth is considered to be a powerful means of advertisement in marketing. Many merchandises have adopted this approach to market their products/services, exploiting their potential customers as the seed. Example may include the schemes to incentivise the recommenders in e-commerce platforms. A classic example is the case of Gmail by Google Inc., which appeared in public on April 1, 2004 with a limited beta release as a private, invitation-only service, when email giants like Yahoo! Mail or MSN Hotmail used to dominate the market. Google incentivised the customers with additional

inbox storage space for successful invitation to “friends”. This strategy turned out to be highly successful, and Gmail turned into an email giant shortly afterwards.³ Georges Harik, a distinguished engineer at Google and Director of Googlettes at that time, exclaimed, Everyone wanted it even more. It was hailed as one of the best marketing decisions in tech history, but it was a little bit unintentional. We observed similar strategy in Dropbox referral program.⁴

1.3.3 Web and Cyberspace

Search Engine Optimisation

Searching the web has become an essential part of our lives nowadays. Search engines (Google, Bing, Baidu, Yahoo! Search, etc.) search the web and suggest numerous websites based on users' queries. Every search engine follows a ranking mechanism to order the search results. Network analysis techniques play a vital role in website ranking, and hence the search performances.

Malware Detection

Malware is a malicious software to harm one's computer. It is a collective name used to denote a wide variety of potentially harmful software classes, which include viruses, Trojan horses, spyware, ransomware, and so on. According to a report by AVTEST, more than 114 million new malware developed in calendar year 2020 as on August 26, 2020, and the same is expected to reach a figure of ~160 million by the end of 2020.⁶ Many of these are serious threats for the systems and for the users; however, detection and subsequent removal of the same is always a challenge due to the massive volume mentioned earlier. Since SNA is designed to efficiently deal with large volume of data, many scientific communities have modeled malware using a graph structure, and then used SNA methods on that graph. Graph instances are system call graphs using the system calls (Jang et al. 2015), malware similarity network by matching the malware features (Kim et al. 2018), etc.

Spam Detection

A spam, unlike malware, is an unwanted, unsolicited digital communication, usually sent as an email in bulk. Although the term usually refers to the email spams due to classical reasons, other similar abuses like mobile phone messaging spam, social spam, online classified ads spam, etc., fall in the category. Spams are annoying and often used as means to commit cybercrime; hence, its detection (and removal) is always warranted. However, the task is challenging due to the reasons more or similar to virus detection.

Scientific community exploited SNA methods for spam detection and filtering as well (DeBarr and Wechsler 2010)

1.3.4 Police and Military

Fighting Cybercrimes

With the growing popularity of the social media applications, the cybercrime has grown exponentially at the cost of the global economy. Typical examples include online fraud, fake message/news propagation, sharing (child) pornography, cyber bullying, etc. The gravity of the situation is enhanced due to the use of fake accounts/profiles during such activities. SNA may also come in rescue for such situations (Kirichenko et al. 2018, Krithiga and Ilavarasan 2019).

Fighting Terrorism

Neutralising the terror plans and capturing terrorists associated with the plan well before they succeed in executing their malicious plan has great impact on state affairs and on the society at large. However, the task is challenging due to numerous reasons: (a) terrorist locations often span neighbouring countries; (b) they often brainwash innocent people who have no past criminal records and deploy them as sleeper cell members; (c) they usually use untraceable devices for communication. SNA may play a crucial role to detect possible communication between terrorists and eventually nab them in action (Ressler 2006). It was often said by the security experts that the 9/11 terror attack would not have happened if SNA of present standard were available during that time.

Network-centric Warfare

The rising popularity of SNA has also influenced military doctrine. It is claimed that Saddam Hussein was captured from his hideout exploiting network analysis techniques.⁷ It is also claimed that US Navy Seal Team Six assassinated Osama Bin Laden when his secret hiding location was tracked using SNA. Based on such ideas, a group of US military think-tank has given rise to a network-centric warfare principle using SNA methods (Wong-Jiru 2006; Council et al. 2007; Knoke 2013).

1.3.5 Scientific Research and Academic Collaboration

Finding impact of a published scholarly article in a research field, deriving the influence of a scholar in a research community, and identifying/establishing the prestige of the conference or journal of which the publication is a part are some highly warranted but critical tasks in a research discipline. Answering these and many more similar questions

effectively solve a great many conflicting situations that often appear to the academic research communities, such as selecting the best paper in a conference/journal, choosing the best candidate for the research project grant, awarding a prestigious life-time research award, and so on. Such choices are often found to flare up controversies in practice till date. However, it is worth noting that scientific entities may yield a number of social networks – citation networks, coauthorship network, and cocitation networks are some typical examples. Researchers suggested that mining these networks using SNA may yield significant insight about the field of research in general (Chakraborty et al. 2013, 2014, 2015a).

1.3.6 Miscellaneous

Computer-supported Collaborative Learning

Activities in which a group of learners interact and are engaged together using information and communication technology (ICT) in order to meet a common learning goal is the objective of computer-supported collaborative learning (CSCL). Extraction of insights regarding learner interactions is necessary to enhance the learning outcome, and the same is possible from the large amount of computer-generated data available in CSCL in the form of system log files, messages, etc. We can apply SNA techniques to find relationship between various actors in CSCL (human actors such as learners and teachers, or non-human actors such as classes, courses, and learning materials) that interact with each other. A detailed discussion regarding this kind of application of SNA may be found in Dado and Bodemer (2017).

Complex Project Management

Credentials of a business house rise with the successful completion of its projects and drops with failure. With the rising complexity of the projects, especially the construction projects, the chance of failure is getting increased day by day. A project is complex if it has a complex structure with a number of elements of varying type with inter dependencies between them, uncertain goals and methods, and is dynamic. Lee et al. (2018) reviewed the potential applications of social networking analysis techniques in managing and strategic planning of such complex projects.

1.4 PRELIMINARIES

1.4.1 Defining a Network

A network or a graph $G(V, E)$ is defined by a set of nodes/vertices/entities (V) and a set of edges/links/relations (E). Depending on the type of applications, we can add more features into a network, such as a node or a link is assigned a set of attributes or features, a timestamp can be associated with each node or edge indicating the creation or existence duration and more.

A simple network does not have any self-loop (an edge whose end nodes are the same) and parallel edges (multiple edges whose end nodes are the same). An edge can be directed (asymmetric, irreversible) or undirected (symmetric, reversible); however, one can think of an undirected edge as a combination of two directed edges in opposite directions.

A weight can also be assigned to an edge, indicating the strength of the edge. For example, a follower–followee network in Twitter is a directed and un-weighted network, whereas a Facebook friendship network is an undirected and unweighted network. A user–user Twitter reply network is a directed and weighted network where nodes are users, a link indicates if a user has replied to posts of another user, and the weight of the edge indicates the number of times such replies appeared in the past.

A network $G(V, E)$ is generally represented or stored using an adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$, a square matrix whose each element a_{ij} indicates an existence of an edge between node i and node j (weight of an edge $\langle i, j \rangle$ in case of weighted network).

The adjacency matrix of an undirected and unweighted network is a symmetric (0, 1) matrix with zeros on its diagonal since there is no self-loop in general. Similarly, an entry a_{ij} in the adjacency matrix of a directed network indicates an existence of an edge from node i to node j . Figure 1.4 shows toy examples of undirected, directed, unweighted, and weighted networks, whose adjacency matrices are shown below

$$A(G_1) = (a_{ij}) = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad A(G_2) = (a_{ij}) = \begin{pmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 \end{pmatrix}$$

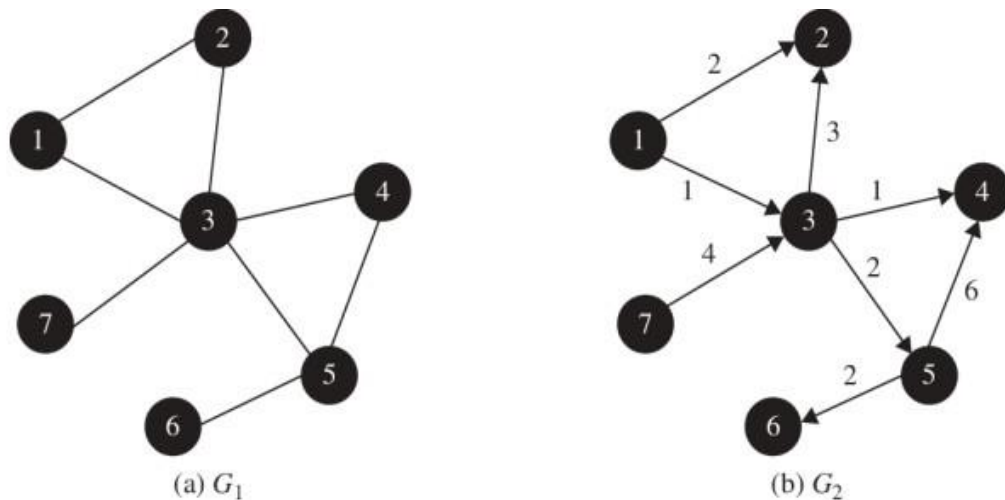


FIGURE 1.4. Toy examples of (a) Unweighted, undirected network G_1 and (b) Weighted, directed network G_2 .

It is worth noting that nodes 1 and 7 in G_2 do not have any incoming link, and nodes 2, 4, and 6 do not have outgoing link.. Node 3 in G_1 has many links attached, thus serving as an important node in the network.

The advantage of an adjacency matrix representation is that it is easier to implement and follow. Removing and querying an edge requires $O(1)$ time. The disadvantage is that it consumes more space, $O(|V|^2)$, irrespective of whether the network is sparse (contains less number of edges) or dense.

Moreover, adding a new node takes $O(|V|^2)$ time. Another way of representing a network is using adjacency list, which is a collection of unordered lists. Each list indicates a set of neighbors of a node in the network. An entry A_i of an array indicates

the list of nodes adjacent to node i . We can also represent a weighted network using the adjacency list. Figure 1.5 shows the adjacency lists of G_1 and G_2 in Figure 1.4.

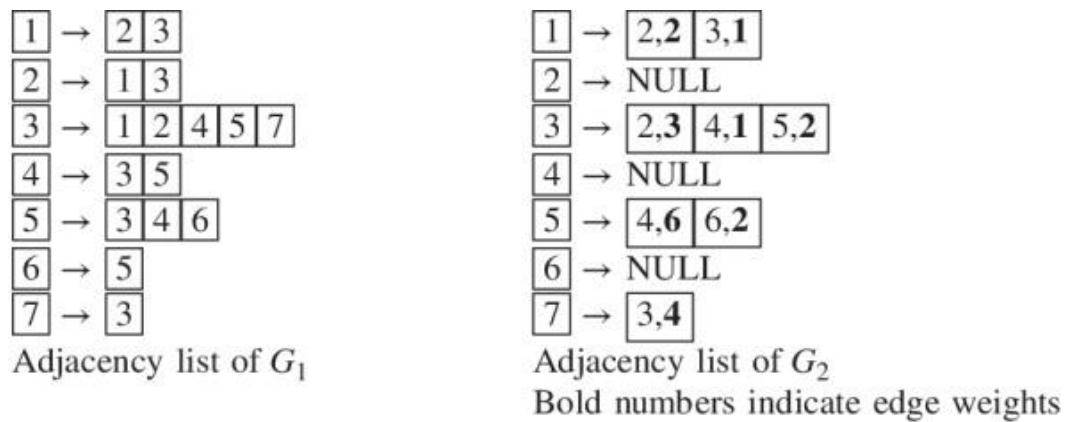


FIGURE 1.5. Examples of adjacency lists.

1.4.2 Types of Networks

Once we collect a dataset, our next task is to construct a network out of it. One can think of multiple ways to construct a network from a single dataset. For instance, from a Twitter dataset, we can form the following three networks:

- (a) nodes are users, and links denote who follows whom;
- (b) nodes are users, and links denote who replies to whom;
- (c) two set of nodes, one constituting users and other constituting tweets, and a link between a user and a tweet indicates if the user acted (posted, liked, replied, shared) on the tweet. The construction of the network depends on the task to be solved.

Depending on various aspects of the components of a network, we can construct a network in different ways. In general, we can broadly divide the networks into five types based on the topological structure as shown in Figure 1.6. In Figure 1.7, we describe each of these types with suitable real-world examples.

