Social Network Overview 3

3.1 Overview

Online social networks (OSNs) denote the online platforms that are used by people to build social connections with the other people, who may share similar personal or career interests, backgrounds, or real-life connections. Online social networking sites vary a lot and there exist a large number of online social sites of different categories, including *online sharing sites*, *online publishing sites*, *online networking sites*, *online messaging sites* and *online collaborating sites*. Each category of these online social networks can provide specific featured services for the customers. For instance, Facebook allows users to socialize with each other via making friends, posting text, sharing photos and videos; Twitter focuses on providing micro-blogging services for users to write/read the latest news and messages; Foursquare is a location-based social network offering location-oriented services; and Instagram is a photo and video sharing social site among friends or to the public. To enjoy different kinds of social network services simultaneously, users nowadays are usually involved in multiple online social sites at the same time, in each of which they will all form social connections and generate social information.

Generally, online social networks can be represented as graphs in mathematics. Besides the users, there usually exist many other types of information entities, like posts, photos, videos, and comments, generated by users' online social activities. Information entities in online social networks are extensively connected, and the connections among different types of nodes usually have different physical meanings. The diverse nodes and connections render online social network to be a very complex graph structure. Meanwhile, depending on the categories of information entities and connections involved, the online social networks can be divided into different types, like *homogeneous network* [48], *bipartite network* [65], and *heterogeneous network* [51]. To model the phenomenon that users are involved in multiple networks, a new concept called "*multiple aligned heterogeneous social networks*" [29, 71–73] has been proposed in recent years.

Different online social networks are usually of different characteristics, which can be quantified with some network measures formally. Users in online social networks can have different numbers of connections, which can be quantified as the user *node degree* [2, 8] mathematically. User nodes of a larger degree will be more important (in terms of social connections) generally. A more formal concept indicating the node importance is called the *node centrality* [11], which can be quantified with many different measures. Connections are very important for online social networks, and node connection measures quantifying the linking behaviors of nodes in the networks are of great interests.

J. Zhang, P. S. Yu, *Broad Learning Through Fusions*,
https://doi.org/10.1007/978-3-030-12528-8_3

Based on the connections among nodes, the social closeness measures between pairs of nodes can be calculated, where user nodes who frequently interact with each other will have a larger closeness score. As to the local social connection patters, they may also follow the social balance theory, e.g., "friends of my friend are my friends."

For the networks with simple structures, like the homogeneous networks merely involving users and friendship links, the social patterns are usually easy to study. However, for the networks with complex structures, like the heterogeneous networks, the nodes can be connected by different types of links sequentially, which are of different physical meanings. One general technique for heterogeneous network studies is "*meta path*" [53, 73], which specifically depicts certain link sequences connecting the nodes based on the network schema. The meta path concept can also be extended to the *multiple aligned social network* scenario [29,73], which can connect the nodes across different social networks. The machine learning approaches introduced in the previous chapter are very general learning models, which take the feature representation data as the input and output the predicted labels of the data instances. There actually also exist some learning algorithms proposed for the network structured data specifically, like the *random walk* approach [36].

In this chapter, we will provide the definitions of some important concepts that are useful for the social network studies, including the basic graph related concepts, and some advanced social network concepts, like *meta path* [53, 73]. A clear categorization of the network types will be provided, and some network measures will be introduced to illustrate the properties of the networks. Finally, an introduction about some network-based models will be provided. These concepts, network categories, network measures, and approaches will be frequently used and mentioned in the following chapters of this book.

3.2 Graph Essentials

In mathematics and computer science, the online social networks are generally represented as graphs [60], where the information entities are denoted as the nodes and the connections among the information entities are represented as the links. In this section, we will provide some basic introductory knowledge about graph, including its representations and the connectivity properties.

3.2.1 Graph Representations

Graphs can be represented in different forms, like a traditional graph definition involving nodes and links, an *adjacency matrix* indicating the connectivity among nodes, *adjacency list* and *link list*.