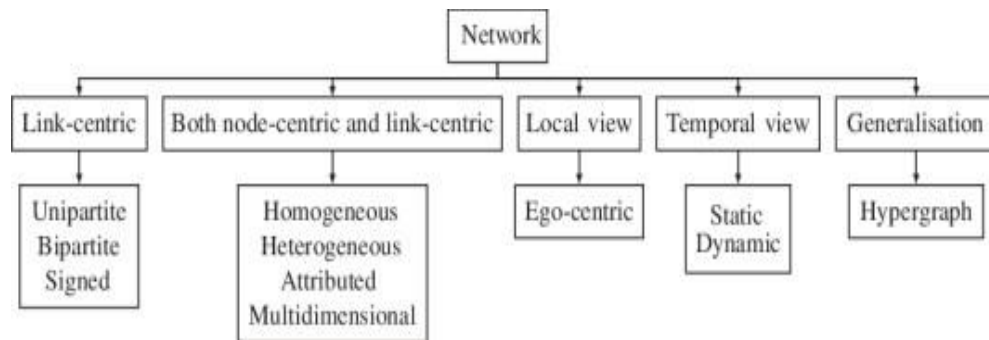


FIGURE 1.6. Categories of networks.

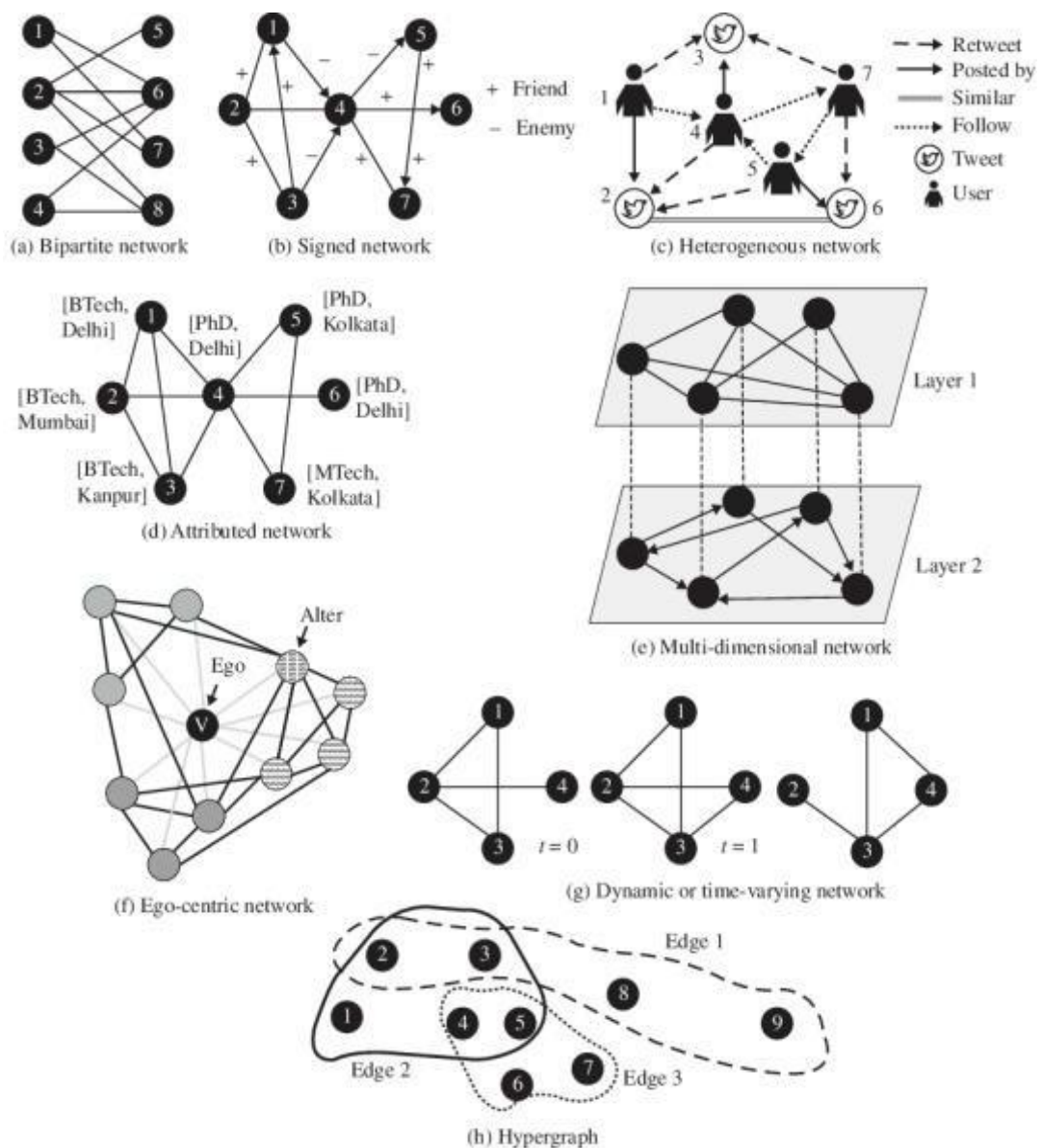


FIGURE 1.7. Toy examples of different types of networks.

1.4.3 Link-centric View

A link can be allowed to be formed between any pairs of nodes. However, in certain cases, due to the restriction on the network design, certain node pairs are not allowed to be connected. Sometimes, links appear in the network with additional information such as the polarity of relation between two nodes. Based on the design decision of the links, we can categorise networks into following three types.

1. Unipartite Network

A unipartite network $G(V, E)$ consists of a vertex set V and an edge set E , and there is no restriction on the formation of edges between nodes.

2. Bipartite Network

A bipartite network or bigraph $G(V, E)$ consists of a vertex set V which is divided into two disjoint and independent sets V_1 and V_2 , $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$, an edge set E where each edge $e \in E$ connects a vertex in V_1 to another vertex in V_2 . V_1 and V_2 are usually called the parts or partitions of the network. In other words, intra-part edges are not allowed; only inter-part edges can be formed.

For example, consider an e-commerce user–product network. One part consists of users, and other part consists of products; edges are formed based on who bought the products. It is not allowed to form edges between users as one user cannot buy another user! Similarly, edges cannot be formed between products.

Figure 1.7(a) shows a toy example of a bipartite network. Note that a bipartite network does not contain any odd-length cycles⁸ (think why?). Bipartite network plays an instrumental role in famous graph-colouring and edge colouring problems in graph theory. A generalisation of bipartite network is n -partite network with n independent node partitions and inter-partite edges.

3. Signed Network

A signed network $G(V, E, f)$ consists of set of nodes V , set of edges E and a function $f : E \rightarrow \{+, -\}$, which assigns each edge a positive or a negative sign. For example, let us consider Slashdot,⁹ a technology-related news website. It allows users to tag each other as friends or foes. Similarly, in Epinions¹⁰ network, members can decide whether to “trust” each other. All trust relations interact and form the network of trust. These are examples of explicit signed networks. One can also construct signed network from

implicit relations between nodes. For instance, if one scientific article negatively or positively cites (criticizes or appreciates) another article, we can add signs accordingly in the citation network.

Figure 1.7(b) shows a toy example of a signed network. Signed network is studied in the context of balance and status theory (Akiyama et al. 1981), which determines the stability or existence of certain types of structural patterns in a network. For instance, it is unlikely that there is an existence of a triangle with all three links marked by negative sign.

1.4.4 Combining Node-centric and Link-centric View

A network can be modeled by observing different meta-information of nodes and edges. Nodes and edges often arrive with additional attributes. Depending on how we incorporate these attributes into the construction of the network, we can further divide networks into following four categories.

1. Homogeneous Network

A homogeneous network $G(V, E)$ consists of a set of nodes V of same type and a set of edges E of same type. For instance, in a follower–followee network, all nodes are of same type (“user”), and all edges are of same type (“who-follows-whom”)

2. Heterogeneous Network

A heterogeneous network $G(V, E, f_v, f_e)$ consists of a set of nodes V and a set of edges E and two associated mapping functions, f_v and f_e , for nodes and edges. f_v maps a node to a node type, $f_v : V \rightarrow O$, where O is a set of object types.

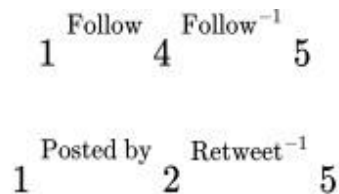
f_e maps an edge to an edge type, $f_e : E \rightarrow R$, where R is a set of object types. Each node belongs to one particular type as $f_v(v_i) \in O$.

Similarly, each edge is mapped to one category, $f_e(e_j) \in R$. If both the functions are available, we call the network both node and edge heterogeneous.

Note that if $|O| = |R| = 1$, the network is homogeneous. Let us take an example of a Twitter network. One can form a heterogeneous network by users and posts as two types of nodes, that is $O = \{\text{user}, \text{post}\}$.

Edges can be of different types as well such as “posted by” (user-post, directed), “followed by” (user-user, directed), “similar” (post-post, undirected), “retweet” (user-post, directed), as shown in Figure 1.7(c).

Recently, a new concept, called 'meta-path' has been proposed for heterogeneous network, which takes into account the edge heterogeneity in extraction of paths between nodes (Sun et al. 2011; Meng et al. 2015). Meta-path is a sequence of node types and edge types between a pair of nodes. In Figure 1.7(c), nodes 1 and 5 are connected through the following meta-paths:



Note that "Follow⁻¹" and "Retweet⁻¹" indicate the opposite directions of the edges labeled "Follow" and "Retweet", respectively.

3. Attributed Network An attributed network $G(V, E, F_v, f_e)$ consists of a set of nodes V , a set of edges E , and two associated mapping functions, f_v and f_e , for nodes and edges. f_v maps a node to an n dimensional attribute/feature space, $f_v : V \rightarrow \mathbb{R}^n$ such that for $v_i \in V$, $f_v(v_i) = \{v_{i1}, v_{i2}, \dots, v_{in}\}$ is the attribute vector of node v_i . Similarly, f_e maps an edge to an m dimensional attribute/feature space, $f_e : E \rightarrow \mathbb{R}^m$ such that for $e_j \in E$, $f_e(e_j) = \{e_{j1}, e_{j2}, \dots, e_{jm}\}$ is the attribute vector of edge e_j .

An attributed network may consist of only node attributes (no edge attributes) or vice versa. Let us take an example of a Facebook friendship network where nodes are users, and links indicate friendships as shown in Figure 1.7(d). It is a node-attributed network, and each node is associated with an attribute vector of size 2 – educational qualification (whose possible values are "BTech" "MTech" and "PhD") and location (whose possible values are "Kolkata" "Kanpur" "Delhi" and "Mumbai"). Note that in most real-world scenarios, only node attributes are available. One can derive the attribute vector of an edge by measuring the correlation/similarity/commonalty between the attribute vectors of two end nodes. For instance, in Figure 1.7(d), $f_v(\text{node } 1) = \{\text{BTech}, \text{Delhi}\}$, $f_v(\text{node } 4) = \{\text{PhD}, \text{Delhi}\}$. The attribute vector of edge $\langle 1, 2 \rangle$ can be derived by checking if the element-wise attribute values in $f_v(\text{node } 1)$ and $f_v(\text{node } 4)$ are same, that is $f_e(\langle 1, 4 \rangle) = \{|f_v(\text{node } 1)[1] \cap f_v(\text{node } 4)[1]|, |f_v(\text{node } 1)[2] \cap f_v(\text{node } 4)[2]|\} = \{0, 1\}$. Note that $f_v(\text{node } 1)[1] = \text{BTech}$.

1.4.4. Multidimensional Network

A multidimensional network is a multilayer network $G = (V, E, L)$ having $|L|$ layers. Each layer l corresponds to a dimension d of the multidimensional network. V denotes a set of N unique nodes.

A node $v_i \in V$ in layer $l \in L$ is denoted by v_{li} ($1 \leq i \leq N$; $1 \leq l \leq |L|$). Each edge $E_{li,j} \in E$ is a tuple (v_i, v_j, l) , representing an edge of type l from node v_{li} to node v_{lj} such that $v_i, v_j \in V$. Essentially, G consists of a total of $|V| \times |L|$ number of nodes.

For simplicity, we assume that all nodes in V are present in all the layers. If a node is absent in a layer l (i.e., no edge of type l connects to that node), we add the node as an isolated node in that layer. Let us take an example of a Twitter user–user network as shown in Figure 1.7(e).

We may think of two types of links – user–user follower links and user–user similarity links in terms of mutual interest. The former type will be directed (Layer 2), and the latter type will be undirected (Layer 1). Note that in the above case, each layer is node-homogeneous. Here, we only model edge heterogeneity in terms of two node-homogeneous layers.