Definition 3.1 (Graph) Formally, a graph can be represented as $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{E}$ represents the set of links in the graph G.

Generally, node is the basic entity unit in graphs, which can represent different types of information entities when using the graph definition to represent social networks. For instance, in online social networks, a node can denote a user, a post, a comment, and a photo. The formal representation of the node set $\mathcal{V}$ can be denoted as

$$\mathcal{V} = \{v_1, v_2, \ldots, v_n\}, \tag{3.1}$$

where v_i $(1 \leq i \leq n)$ represents a single node in the graph and the node set size (i.e., the size of the graph) is $|\mathcal{V}| = n$.

Meanwhile, the different kinds of connections among the information entities are represented as the links in the graphs, which bear various physical meanings. For instance, in online social networks, the links among users can denote their friend/follow relationships, the links between users and posts denote the post-writing action, and the links between posts and spatial "(latitude, longitude)" coordinate pairs denote the check-ins attached to the posts. Formally, the set of links in the network can be represented as

$$\mathcal{E} = \{e_1, e_2, \ldots, e_m\}, \tag{3.2}$$

where $e_j = (v_o, v_p)$ $(1 \leq j \leq m)$ denotes a link/node pair in the graph. The size of the link set in the network can be represented as $|\mathcal{E}| = m$.

Besides the aforementioned regular graph definition, a graph can also be represented as an *adjacency matrix*, which indicates the connectivity among the nodes.

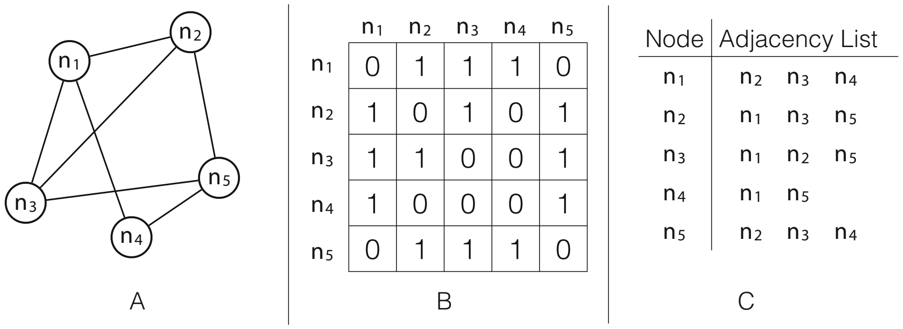
Definition 3.2 (Adjacency Matrix) Given a graph $G = (\mathcal{V}, \mathcal{E})$, we can represent its corresponding adjacency matrix as a binary matrix $\mathbf{A} = \{0, 1\}^{n \times n}$, where the rows and columns of the matrix correspond to the nodes in G and entry $A(i, j) = 1$ iff link $(v_i, v_j) \in \mathcal{E}$.

The graph definition and its adjacency matrix representation actually have equivalent representation capacity, and the transformation between which can be achieved very easily. Various properties of the graphs can also be revealed by their adjacency matrices as well. For instance, if a graph has a very small number of connections compared with the number of nodes in it, the corresponding adjacency matrix of the graph will be very sparse [43]. Meanwhile, if the nodes in the graph actually form several communities where the nodes in each community tend to have dense connections compared with those outside the communities, the corresponding graph adjacency matrix will have a lower rank [47].

Besides the *adjacency matrix*, the other graph representations include adjacency list. Let set $\Gamma(u_i) = \{u_j | u_j \in \mathcal{V}, (u_i, u_j) \in \mathcal{E}\} \subset \mathcal{V}$ denote the neighbors that user u_i connects to. The adjacency list representation of graph G can be represented as $\{(u_i, \Gamma(u_i))\}_{u_i \in \mathcal{V}}$.