Same nodes are connected across layers (as shown by the dotted lines in Figure 1.7(e)). In certain cases, each layer can also be node-heterogeneous. In this case, inter-layer edges indicate another type of relations. For example, a scientific publication network can be modeled as a two-layer network where one layer corresponds to coauthorship network, and other layer indicates paper–paper citation network.

Two layers are connected by edges, indicating who wrote which papers. This is another way to model a heterogeneous network. Readers are highly encouraged to read the two seminal papers (Kivelä et al. 2014; Boccaletti et al. 2014) on multilayer network.

1.4.5 Local View

Sometimes, a large network is difficult to analyse at a time. We then examine node-by-node and aggregate the measures to get a sense of the entire network. Such local view of a network plays an important role in certain tasks. For instance, in a phone call network, if you are interested to know how often one person calls her acquaintances and if there is any calling pattern as such, we would look at her personal call network for further inspections, rather than analysing the entire network.

Ego-centric Network

An ego-centric network $G(V, E, u)$ corresponding to a node $u \in V$ consists of a central node u (also known as “ego”), nodes to which u is directly connected by an edge (also known as “alters”) and their induced subgraph.¹² For example, in a Facebook friendship network, an ego-centric network can be formed around a user. The alters can be her friends in school days; some of them can be office collages, and so on. Figure 1.7(f) shows a toy example of the same.

1.4.6 Temporal View

Most of the real-world networks change over time. New nodes arrive at different times and attach with existing nodes through new edges. Old nodes also leave the network, which in turn remove existing edges. Therefore, nodes and links only carry certain information when they are active in the network. For instance, Facebook friendship network evolves drastically over time; however, it is still considered as a growing network as new users are mostly added to the network; it is less likely that existing users leave Facebook. In that sense, a paper–paper citation network is a strictly growing network. On the other hand, the frequency of change in the topological structure of a phone call network is very frequent – a phone call network consists of callers as nodes, and if one caller calls other, an edge is formed; an edge persists in the network as long as the call continues. To model such phenomenon, we need such a network structure that is dependent on the time. We call such network temporal, time-varying, or dynamic network.

Time-varying Network

A time-varying network $G(V, E)$ consists of a set of nodes V and a set of edges E where each edge $e_{i,j} \in E$ is represented by a tuple $e_{i,j} = \{v_i, v_j, t_{ij}\}$. Here, v_i and v_j are two end-points, and t_{ij} indicates the persistence duration (or “age”) of $e_{i,j}$.

Figure 1.7(g) shows a toy example of a time-varying network. Examples of such networks include person-to-person communication network (such as networks of email messages, mobile phone text messages, instant messages, and messages in online forums), cell biology network (such as molecular interaction network), neural and brain network, ecological network (interactions of species or other categories of organisms), and so on. Readers are highly encouraged to read a structured survey by Holme and Saramäki (2012). This is one way to represent a temporal network. This is called contact sequences where the entire network is represented as a set of contacts C of the form $(v_i,$

v_j, t_{ij}). Alternatively, it can be represented by an edge list E where each edge e_{ij} is a pair of nodes (v_i, v_j) and has a set of activation time of the edge $Te_{ij} = \{t_1, t_2, \dots\}$. Other two representations are as follows:

1. Interval network: If the interaction duration is non-negligible, Te_{ij} will denote a set of discrete intervals $Te_{ij} = \{(t_1, t'_1), (t_2, t'_2), \dots\}$.
2. Snapshots: The entire network is represented by a series of static networks. In this case, it is important to determine the interval gap after which we take the snapshot and the duration of each snapshot.

1.4.7 Generalised View

One can even generalise this structure by considering each edge as a hyperplane, which connects multiple nodes together. Such structure is called hypergraph.

Hypergraph A hypergraph $G(V, E)$ is defined by a set of nodes V and a set of edge or hyperedges E , where each hyperedge e connects multiple nodes. 1. 2. Two nodes are said to be adjacent if there exists an edge containing these nodes. Figure 1.7(h) shows an example of a hypergraph, containing three edges: Edge 1 connects $\{2, 3, 8, 9\}$, Edge 2 connects $\{1, 2, 3, 4, 5\}$, and Edge 3 connects $\{4, 5, 6, 7\}$. A coauthorship network can be modeled as a hypergraph where nodes are authors, and each paper represents a hyperedge, connecting its corresponding authors. If same set of authors wrote multiple papers, the count of papers can be used as the weight of the corresponding hyperedge. Readers are encouraged to go through the survey on hypergraph by Bretto (2013).

1.4.8 Popular Real-world Networks

Earlier studies in network science consider small networks; in most cases, they were handcrafted, for example, the famous Karate Club network (Zachary 1977), consisting of 34 members of a karate club representing nodes, and links between pairs of members are formed based on who interacted with whom outside the club (Figure 1.8). In 2002, Mark Newman took this network to the limelight by exploring two strong communities in the network (Girvan and Newman 2002). In the same study, Newman also experimented with American college football network, which is a network of American football games between Division IA colleges during regular season Fall 2000. There are 115 nodes, representing football teams (represented by the collages), and 613 edges indicate regular season games between the two teams they connect. Similar such small handcrafted networks have been curated and studied extensively as their topological structure is easy to analyse manually¹³. A few such popular networks which have gained

significant attention are briefly described here. Brief statistics of these networks are listed in Table 1.1 (Newman et al. 2006).

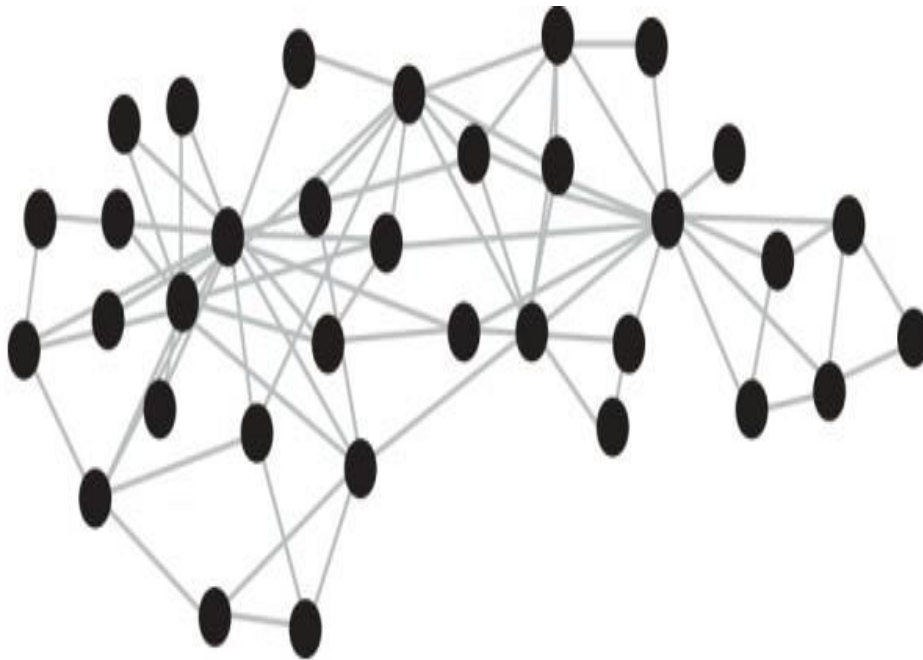


FIGURE 1.8. The famous Zachary's karate club network.

Table 1.1. A brief statistics of a few publicly available networks.

	Network	n	m	z	C	Reference(s)
Social	Film actors	449913	25516482	113.43	0.78	Watts and Strogatz (1998) ; Amaral et al. (2000)
	Company directors	7673	55392	14.44	0.88	Davis et al. (2001) ; Newman et al. (2001)
	Math coauthorship	253339	496489	3.92	0.34	De Castro and Grossman (1999) ; Gaskó et al. (2016)
	Physics coauthorship	52909	245300	9.27	0.56	Newman (2001, 2004)
	Biology coauthorship	1520251	11803064	15.53	0.60	Newman (2001, 2004)
	Telephone cell graph	47000000	80000000	3.16		Aiello et al. (2000, 2002)

	Network	n	m	z	C	Reference(s)
	Email messages	59912	86300	1.44	0.16	Ebel et al. (2002)
	Email address book	16881	57029	3.38	0.13	Newman et al. (2002)
	Student relationship	573	477	1.66	0.001	Bearman et al. (2004)
Information	WWW.nd.edu	269504	1497135	5.55	0.29	Barabási et al. (2000)
	WWW Altavista	203549046	2130000000	10.46		Broder et al. (2000)
	Citation network	783339	6716198	8.57		Egghe and Rousseau (1990); Redner (1998)
	Rogets Thesaurus	1022	5103	4.99	0.15	Knuth (1993)
	Word co-occurrence	460902	17000000	70.13	0.44	Cancho and Solé (2001)
Technological	Internet	10697	31992	5.98	0.39	Faloutsos et al. (1999); Chen et al. (2002)
	Power grid	4941	6594	2.67	0.080	Watts (2004); Leskovec et al. (2008)
	Train routes	587	19603	66.79	0.69	Latora and Marchiori (2002)
	Software packages	1439	1723	1.20	0.082	Newman (2003)
	Software classes	1377	2213	1.61	0.012	Valverde et al. (2002)
	Electronic circuits	24097	53248	4.34	0.030	Cancho et al. (2001)
	Peer-to-peer network	880	1296	1.47	0.011	Adamic et al. (2001); Ripeanu et al. (2002)
Biological	Metabolic network	765	3686	9.64	0.67	Jeong et al. (2000); Fell and Wagner (2000)
	Protein interactions	2115	2240	2.12	0.071	Vazquez et al. (2003); Szklarczyk et al. (2015)
	Marine food web	135	598	4.43	0.23	Dunne et al. (2004)
	Freshwater food web	92	997	10.84	0.087	Martinez (1991)
	Neural network	307	2359	7.68	0.28	Watts and Strogatz (1998)

Notations: n : number of nodes, m : number of edges, z : mean degree of nodes, C : mean clustering coefficient of nodes (see Chapter 2 for the definition of clustering coefficient). The table is adapted from (Newman et al. 2006).

Social Network

A social network is a social structure made up of a set of people or groups of people with some pattern of interconnections between them. These types of networks are very commonly seen in our day-to-day life. For example, a telephone call network keeps track of personal interaction via cellular connection among a group of people. In this network, each node represents a telephone number and each directed edge between two nodes represents a call. Aiello et al. (2000, 2002) were the first to analyse such telephone networks. They monitored all the calls made over the AT & T long-distance network in a single day. Presumably, even for just a single day, their network became enormous, having about 50 million nodes. Similar to a telephone call network, an email message network (Ebel et al. 2002) can be constructed to record emails sent among a set of users. Here a node and a directed edge represent an email id and one sent email, respectively. Email message network can also expand very quickly. Two other examples of relatively smaller social networks are film actor collaboration networks and academic coauthorship networks. In film actor collaboration network (Watts and Strogatz 1998; Amaral et al. 2000), every node represents a film actor, and two actors are considered to be connected if they have appeared in a film together. One classic example of such a network is neatly documented in the online IMDB Internet Movie Database.¹⁴ In academic coauthorship networks (Melin and Persson 1996; Newman 2004), authors of academic papers are linked if they have coauthored one or more papers. A follower–followee network, as shown in Figure 1.9(a), is also another example of social network.

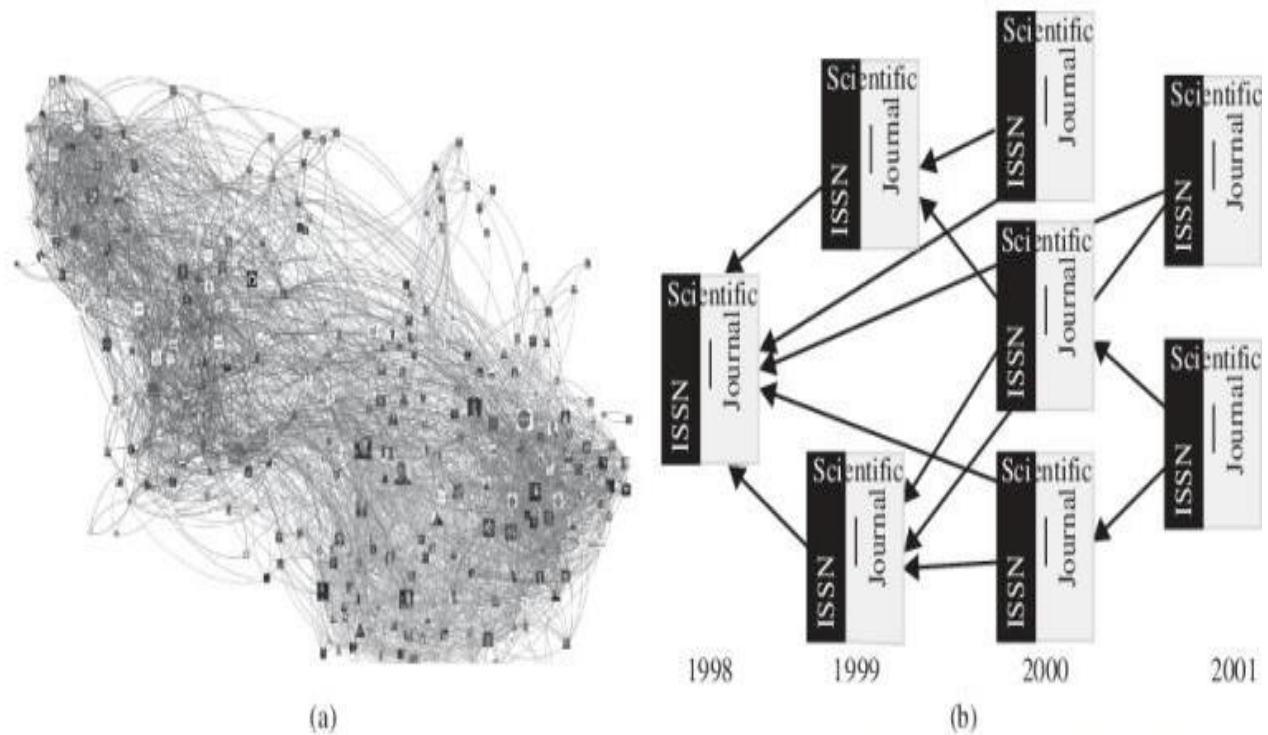


FIGURE 1.9. (a) A sample of Twitter follower-followee network (image source: <http://allthingsgraphed.com/2014/11/02/twitter-friends-network/>). (b) A toy scientific paper citation network.

Information Network

An information network or a knowledge network is a rich and dynamic real-world system that monitors how knowledge is shared, developed, and evolved over different sources. Two most popular information networks are citation networks and the World Wide Web (WWW). Citation network (Egghe and Rousseau 1990; Redner 1998) is a directed graph in which each node represents an academic paper, and each edge represents a citation from the citing publication to the cited publication. Figure 1.9(b) shows a toy example of a citation network. Since an academic paper can only cite previous papers, citation networks are acyclic, and all edges in this network point backward in time. It is also a growing network as nodes and edges can only be added; they will never be deleted. One can also construct an author-to-author citation network where nodes are authors, and each link indicates whether an author has cited papers of another author. The links can also be weighted depending on the number of times one author cites papers of another author.¹⁵ On the other hand, WWW (Kleinberg et al. 1999; Barabási et al. 2000; Barabási et al. 2002) is another information network which allows

a web page to be connected to other web pages by hypertext links, enabling the user to search for information by moving from one web page to another. In contrast to the citation network, the WWW network may be cyclic in nature, as there is no constraint of web pages to contain hyperlinks of other pages.

Biological Network

Biological systems are often represented as networks. Different types of biological systems result in different network characteristics in terms of connectivity, complexity, and structure of nodes and edges. Some popular examples of biological networks are metabolic networks, protein–protein interaction networks, genetic regulatory networks, cell signaling networks, neural networks, and the food web. A metabolic network is the complete representation of all the chemical reactions of metabolism, the metabolic pathways, as well as the regulatory interactions that guide these reactions. In this network, the substrates and products are represented by vertices, and an undirected edge is drawn between substrate and product if a known metabolic reaction exists to act on the given substrate and produces the given product. Jeong et al. (2000); Fell and Wagner (2000); Wagner and Fell (2001) conducted extensive research to explore the statistical properties of metabolic networks. Similarly, protein–protein interaction (PPI) networks (Vazquez et al. 2003; Hakes et al. 2008; Szklarczyk et al. 2015) are the complete mathematical representations of the physical contacts between proteins in the cell. In these networks, an undirected edge is drawn between two proteins if they are known to interact with each other. Genetic regulatory network (Guelzim et al. 2002; Warren and Ten Wolde 2004) monitors the interaction of molecular regulators with each other. Cell signaling networks (Eungdamrong and Iyengar 2004; Morris et al. 2010) outline cell-to-cell communication to govern and coordinate multiple cell actions.

Another much-studied example of a biological network is the food web which represents the natural interconnection of the food chain to interpret the predators and prey relationship in an ecological community. Generally, in this network, nodes represent various species in an ecosystem, and a directed edge from A to B indicates that species A preys on species B. However, some ecologists, who tend to think in terms of energy or carbon flows through food webs, draw this direction in the other way round. In recent years, Camacho et al. (2002); Petchey et al. (2008); Layman et al. (2012) performed extensive studies to understand the statistical properties of the topology and structure of food webs.

Technological Network

Technological networks are typically man-made networks for the distribution and collection of information and commodity. Some extremely popular examples of technological networks are electric power grids, networks of airline routes, roads, railways and electronic circuits, delivery networks of post-office, and the Internet. In high-voltage electric power grids, the generating stations and electric substations are represented by nodes and the high voltage transmission lines by edges. Statistical properties of such power grids were studied by Watts and Strogatz (1998); Watts (2004); Amaral et al. (2000); Leskovec et al. (2008). For road, rail, subway, and airlines routes (Kalapala et al. 2003; Latora and Marchiori 2002), different cities, stations, and airports can be seen as nodes, whereas the roads and railway tracks are the edges. The Internet (Faloutsos et al. 1999; Chen et al. 2002), which is the electronic communication system that interconnects all computer networks, devices, and organisational facilities around the world, is an enormous example of a man-made technological network. Some examples of natural communication networks are river networks of a continent and vascular networks in animal and plant bodies.

Language Network

Language network is one of the earliest human networks. A language network is formed with a group of people who share the same language – be it some symbolic language, which they know early on from their infancy, or some spoken language which they adopt in their later life. Language network can be enormous or tiny; for example, the English network is vast and expansive; but in contrast, the Aikana16 language network is really tiny. Language networks are fundamental to human history – it is the means by which ideas are shared, evolved, and built upon. Language networks can grow or sink over time. In some cases, people outside of a language network join it to learn that language, particularly to facilitate trade and exchange of different types. On the other hand, people of a particular language can leave the network. Again, Yiddish is an example of semiextinct language which is recently facing a revival among the second or third generation of American Jews. So there can be this waxing and waning of language networks. Statistical studies on the structure, function, and evolution of language networks have been made by many researchers, including Solé et al. (2010); Friederici and Gierhan (2013); Seoane and Solé (2018).

1.5 THREE LEVELS OF SOCIAL NETWORK ANALYSIS

In general, social networks follow self-organising mechanism where the entire structure emerges from local interactions between pairs of entities (nodes) which otherwise are disordered. The emerging process is spontaneous, or sometimes it is driven by underlying latent dynamics, not by any control of external agents. Social networks follow certain structure; they are not random collections of nodes and edges. They are dynamic in nature. The dynamic behaviour of a network is often the result of a series of evolutionary steps followed by nodes and their groups. Networks grow from bottom-to-top where nodes first interact at local level and move to the top or global level. Therefore, to understand global patterns, one should delve deeper into the local and/or semi-local level interactions. Moreover, due to the sheer volume of modern real-world networks, it is impossible to extract the global characteristics of a network in one go since we have limited access to the computational resources to handle such huge networks. However, the granular details of local interactions may be lost as we move from bottom to top. Therefore, there is a trade-off between how much local information we need for the analysis and what scale of understanding one should require about a network. Hence, it is a common practice to analyse a network into three levels: microscopic, mesoscopic and macroscopic.

Microscopic Analysis

At the microscopic level, we begin by analysing how a pair of nodes interacts and gradually trace the interactions at the group level or subgraph level. At the dyadic level, we observe interaction patterns among two nodes to examine several properties such as homophily, reciprocity, social equality, mutuality and derive global statistics of the network such as assortativity, mixing coefficient. A one level higher, we examine the interactions among three nodes. We call it triadic level interaction. This level of analysis reveals local interaction properties such as clustering coefficient. The concept of triadic closure has led to the formulation of clustering coefficient and local bridges. The theory of balance and status in signed networks has also emerged from the triadic level interactions. One may also magnify the interactions of an ego to its alters and study how alters form ego-centric circles (Chakraborty et al. 2015b) in the ego network.

Mesoscopic Analysis

Mesoscopic analysis is an intermediary between microscopic and macroscopic analyses, which mostly deals with a subset of the entire population. Several different substructures of the network play important roles at this level. In particular, communities or network

clusters act as major points of interest. Communities are the sets of nodes which are formed due to frequent interactions among homogeneous nodes in the network. Thus within a community, nodes exhibit a separate dynamical behaviour, whereas across communities, these behaviours may differ. Communities are known to present the functionalities of different organisational units of a system by exploring the intra-group and inter-group relations, especially in large organisations with multiple diverse branches. The second important mesoscopic level structure is network motifs. Motifs are subgraphs that repeat themselves frequently within a network or across networks. Figure 1.10 shows examples of motifs. Motifs are shown to be highly effective in capturing functional properties of a network, particularly biological network (Masoudi-Nejad et al. 2012). Each network motif can carry out defined information-processing functions, which have been studied mostly for *Escherichia coli* (Alon 2007).

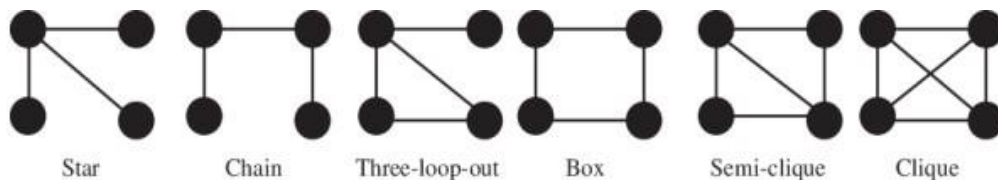


FIGURE 1.10. Undirected motifs with size 4 and their names.

Macroscopic Analysis

At macroscopic level, we deal with the entire network as a whole and try to understand the microlevel dynamics by exploring the overall graph property. For instance, the properties like connectedness, diameter and average path length between pairs of nodes, degree distribution, edge density etc., although being network-level properties, often explain the underlying structure and the interactions of nodes. Let us assume that the diameter of a network is very small. We can anticipate that the network may look like a star or a clique (a complete graph). In addition, if we come to know that the overall edge density is very high, we can anticipate that the network looks like a clique. The estimation of low-level interactions between nodes by looking at the high-level properties of the network is called as top-down approach.

1.6 NETWORK BASICS

The concept of a social network is based on the theory of graphs. A network (or graph) consists of a set of entities, called nodes and a set of links between these entities, called

edges. We often use the terms “network” and “graph” interchangeably to refer to the same concept. Mathematically, a network G is represented by an ordered pair (V, E) , where V is the set of nodes or vertices or entities, and E is the set of edges or links or relations. There are a variety of network types depending on the nature of the edges between the nodes. Based on applications, edges can be directed or undirected. Edges may also be associated with a real number, called the weight of the edge. When edges do not have any direction, we call the resultant network an undirected network or simply, a network. When edges are directed from one node to another, we call the resultant network a directed network. In a similar manner, if the edges of the network are not associated with weights, we call it an unweighted network; if they are associated with weight values, we call the network a weighted network. The network in Figure 2.3(a) is undirected and unweighted; the same in Figure 2.3(b) is undirected and weighted; and the network in Figure 2.3(c) is directed and unweighted.