Example 3.1 For instance, given a graph illustrated in Fig. 3.1a, we can represent the graph as

$$G = (\mathcal{V}, \mathcal{E}), \tag{3.3}$$



$$\begin{array}{c|ccccc} & n_1 & n_2 & n_3 & n_4 & n_5 \\ \hline n_1 & 0 & 1 & 1 & 1 & 0 \\ n_2 & 1 & 0 & 1 & 0 & 1 \\ n_3 & 1 & 1 & 0 & 0 & 1 \\ n_4 & 1 & 0 & 0 & 0 & 1 \\ n_5 & 0 & 1 & 1 & 1 & 0 \end{array}$$

Node	Adjacency List
n1	n2 n3 n4
n2	n1 n3 n5
n3	n1 n2 n5
n4	n1 n5
n5	n2 n3 n4

Fig. 3.1 An example of different graph representations. ((**a**) Graph; (**b**) adjacency matrix; (**c**) adjacency list)

where the node set $\mathcal{V} = \{n_1, n_2, n_3, n_4, n_5\}$ contains five nodes and the link set $\mathcal{E} = \{(n_1, n_2), (n_1, n_3), (n_1, n_4), (n_2, n_3), (n_2, n_5), (n_3, n_5), (n_4, n_5)\}$ covers seven links. In the graph, there are five different nodes $\{n_1, n_2, n_3, n_4, n_5\}$, where they are connected by seven links. In the graph, all the nodes are connected with three other nodes, except n_4 which is connected to n_1 and n_5 only. We show its adjacency matrix and adjacency list representations in Fig. 3.1b, c, respectively. For any connected node pairs in the graph, the corresponding entry in the matrix will be filled with value 1; otherwise they have value 0 instead. For instance, link (n_2, n_3) connects nodes n_2 and n_3. In the adjacency matrix, the (2nd, 3rd) entry and the (3rd, 2nd) are both filled with value 1. The graph is also represented as an adjacency list as shown in Fig. 3.1c. For each node in the graph, we provide a list of nodes connected with the nodes. For instance, node n_3 is connected with nodes n_1, n_2, and n_5 simultaneously, which will form the adjacency list of node n_3.

3.2.2 Connectivity in Graphs

Connectivity [8] is an important property of graphs, where nodes are connected with each other via either direct connections or paths consisting of a sequence of links. Formally, given a graph G and a node n in the graph, the set of nodes that are adjacent to n in the graph are called the *adjacent neighbors* of n in the graph G.

Definition 3.3 (Adjacent Neighbor) Given a graph $G = (\mathcal{V}, \mathcal{E})$, the *adjacent neighbors* of node n in G can be represented as $\Gamma(n) = \{n' | n' \in \mathcal{V} \wedge (n, n') \in \mathcal{E}\}$.

Adjacent neighbor set is an important concept in social network studies. For instance, given a social network, the *adjacent neighbor* set of a user denote the online friends that the user is connected to, which is very useful for analyzing the socialization patterns and preference of users in the social network.

Meanwhile, given a node n in a network G, we can call the set of links incident to n in the graph as the *incident links* of node n.

Definition 3.4 (Incident Link) Given a graph $G = (\mathcal{V}, \mathcal{E})$ and a node $n \in \mathcal{V}$, the set of *incident link* set of n in G can be represented as $\Delta(n) = \{e | e \in \mathcal{E} \wedge \exists n' \in \mathcal{V}, (n, n') = e\}$.

Furthermore, we can also define the *incident relationships* between two links. Formally, given two links (a, b) and (c, d) in graph G, (a, b) is said to be incident to (c, d) iff $a = c \vee a = d \vee b = c \vee b = d$, i.e., they share a common node. Based on this definition, we can define the concepts of *walk*, *path*, *trail*, *tour*, and *cycle* of graph G as follows:

- **Walk**: Formally, a *walk* can be denoted as a sequence of nodes $n_1, n_2, \ldots, n_k$ from set $\mathcal{V}$, where there exists a link between any sequential pairs of nodes in the graph. For any three sequential nodes in the sequence, e.g., n_i, n_{i+1}, n_{i+2}, the links (n_i, n_{i+1}) and (n_{i+1}, n_{i+2}) are *incident* to each other sharing a common node n_{i+1}. Furthermore, if the ending node n_k is the same as the starting node n_1 in the *walk*, then it will be called a *closed walk*; otherwise, it is called an *open walk*. The length of the walk is formally defined as the number of links involved in the walk. For instance, sequence $n_1, n_2, \ldots, n_k$ forms a walk of length $k - 1$.
- **Trail**: A *trail* denotes a *walk* in the graph G, where all the links are distinct. By traveling along a *trail*, each link in the *trail* can be visited once, but the nodes can be visited multiple times. The shortest *trail* in graph G can be just one link in the graph.