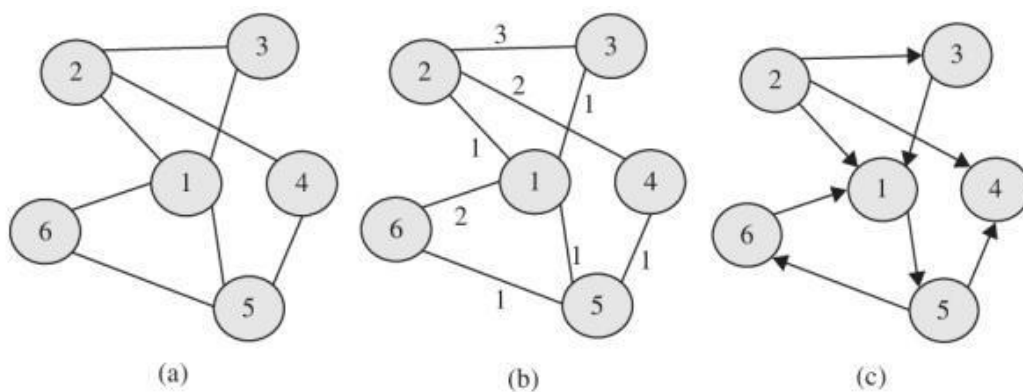


FIGURE 2.3. Examples of networks – (a) undirected and unweighted, (b) undirected and weighted, and (c) directed and unweighted.

by a network G , we mean an undirected and unweighted network. As we observe from the preceding examples, a node in a network may or may not have direct connection (edge) with another node of the network. The following subsection introduces an important concept involving the number of edges associated with a node in the network, called the degree of a node.

1.6.1 Degree and Degree Distribution

Degree

Degree of a node v , denoted by $\deg(v)$, in an undirected and unweighted network G , is the number of other nodes in the network to which v has an edge. Degree is typically defined for unweighted and undirected networks. Hence, alternatively, we state that, in an unweighted and undirected network, the degree of a node is simply the number of edges incident to that node. For example, the number of friends you have in your Facebook account is your degree in the Facebook social network. If two nodes in an undirected network are linked by an edge, they are called neighbours. Therefore, the degree of a node in an undirected and unweighted network is the number of neighbours of the node.

Example 2.1

Find the degrees of all the nodes of the network shown in Figure 2.3(a).

Solution The degrees of six nodes 1, 2, 3, 4, 5, 6 are 4, 3, 2, 2, 3, 2, respectively.

Many real-world applications require networks with weighted edges. For example, let us consider the instance of a typical road network where nodes represent the metro cities, and edges are the roads connecting pairs of cities. Here, we may represent the (road) distance between a pair of cities as the weight of the corresponding edge in the network. **Weighted Degree** Weighted degree of a node is the sum of the edge weights of the corresponding edges attached to the node.

Example 2.2

Find the weighted degrees of all the nodes of the weighted network shown in Figure 2.3(b).

Solution The weighted degree of node 1 is the sum of edge weights 1, 2, 1, 1, which is 5. In the same fashion, the weighted degrees of nodes 2, 3, 4, 5, 6 are 6, 4, 3, 3, 3, respectively.

Directed networks are common in applications where the linkages between a pair of entities are asymmetric. For example, the social media platforms such as Twitter, wherein a follower–followee relation is asymmetric in the sense that node A follows node B does not necessarily imply that B follows A, and viceversa. Such a network can be portrayed as a directed network.

We now extend the concept of degree of a node for directed networks as well. Due to the asymmetric nature of the directed edges, in a directed network, we define two types of degrees of a node: In-degree and Out-degree.

In-degree In-degree of a node in a directed network is defined as the number of incoming edges to the node.

Out-degree Out-degree of a node in a directed network is defined as the number of outgoing edges from the node.

Example 2.3

Find in-degrees and out-degrees of all the nodes of the network shown in Figure 2.3(c).

Solution The in-degrees of nodes 1, 2, 3, 4, 5, 6 are 3, 0, 1, 2, 1, 1, respectively. Similarly, the out-degrees of the same are 1, 3, 1, 0, 2, 1, respectively.

We also can define the degree of a node in a directed network as the sum of the in-degree and the out-degree of the node. However, this particular notion has limited significance and hence is rarely used in practice.

Observation 2.1.

For an undirected and unweighted network, the sum of the degrees of all the nodes in a network is twice the number of edges in the network. This is because in deriving the degree of a node, we are counting the number of edges incident on that node. We also note that every edge in the network is incident to exactly two nodes in the network. So, during the calculation of the sum of the degrees of all the nodes in the network, we are counting every edge twice, one each for degree of the node it is incident to. This particular observation, though simple, has significant impact on the theory of networks, and hence worth remembering.

Observation 2.2.

The observation that follows immediately from the Observation 2.1 is that the number of odd degree nodes in an undirected network is always even. This follows from the following two facts: (a) sum of any number (odd or even) of even numbers is even; (b) sum of an odd number of odd numbers is odd, that is, $\sum_{i=1}^n a_i$ is odd when n is odd and all a_i is odd.

Observation 2.3.

Extending the result towards directed networks, we may find that the sum of indegrees of all the nodes in a directed network is same as the sum of outdegrees of all the nodes in the network. This follows immediately from the observation that a directed edge can contribute once each for the in-degree and the out-degree calculations. We have noticed that finding the degree of a node in a network is nothing but counting the edges incident on the same. As we have done in the preceding examples, when finding the degrees of all the nodes in a network, one may think of listing them as a sequence of integer-valued degrees of nodes. This approach may work for small networks; but, for large networks, we require a sophisticated means of representation, as follows.

Degree Distribution

Degree distribution of a network is the (probability) distribution of the degrees of nodes over the whole network.

Suppose a network $G(V, E)$ has $N = |V|$ nodes. Let us also suppose that $P(k)$ denotes the probability that a randomly chosen node from the network has degree k . If N_k is the fraction of nodes having degree k in the network, then $P(k) = N_k/N$. The set of all pairs $(k, P(k))$ presents the degree distribution of the network G . One convenient way of representing the degree distribution of a network is to plot the distribution as a histogram, where x-axis represents degree k , and y-axis represents the corresponding $P(k)$ values.

Note that $\sum_k P(k) = 1$, $0 \leq P(k) \leq 1$, and $P(k)$ is a discrete distribution. Therefore, the average degree of nodes $\langle k \rangle$ can be written as $\langle k \rangle = \sum_k kP(k)$.

The same information can also be presented using *cumulative degree distribution* (CDD), indicating the fraction of nodes with degree smaller than k , that is, $C_k = \frac{\sum_{k' < k} N_{k'}}{N}$. Even the *complementary cumulative degree distribution* (CCDD) is also often used, indicating the fraction of nodes with degree greater than or equal to k , that is, $CC_k = 1 - C_k$.

Example 2.4

Find the degree distribution of the network shown in Figure 2.3(a). Solution For the network shown in Figure 2.3(a), the number of nodes, $N = 6$. Then, the degree distribution of the network with degrees, $k = \{1, 2, 3, 4\}$ are given by $P(1)$, $P(2)$, $P(3)$, $P(4)$ as follows.

$$\begin{aligned}
 N_1 = 0 &\implies P(1) = \frac{0}{6} = 0.0 \\
 N_2 = 3 &\implies P(2) = \frac{3}{6} = 0.5 \\
 N_3 = 2 &\implies P(3) = \frac{2}{6} = 0.33 \\
 N_4 = 1 &\implies P(4) = \frac{1}{6} = 0.17
 \end{aligned}$$

Example 2.5

Draw the degree distribution of the network shown in [Figure 2.3\(a\)](#).

Solution

The degree distribution of the network shown in [Figure 2.3\(a\)](#) is represented graphically in [Figure 2.4\(a\)](#). [Figure 2.4\(b\)](#) portrays CDD and CCDD of the same network – you may notice that one is the mirror image of other (think why?).

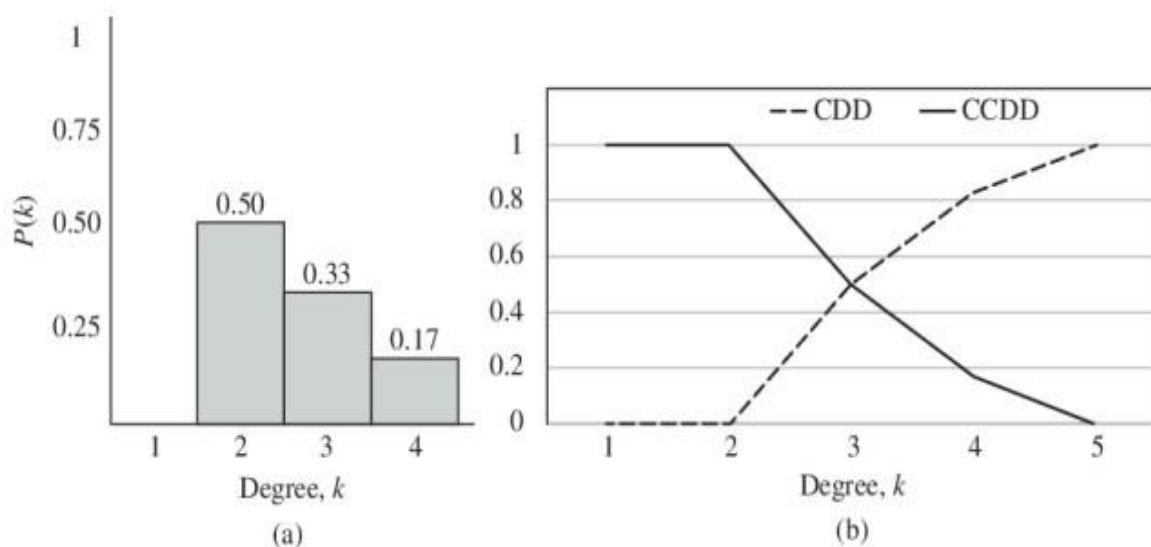


FIGURE 2.4. (a) The degree distribution, and (b) CCD and CCDD of the network shown in [Figure 2.3\(a\)](#).

degree distribution at the asymptotic level, that is, $P(k) \sim k^{-\gamma}$, where γ is a parameter that ranges between 2 and 3 for most of the networks. Therefore, in the log–log scale, the degree distribution looks like a straight line (as $\log P(k) \sim -\gamma \log k$).

1.6.2 Paths

The next notion that we are interested in a network is its connectivity, which tells us how nodes are connected via a sequence of edges.

Adjacent Nodes

A pair of nodes v_i and v_j in a network $G(V, E)$ are called adjacent if they are linked by an edge, e . In other words, v_i and v_j are the two endpoints of the edge e in network G .

Incident Edges

A pair of edges e_1 and e_2 in a network $G(V, E)$ are called incident if they have a common end node.

Walk

A walk in a network is a sequence of nodes from the network such that every consecutive node pair in the sequence is adjacent. A walk can pass through the same node or edge more than once.

Walk Length

The length of a walk is the number of edges in the walk. A closed walk is simply a walk which starts and ends at the same node. Correspondingly, an open walk is a walk which starts and ends at different nodes.

Example 2.6

Show some examples of walks from the network shown in Figure 2.3(a). Solution Some examples of walks from the network shown in Figure 2.3(a) are as follows:

Example 2.7

What is the length of the walk in the network of Figure 2.3(a)? Solution Since there are four edges in the mentioned walk, its length is 4.

Trail

A trail in a network is a walk with no edge repeated.

Path

A path in a network is an open trail with no node repeated. In other words, a path in a network is a walk in which all the nodes and edges are distinct.

Cycle

A cycle in a network is a closed path. In other words, a cycle in a network is a closed walk in which (a) all the edges are distinct, (b) all the nodes, leaving the start and end nodes, are also distinct.

Example 2.8

Example 2.8 Show some examples of trails, paths, and cycles in the network shown in Figure 2.3(a). Solution Some examples of trails, paths, and cycles from the network shown in Figure 2.3(a) are as following:

Trails	Paths	Cycles
$1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \rightarrow 5$	$1 \rightarrow 2 \rightarrow 3$	$1 \rightarrow 2 \rightarrow 3 \rightarrow 1$
$1 \rightarrow 2 \rightarrow 3 \rightarrow 1$	$1 \rightarrow 5 \rightarrow 4 \rightarrow 2$	$1 \rightarrow 5 \rightarrow 4 \rightarrow 2 \rightarrow 1$
$1 \rightarrow 5 \rightarrow 4 \rightarrow 2$		$1 \rightarrow 5 \rightarrow 6 \rightarrow 1$

Observation 2.4.

We can extend all the above concepts for directed networks by maintaining the direction of the edges. In particular, in a directed network, a path can only follow the direction of arrow of the corresponding directed edges.

For example, for the directed network shown in Figure 2.3(c), $1 \rightarrow 5 \rightarrow 6 \rightarrow 1$ is a valid path; but, $1 \rightarrow 5 \rightarrow 6 \rightarrow 1 \rightarrow 2$ is not.

Distance

The distance between any two nodes v_i and v_j , denoted by $d_{i,j}$, in a network G , is the length of the shortest path between v_i and v_j . If there is no path between v_i and v_j , conventionally the distance between v_i and v_j is considered to be infinity.

Example 2.9

For the network shown in Figure 2.3(a), the distance between nodes 1 and 4 is 2 as the length of the shortest path ($1 \rightarrow 2 \rightarrow 4$, or $1 \rightarrow 5 \rightarrow 4$) is 2. Note that, there exist multiple

longer paths connecting 1 and 4 ($1 \rightarrow 6 \rightarrow 5 \rightarrow 4$ or $1 \rightarrow 3 \rightarrow 2 \rightarrow 4$), which will not be considered while calculating the distance.

Diameter

The diameter of a network G , denoted by $\text{Dia}(G)$, is defined as the length of the longest of all the calculated shortest paths between all possible pairs of nodes in the network. Trails $1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \rightarrow 5$ $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ $1 \rightarrow 5 \rightarrow 4 \rightarrow 2$ Paths $1 \rightarrow 2 \rightarrow 3$ $1 \rightarrow 5 \rightarrow 4 \rightarrow 2$ Cycles $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ $1 \rightarrow 5 \rightarrow 4 \rightarrow 2 \rightarrow 1$ $1 \rightarrow 5 \rightarrow 6 \rightarrow 1$ (2.1)

To measure the diameter for a network, we make a list of lengths of the shortest path distances between all pairs of nodes in the network. The maximum value from this list would be the diameter of the network.

Example 2.10

Find the diameter of the network shown in [Figure 2.3\(a\)](#).

Solution

Let us first list the distance $d_{i,j}$ between every pair of nodes (i,j) of the network, as follows:

$$\begin{aligned} d_{1,2} &= 1; d_{1,3} = 1; d_{1,4} = 2; \\ d_{1,5} &= 1; d_{1,6} = 1; d_{2,3} = 1; \\ d_{2,4} &= 1; d_{2,5} = 2; d_{2,6} = 2; \\ d_{3,4} &= 2; d_{3,5} = 2; d_{3,6} = 2; \\ d_{4,5} &= 1; d_{4,6} = 2; d_{5,6} = 1. \end{aligned}$$

Therefore, the diameter of the network is 2.

Average Path Length

The average path length of a network G , denoted by l_G , is defined as the average number of steps along the shortest paths for all possible pairs of nodes in the network. In other words, it may be defined as the sum of lengths of all the shortest paths divided by the maximum possible edges between all the nodes of the network.

Mathematically,

$$l_G = \frac{1}{2 \times E_{max}} \sum_{i \neq j} d_{i,j} \quad (2.1)$$

where $d_{i,j}$ is the distance (or the length of the shortest path) between nodes v_i and v_j in G , and E_{max} is the maximum possible edges between all the nodes in G . For a network with n nodes,

$$E_{max} = {}^nC_2 = \frac{n(n-1)}{2}$$

Example 2.11

Find the average path length of the network shown in [Figure 2.3\(a\)](#).

Solution

Let us first list the distance between every pair of nodes of the network shown in [Figure 2.3\(a\)](#).

$$d_{1,2} = d_{2,1} = 1; d_{1,3} = d_{3,1} = 1; d_{1,4} = d_{4,1} = 2; d_{1,5} = d_{5,1} = 1; d_{1,6} = d_{6,1} = 1;$$

$$d_{2,3} = d_{3,2} = 1; d_{2,4} = d_{4,2} = 1; d_{2,5} = d_{5,2} = 2; d_{2,6} = d_{6,2} = 2; d_{3,4} = d_{4,3} = 2;$$

$$d_{3,5} = d_{5,3} = 2; d_{3,6} = d_{6,3} = 2; d_{4,5} = d_{5,4} = 1; d_{4,6} = d_{6,4} = 2; d_{5,6} = d_{6,5} = 1.$$

Summing all over the above = $22 \times 2 = 44$, $E_{max} = {}^6C_2 = 15$.

Therefore, the average path length would be = $\frac{44}{2 \times 15} = 1.47$.

Density The density of a network is defined as the ratio of the number of actual edges in the network to the total number of possible edges in the network. Mathematically, for a network $G(V, E)$, the density, denoted by $\rho(G)$, can be written as,

$$\rho(G) = \frac{|E|}{E_{max}} = \frac{|E|}{|V|C_2} \quad (2.2)$$

where $|E|$ and $|V|$ are the number of edges and nodes in the network, respectively. Density is typically defined for an undirected network.

Example 2.12

Find the density of the network shown in Figure 2.3(a).

Solution

For the network shown in Figure 2.3(a): $|V| = 6$, $|E| = 8$. Then,

$$|V|C_2 = {}^6C_2 = 15$$

Therefore, the density of the network = $\frac{8}{15} = 0.533$.

Observation 2.5.

For the network shown in Figure 2.3(a), the maximum possible value of $\rho(G)$ is 1 (why?). This would only happen if, between every pair of nodes in G , there is an edge. The network shown in Figure 2.5 is such a network, and hence, it has density 1. We call such a network a *complete network* or a *clique*.

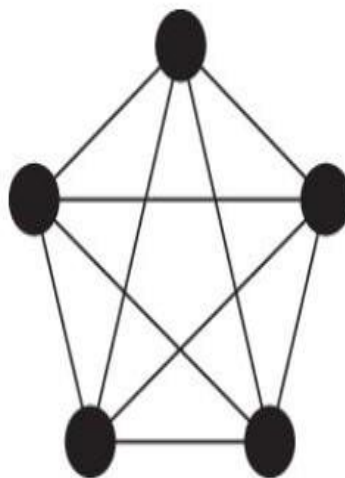


FIGURE 2.5. An example of a complete network (clique).

1.6.3 Clustering Coefficient

In such a network, we often find tightly knit groups here and there with less dense ties away from these groups. The same phenomenon is observed in most of the networks and heavily on social networks.

The network measure, called clustering coefficient, we introduce here can characterise the above phenomenon. Clustering coefficient measures (2.3) (2.4) the local connectivity of a node by considering the connectivity between its immediate neighbours.

There are two ways to measure clustering coefficient, namely, local clustering coefficient (LCC) and global clustering coefficient (GCC).

Local Clustering Coefficient

The local clustering coefficient (LCC) of a node is the ratio of the number of edges to the maximum possible edges between the neighbours of the node. Formally, in a network $G(V, E)$, the LCC of node $v_i \in V$ is defined as follows:

$$C_i = \frac{\text{Number of edges between neighbours of } v_i}{\text{Number of maximum possible edges between neighbours of } v_i}$$

Global clustering coefficient (GCC) is based upon triplets of nodes. A(2,3) triplet consists of three nodes connected by either two (open) or three (closed) undirected edges (see [Figures 2.6\(a\)](#) and [2.6\(b\)](#)). A triangle network with three nodes 1, 2, 3 as in [Figure 2.6\(a\)](#) has three closed triplets enclosed: (1,2,3), (2,3,1), and (3,1,2).

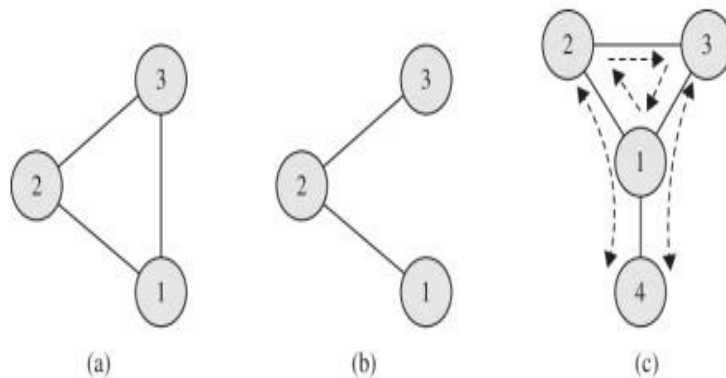


FIGURE 2.6. Examples of networks with (a) closed triplet, (b) open triplet, (c) both open and closed triplets.

Global Clustering Coefficient

The global clustering coefficient (GCC) is the ratio of the number of closed triplets over the total number of triplets (both open and closed). Mathematically, the GCC, denoted by C , of a network G is defined as,

$$C = \frac{\text{Number of closed triplets in } G}{\text{Total number of triplets in } G} \quad (2.4)$$

Example 2.13

Find the LCC of node 1 in the network shown in [Figure 2.3\(a\)](#).

Solution

Node 1 has four neighbours, {2, 3, 5, 6}. There are two edges between these neighbours given by $\langle 2,3 \rangle, \langle 5,6 \rangle$. There can be a maximum of

${}^4C_2 = \frac{4 \times 3}{2} = 6$ edges between these neighbours. Therefore, the LCC is $\frac{2}{6} = 0.33$.

Example 2.14

Find the GCC of the network shown in [Figure 2.6\(c\)](#).

Solution

In [Figure 2.6\(c\)](#), there are three closed triplets enclosed between nodes 1, 2, and 3. In addition to this, there are two (quite deceiving) open triplets – (2,1,4) and (3,1,4). Therefore, there are a total of five triplets. Hence, the

GCC of the network is $= \frac{3}{3+2} = 0.6$.

The clustering coefficient has been shown to have applications in various domains of network science. For example, in neuroscience, Masuda et al. (2018) utilised clustering coefficient in measuring correlations in function– structure associations in the human brain. In social analytics, clustering coefficient is an important feature utilised for recommendation and finding echo chambers (Barberá et al. 2015). Han and Mao (2018) showed that clustering coefficient could play a significant role in network reconstruction.

1.6.4 Connected Components

Connectedness, in some sense, establishes the presence of such nodes in the network.

Connected Network

In an undirected network G , two nodes v_i and v_j are said to be connected, if there exists a finite path between v_i and v_j . An entire network G is said to be connected if every node pair in G is connected. A network may consist of smaller connected subnetworks. These connected subnetworks are called as components of the network. In the real-world networks such as a social network or a brain network, usually there is a giant component (consuming major chunk of nodes) and many smaller components. Figure 2.7 shows an example of the same. Here nodes 1, 2, 3, 7, 9 form the giant component.

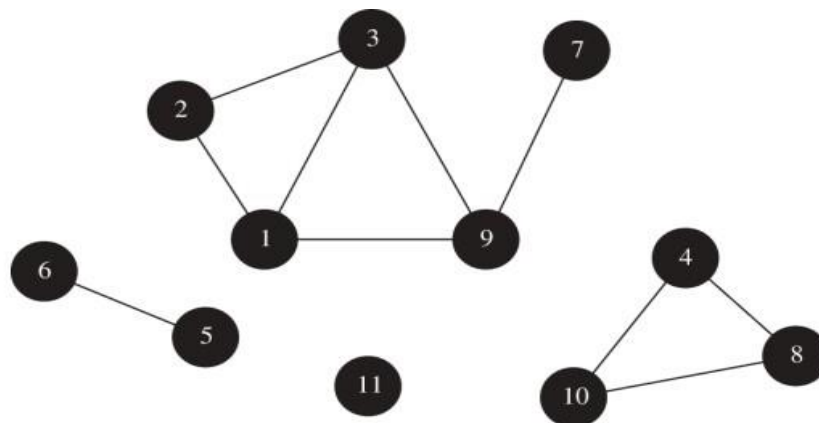


FIGURE 2.7. A toy example depicting the components of a network.

How do we find connected components in a network of size N ?

1. Set $n = 1$; colour all the nodes by red.
2. Choose an initial node v from the red coloured nodes.
3. Run breadth first search (BFS) from v ; colour all the nodes reached this way by blue; increase n by 1 upon reaching each node.
4. All the blue-coloured nodes form a component.
5. If $n = N$ then exit; Else go to Step 2.

For a directed network, if there exists a path from node A to node B, it is not necessary that there also exists a path from node B to node A. Based on this, the connectivity in directed networks has two notions: (a) strongly connected network and (b) weakly connected network.

Strongly Connected Network

A directed network G is strongly connected if there exists a (directed) path for every pair of nodes in G . Hence, if there exists a path from node v_i to node v_j , then there must also exist a directed path from node v_j to node v_i .

Example 2.15

Any undirected network in which all corresponding edges are replaced by a pair of directed edges in opposite directions forms a directed network. Such kind of directed networks are default examples of strongly connected networks. However, such networks are rare in real life, and hence has lesser significance.

Weakly Connected Network

If we replace all the directed edges of a directed network G with undirected edges, then the resultant network is called an undirected version of the directed network G . A directed network G is said to be weakly connected if its undirected version is connected.

Example 2.16

Consider the directed network shown in Figure 2.3(c). The undirected version of it would be the network shown in Figure 2.3(a). Since the network shown in Figure 2.3(a) is connected, the directed network shown in Figure 2.3(c) is weakly connected. Check yourself whether it is strongly connected or not.

1.7 NODE CENTRALITY

In the early 2020, there has been a huge eruption of controversies in Indian media regarding the White House official Twitter handle following and later (2.5) unfollowing top diplomatic Twitter handles from India during US President's visit to India.⁴ Such incidents often indicate the presence of "influential players" and the impact of their activities in a social network. Network centrality defines measures that help in finding such influential nodes in a network. Centrality is one of the most widely applied measures in network science. Typically, it is applied to measure how "central" a node is in the network.

There are a variety of measures in this category. This section introduces some of them that are prominent in the literature. It is worth noting at this point that the interpretation

of these measures usually varies with (a) the type of centrality measure used, and (b) the nature of the network we consider.

For example, PageRank that exploits eigenvector centrality, is used to find the important web pages based on the visiting patterns of users to the web pages on the Internet. Let us take a closer look into each centrality measure and its applications to real-world networks.

1.7.1 Degree Centrality

Degree centrality measures a node's degree of connectedness. It quantifies the direct influence of a node on its local neighbourhood. Mathematically, the degree centrality of a node v in a network $G(V, E)$ is defined as,

$$C_d(v) = \frac{\deg(v)}{\max_{u \in V}(\deg(u))} \quad (2.5)$$

Degree centrality has been applied in various domains. Tang et al. (2013) showed that essential proteins can be detected in protein networks using degree centrality. In online social networks such as Twitter or Facebook, degree centrality is often used to identify the most influential users. This is particularly useful for marketing wherein the detected influential user can promote a product to her followers. Bródka et al. (2011) used degree centrality on a multi-layered social network to find influential nodes across different platforms over the Internet.

Example 2.17

What is the degree centrality of nodes 1, 2, 3, 4, 5, 6 of the network shown in Figure 2.3(a)?

Solution The degree centrality values of nodes are as follows:

$$C_d(1) = \frac{4}{4} = 1.0, C_d(2) = \frac{3}{4} = 0.75, C_d(3) = \frac{2}{4} = 0.5, \\ C_d(4) = \frac{2}{4} = 0.5, C_d(5) = \frac{3}{4} = 0.75, C_d(6) = \frac{2}{4} = 0.5$$

1.7.2 Closeness Centrality

Closeness centrality is a measure of reach, that is, it tells us how close a particular node is from the rest of the network. The closeness is measured by reciprocating the mean shortest path from the node to all other connected nodes in the network. Mathematically, closeness centrality of a node v in a network $G(V, E)$ is defined as,

$$C_c(v) = \frac{|V| - 1}{\sum_{u \in V, u \neq v} d_{u,v}} \quad (2.6)$$

where $d_{u,v}$ indicates the length of the shortest path between nodes u and v .

Equation 2.6 might look non-intuitive as one may expect it to be $\frac{\sum_{u \in V, u \neq v} d_{u,v}}{|V| - 1}$.

However, in order to make “higher the closeness centrality, better it is” (which usually is the case for other centrality measures), we generally reciprocate the fraction.

Closeness centrality is often used for a social network as a critical feature to examine the extent of information spread in the network. This is particularly useful while examining the spread of fake news/misinformation on Twitter, Facebook, etc. Another area where closeness centrality is used is epidemic modelling wherein the spread of a particular disease is examined on a patient network.

Example 2.18

Calculate the closeness centrality of node 1 shown in Figure 2.3(a). Solution

There are four nodes 2, 3, 5, 6 with direct edges to node 1. Therefore, each of their corresponding shortest path length from node 1 is 1. Node 4 has a shortest path via 1 – 5 – 4 and hence has a shortest path length of 2. Therefore, the total length of the shortest paths is $1 + 1 + 1 + 1 + 2 = 6$. Since there are $|V| = 6$ nodes in the network, from Equation 2.6, we calculate the closeness centrality of node 1 as $C_c(\text{node } 1) = 5/6$.

1.7.3 Betweenness Centrality

Betweenness centrality, as the name suggests, is a measure to compute how central a node is in between paths of the network. In simpler terms, how many (shortest) paths of

the network pass through the node? To compute the betweenness centrality of node v , we first calculate all the shortest paths between all pairs of nodes in the network. The more the paths pass through node v , the higher the betweenness centrality of the node. Mathematically, betweenness centrality of node v in a network $G(V, E)$ is defined as,

$$C_B(v) = \sum_{x, y \in V, x \neq v \neq y} \frac{\sigma_{xy}(v)}{\sigma_{xy}} \quad (2.7)$$

where σ_{xy} denotes the number of shortest paths between nodes x and y in the network, and $\sigma_{xy}(v)$ denotes the number of shortest paths between nodes x and y in the network, passing through v . If $x = y$, then $\sigma_{xy} = 1$.

Betweenness centrality is useful in identifying points in a network which, if removed, may disconnect the network. Such points are known as articulation points. In communication networks, betweenness centrality helps to identify nodes which are parts of various communication channels, and are hence vulnerable to attacks. It is also useful in analysing disease spreading in epidemiology, where super spreaders are identified as nodes with a higher betweenness centrality. In security networks, betweenness centrality is used as an indicator to spot suspected spies. Leydesdorff (2007) showed that betweenness centrality can be used as an indicator of cross-disciplinary journals.

Example 2.19

Calculate the betweenness centrality of node 4 shown in Figure 2.3(a). **Solution** In the following matrix, the cell (i, j) is of the form $x|y$, where x is the number of the shortest paths between nodes i and j involving node 4; and y is the number of the shortest paths between nodes i and j . Summing all the y values will give the total number of paths. From this, we obtain

Calculate the betweenness centrality of node 4 shown in Figure 2.3(a).

Solution

In the following matrix, the cell (i, j) is of the form $x|y$, where x is the number of the shortest paths between nodes i and j involving node 4; and y is the number of the shortest paths between nodes i and j . Summing all the y values will give the total number of paths. From this, we obtain

$$C_B(4) = \frac{1}{2} + \frac{1}{2} = 1.$$

	1	2	3	4	5	6
1	0 1	0 1	0 1	–	0 1	0 1
2	0 1	0 1	0 1	–	1 2	0 1
3	0 1	0 1	0 1	–	0 1	0 1
4	–	–	–	–	–	–
5	0 1	1 2	0 1	–	0 1	0 1
6	0 1	0 1	0 1	–	0 1	0 1

1.7.4 Edge Betweenness

Centrality In the simpler terms, edge betweenness centrality refers to the fraction of all pairs of the shortest paths of the network that pass through a given edge. As before, to compute the edge betweenness centrality of edge e , we first calculate all the shortest paths between all pairs of nodes in the network. The more the paths passing through edge e , the higher the edge betweenness centrality of e .

1.7.5 Flow Betweenness

It is possible that, to communicate with another node, a node may follow a path other than the shortest path. Hence, in flow betweenness, we examine all paths between the pair of nodes instead of only the shortest paths. It is computationally more extensive than edge betweenness.

Flow betweenness is typically applied when edge betweenness fails. For example, if nodes with a higher edge betweenness fail, that is, they cannot communicate with the rest of the network or act as “reluctant brokers”, flow betweenness can be applied.

1.7.6 Eigenvector Centrality

Eigenvector centrality measures a node's importance by taking into consideration the preference of its neighbours. Thus, a node's importance is defined based on its local neighbourhood. The basic principle of finding eigenvector centrality of a node is through a recursive approach. In this approach, a node has a higher eigenvector centrality, if it itself is recommended by (or directly connected to) other nodes having high-eigenvector centrality. Eigenvector centrality is generally applied on directed networks. Mathematically, for a network $G(V, E)$, let $A = (a_{ij})$ be the adjacency matrix of G . The eigenvector centrality x_v of node v of the network is given by

$$x_v = \frac{1}{\lambda_1} \sum_{t \in N(v)} x_t = \frac{1}{\lambda_1} \sum_{t \in V} a_{v,t} \times x_t \quad (2.8)$$

where $N(v)$ is the neighbourhood of node v in G , and λ_1 is the largest eigenvalue obtained by solving the following equation:

$$\mathbf{A} \cdot \mathbf{X} = \lambda_1 \cdot \mathbf{X} \quad (2.9)$$

where $\mathbf{X}$ is a one-dimensional matrix, whose v^{th} entry is x_v , the eigenvector centrality of v .

Eigenvector centrality is one of the most widely applied centrality measures. PageRank, which is based upon eigenvector centrality, is one of the earliest algorithms for web page recommendation based on the visiting patterns of millions of users. Eigenvector centrality is also often touted to find the most “significant” nodes in a network. In online social networks such as Twitter, Facebook, Instagram, users are often recommended based on eigenvector centrality. Elliott and Golub (2019) showed that eigenvector centrality can be used as a metric to show that a person's choice influences an efficient social outcome.

1.7.7 Katz Centrality

Katz centrality computes the relative influence of a node in a network by considering all immediate neighbours and all further nodes connected to the node. Katz centrality is an extension of eigenvector centrality. Connections with distant neighbours are penalised by an attenuation factor α , $0 < \alpha < 1$. Katz centrality of a node v_i in a network $G(V, E)$, denoted $CKatz(i)$, is defined as follows:

$$C_{Katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{|V|} \alpha^k \times A_{j,i}^k \quad (2.10)$$

The powers of A indicate the presence (or absence) of links between two nodes through intermediaries. For instance, in matrix A₄, if element a_{1,2} = 1, it indicates that nodes 1 and 2 are connected through a walk of length 4. Therefore, in Equation 2.10, A_{j,i}^k indicates the total number of k-hop connections between nodes j and i.

1.7.8 PageRank

PageRank (Brin and Page 1998) is an algorithm which rates importance of web pages. This is done by a recursively defined measure, where a page's importance is determined by the importance of the web pages linked to the particular page. This is recursive because the page further contributes to the importance of the web pages linked to it. PageRank scores are typically defined for directed networks such as the World Wide Web (WWW) Network, wherein an out-edge from node A to node B is defined if there exists a hyperlink from web page A directing it to web page B. Mathematically, for a network G(V, E), the PageRank score of node v_i, denoted by PG(v_i), is defined as,

$$PG(v_i) = \frac{(1-d)}{|V|} + d \sum_{v_t \in In(v_i)} \frac{PG(v_t)}{outdeg(v_t)} \quad (2.11)$$

where d is the damping factor, and $In(v_i)$ is the set of nodes pointing to v_i .

A good way to think about PageRank is by imagining a random web surfer surfing through the Internet by going from one link to another. Suppose a random walker is on a web page which has three hyperlinks to other web pages (out-edges). The random walker has an equal probability, that is, 1/3 of visiting any of the three web pages. Hence, we normalise the PageRank scores of nodes by the respective out-degrees of the nodes (the second term of Equation 2.11).

The damping factor d is a measure which reflects the probability of the random walker to continue its current walk by clicking on a hyperlink on the current web page. Although

various studies have been done to figure out the optimal value of d , it is generally assumed that $d = 0.85$. The first term of Equation 2.11, $(1 - d)/|V|$, is the probability of the random walker starting a new walk from a random node. Chapter 4 elaborates more on PageRank.

Example 2.20

Calculate the eigenvector, Katz centrality and PageRank scores of all the nodes shown in Figure 2.3(c) keeping $\alpha = 0.1$, $d = 0.85$ and the maximum number of iterations as 100.

Solution

The centrality values are mentioned in Table 2.1.

Table 2.1. Centrality values for Example 2.20.

Node	Eigenvector	Katz	PageRank
1	0.500	0.468	0.262
2	1.192×10^{-7}	0.355	0.051
3	2.623×10^{-6}	0.390	0.065
4	0.500	0.430	0.181
5	0.500	0.402	0.274
6	0.499	0.395	0.167

While PageRank is highly effective in identifying the prestigious nodes with high in-degree in a network, it does not give separate attention to those nodes which point to important nodes, that is, those whose out-degree is high. Nodes with high out-degree are often called as hubs in a network. In 1999, Jon Kleinberg, an eminent American computer scientist, proposed a new notion to measure the importance of nodes by considering two different factors of a node hubness and authoritativeness. While the former factor indicates the inclination of a node to point to many other nodes, the latter emphasises on its ability to be directed by many other nodes. Kleinberg proposed HyperlinkInduced Topic Search algorithm (Kleinberg 1999) that analyses links in a directed network such as web network, citation network and ranks nodes based on the above two factors separately. In the context of citation networks, nodes with high hubness are the survey papers, and nodes with high authoritativeness are the seminal

papers. In a web network, high hub nodes are those web pages that serve as large directories of essential web pages.

1.7.9 Hub and Authority

For node v , its hubness is determined by the cumulative authoritativeness of nodes that v points to. Similarly, its authoritativeness is computed by the cumulative hubness of the nodes pointing to v .

$$\begin{aligned} \text{hub}(v) &= \sum_{u \in \text{out}(v)} \text{auth}(u) \\ \text{auth}(v) &= \sum_{u \in \text{in}(v)} \text{hub}(u) \end{aligned} \quad (2.12)$$

where $\text{in}(v)$ is the set of nodes pointing to v , and $\text{out}(v)$ is the set of nodes pointed by v .

1.10 ASSORTATIVITY

Assortativity or assortative mixing is a measure which is based on the principle that two similar nodes have a high preference for each other. The assortativity of a network can be defined in different ways. However, network scientists often like to express assortativity in terms of the degrees of the nodes in a network.

For example, it may be something as simple as two nodes sharing a lot of common neighbours. A common method is to use Pearson correlation coefficient to find similarity between two nodes. Mathematically, Pearson correlation coefficient, for two data distributions x and y , is given by,

$$r_{xy} = \frac{N \sum x \cdot y - \sum x \sum y}{\sqrt{\left(N \sum x^2 - \left(\sum x\right)^2\right) \cdot \left(N \sum y^2 - \left(\sum y\right)^2\right)}} \quad (2.13)$$

where N is the number of samples in each distribution. We can think of x and y as the degree distributions of nodes in two end points of edges. If $r_{xy} = 1$, the two data distributions are said to be perfectly assortative. If $r_{xy} = -1$, the $\text{hub}(v) = \sum_{u \in \text{out}(v)}$

$\text{auth}(u) \text{ auth}(v) = \sum_{u \in \text{in}(v)} \text{hub}(u)$ two data distributions are said to be perfectly disassortative. If $r_{xy} = 0$, the two data distributions are non-assortative.

1.11 TRANSITIVITY AND RECIPROCITY

Transitivity and reciprocity are metrics used to determine the linkage between pairs of nodes. By linking behaviour, we refer to the nature of forming edges between nodes in a network. For example, we often find new friends among the friends of our friends in real life, or we often follow back the accounts on social networks who follow us. The metrics, we shall discuss here, can quantify these phenomenon in a systematic way.

1.11.1 Transitivity

In abstract mathematics, transitivity in a relation means that when there exists a tie from entity a to entity b, and also from b to c, then there also exists a tie from a to c. The same analogy is adopted in the theory of social network as follows: “Friends of one’s friends are her friends as well”.

Formally, in network $G(V, E)$, when there is an edge between nodes v_1 and v_2 , and also an edge between v_2 and v_3 , if we find another edge between v_1 and v_3 in G , then we observe a transitive linking in G .

Clearly, transitivity involves three edges. These three edges along with their corresponding end nodes form a triangle or a triad. The presence of such triangles provides a measure of transitive behaviour in the network. More the triangles in the network, higher is its transitivity.

One example of a (completely) transitive network is evident the complete network with three or more nodes. In a nutshell, the measure of transitivity intends to capture how close a network is to a complete network. In the network shown in Figure 2.3(a), nodes 1, 2, and 3 show a transitive behaviour; the same holds for nodes 1, 5, and 6. Transitivity in a network may be measured using clustering coefficient introduced in Section 2.1.3. Though we have discussed transitivity so far for undirected networks, the same may be applied to directed networks as well, where instead of triangles, we consider the loops of size 3 (or motifs (Lecca et al. 2016)).

1.11.2 Reciprocity

Reciprocity is a measure of the likelihood of nodes in a directed network to be linked to each other. Formally, given a directed edge from node u to node v , if we find another

directed edge from v to u in the network G , then we notice an instance of reciprocal behaviour in G . Informally, reciprocity refers to — “If you would follow me, I shall follow you back.”

$$R = \frac{2}{|E|} \sum_{i,j; i < j} a_{ij} \times a_{ji} \quad (2.14)$$

where $A = (a_{ij})$ is the adjacency matrix of G . On simplification, R reduces to

$$R = \frac{2}{|E|} \times \frac{1}{2} \text{Tr}(A^2) = \frac{1}{|E|} \text{Tr}(A^2) \quad (2.15)$$

where $\text{Tr}(X)$ refers to the *trace* of a square matrix X , which is the sum of the diagonal elements of X .

1.12 SIMILARITY

Similarity measures are used to check when a pair of nodes in a network are part of the same equivalence class. Similarity measures can be divided into the following two classes: structural equivalence, and regular equivalence

1.12.1 Structural Equivalence

In the structural similarity, we look at the nodes which are in the common neighbourhood to the given two nodes. For example, in a network of your friends, you and your best friend might be most similar since you may share a lot of common friends with your best friend. Let a and b be two nodes in the network G . Let $N(a)$ and $N(b)$ be the neighbourhoods of nodes a and b , respectively. Different measures through which we can measure structural similarity between nodes a and b are as follows.

1. **Common neighbours:** It is simply the number of common neighbours shared in the neighbourhoods $N(a)$ and $N(b)$. Mathematically, it is defined as,

$$\sigma_{CN}(a,b) = |N(a) \cap N(b)| \quad (2.16)$$

2. **Jaccard similarity:** Here we simply normalise common neighbours by the combined size of the neighbourhoods of the two nodes, $|N(a) \cup N(b)|$. Mathematically, it is defined as,

$$\sigma_{JS}(a,b) = \frac{|N(a) \cap N(b)|}{|N(a) \cup N(b)|} \quad (2.17)$$

3. **Cosine similarity:** Here, we normalise the common neighbours by the individual sizes of the neighbourhoods, $|N(a)|$ and $|N(b)|$.

$$\sigma_{CS}(a,b) = \frac{|N(a) \cap N(b)|}{\sqrt{|N(a)||N(b)|}} \quad (2.18)$$

1.12.2 Regular Equivalence

In regular equivalence, we measure the overall similarity of the neighbourhoods around the nodes. If the corresponding structures are similar, we say that the nodes have a high regular equivalence. It is important to note that we do not consider the specific nodes which overlap in both the neighbourhoods, rather the roles each of the nodes plays in the neighbourhood are considered

Example 2.21

One day, you open your Facebook account and notice a friend request from a user from Australia. You are surprised as you do not know the person. Upon examining her account, you realise that two of her friends are also your friends, that is, both of you have two mutual friends. You also realise that the mutual friends study in Australia with the person who sent you the friend request. Therefore, it is not that two users always need to be connected to be similar, it is the mutual neighbours that they share and how similar the mutual neighbours are, which matters in determining the similarity of the given pair of nodes.

Mathematically, for nodes v_i and v_j in network G with an adjacency matrix A , the regular equivalence between them is defined as,

$$\sigma_{reg}(v_i, v_j) = \alpha \sum_k A_{i,k} A_{j,k} \sigma_{reg}(v_k, v_l) \quad (2.19)$$

Equation 2.19 is self-referential as in order to solve it for v_i and v_j , we look at their neighbours v_k and v_l . Similarly, to solve it for v_k and v_l , we need to consider v_i and v_j . One can further relax Equation 2.19 as follows:

$$\sigma_{reg}(v_i, v_j) = \alpha \sum_k A_{i,k} \sigma_{reg}(v_k, v_j) \quad (2.20)$$

This can further be re-written as,

$$\sigma_{reg} = \alpha A \sigma_{reg} \quad (2.21)$$

We can add an identity matrix to the right-hand side, to make sure that a node is highly similar to itself. Finally we get

$$\sigma_{reg} = \alpha A \sigma_{reg} + \mathbf{I} \quad (2.22)$$

$$\sigma_{reg} = (\mathbf{I} - \alpha A)^{-1} \quad (2.23)$$

For convergence, $\alpha < 1/\lambda_1$, where λ_1 is the largest eigenvalue of A .

1.13 DEGENERACY

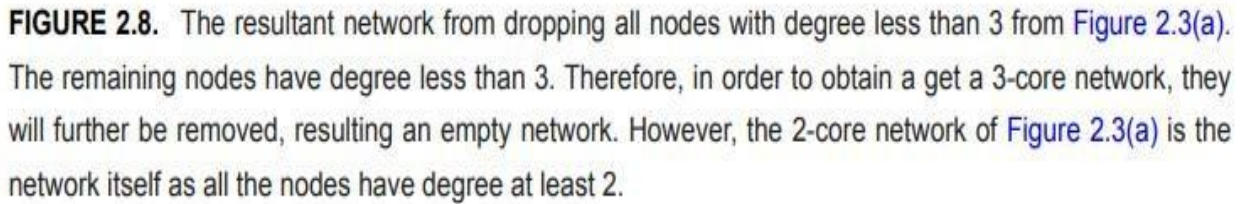
Network degeneracy is an indirect indication of the sparsity of a network.

1.13.1 k-core

For an undirected network G , the k -core is the maximal subgraph of G in which every node has at least k neighbours (or are of degree at least k). How would we construct the k -core of a network G ?

The first thought that may come to our mind is to simply drop all the nodes with degree less than k . Consider the network shown in Figure 2.3(a). To form a 3-core of this network, we drop all the nodes with degree less than 3 (nodes 3, 4, and 6) from the network.

Consequently, we obtain the network as shown in Figure 2.8. However, this is not the desired 3-core, since all the remaining nodes also have degree less than 3. As you can see, dropping all nodes of degree less than k may also decrease the degree of their (undropped) neighbours to below k , thereby making them ineligible to be part of the k -core subgraph. Therefore, to form the k -core subgraph we recursively “shave” nodes with degree less than k . Figure 2.9 shows k -core subgraphs for $k = 0, 1, 2, 3, 4$ for a toy network.



1.13.2 Coreness

The coreness or core number of a node is the order of the highest order core that the node belongs to. A node has a coreness k in network G if it belongs to the k -core subgraph, but does not belong to the $(k + 1)$ -core subgraph of G . For example, in Figure 2.10, nodes inside the central most 4-core subgraph have coreness 4. Coreness is similar to centrality in a sense that it also measures the most significant or prestigious nodes in the network. This is based on the intuition that nodes with higher coreness tend to be more central and connected and hence more influential in the network.

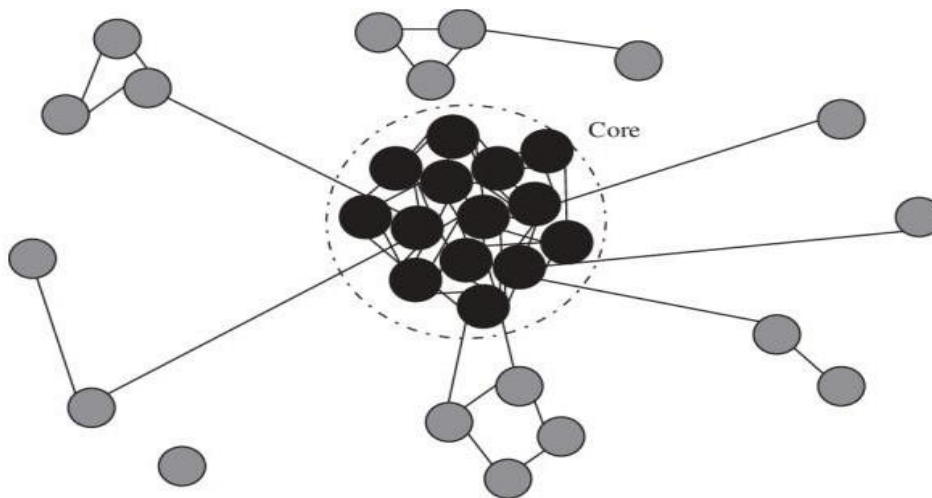


FIGURE 2.10. A visualisation of the core-periphery network model. The black nodes represent the core nodes, while the gray nodes form the periphery.

1.13.3 Core-periphery

A common structure observed in real-world networks is the core-periphery structure. The core-periphery structure consists of a dense and connected core surrounded by disconnected and scrambled periphery. The basic structure of a core-periphery contains:

- Dense core:** The nodes inside core are connected. Moreover, the core has an incredibly high-density and low-average path length.
- Disconnected periphery:** The periphery is disconnected into several smaller components. The components individually may be connected to the core; however, there is little to no interaction between the smaller periphery components.