- **Tour**: A *closed trail* (i.e., the starting and ending nodes of the *trail* are the same) is called a *tour*.
- **Path**: Given a *walk* in the graph G, if all the nodes and links in the *walk* are distinct, the *walk* will be a *path* in the graph. A *path* is also a *trail* in the graph.
- **Cycle**: A *closed path* is defined as a *cycle* in graphs. A *cycle* is also a special type of *tour*.

To help explain the above concepts, we also provide an example as follows, which lists the *walk*, *trail*, *tour*, *path*, and *cycle* instances from the input graph.

Example 3.2 For instance, based on the graph illustrated in Fig. 3.2, the node sequences

1. "$n_1, n_2, n_3, n_5, n_4, n_1, n_2$" is a *walk* of length 6,
2. "n_1, n_2, n_3, n_1, n_4" is a *trail* of length 4,
3. "n_1, n_2, n_5, n_4, n_1" is a *tour* of length 4,
4. "n_1, n_3, n_5, n_2" is a *path* of length 3,
5. "$n_1, n_2, n_3, n_5, n_4, n_1$" is a *cycle* of length 5

in the graph, respectively.

The above concepts can help correlate the nodes in the graphs which are not directly connected with each other.

Definition 3.5 (Reachable) Formally, given two nodes n_i and n_j in the graph G, n_i is said to be *reachable* from n_j iff there is a *path* from n_j to n_i.

For a subset of nodes, which are *reachable* from each other, they together with the links among them will form a *connected component* in the graph.

Definition 3.6 (Connected Component) Given a graph $G = (\mathcal{V}, \mathcal{E})$, the subgraph $G' = (\mathcal{V}', \mathcal{E}')$ is said to be a connected component of G iff $\mathcal{V}' \subset \mathcal{V}$, $\mathcal{E}' \subset \mathcal{E}$, and for any pair of nodes in $\mathcal{V}'$ they are *reachable* via the links in $\mathcal{E}'$.

Example 3.3 For instance, based on the input graph illustrated in Fig. 3.2, the subgraph $G' = (\{n_1, n_2, n_4, n_5\}, \{(n_1, n_2), (n_2, n_5), (n_4, n_5), (n_1, n_4)\})$ will be a *connected component* of the input graph. Meanwhile, considering that all the nodes in the network are *reachable* to each other, and the original network itself is also a *connected component* actually.

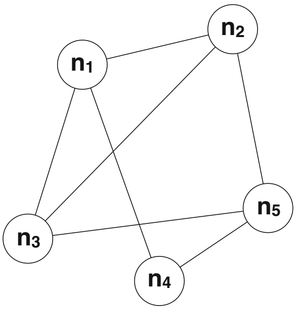


Fig. 3.2 An input graph example

Based on the graph links $\mathcal{E}$, there may exist multiple *paths* of different lengths connecting a certain pair of nodes (e.g., n_i, n_j). Meanwhile, the *path* of the shortest length can be of great importance and has concrete applications in many research problems, like *traffic route planning* [6]. Formally, such a *path* is also named as the *shortest path* in graphs.

Definition 3.7 (Shortest Path) Given a pair of nodes $n_i, n_j \in \mathcal{V}$ in the graph G, the set of *paths* connecting n_i and n_j based on G can be represented as $\mathcal{P}$, in which one of the shortest lengths is called the *shortest path* between n_i and n_j:

$$SP(n_i, n_j) = \min_{p \in \mathcal{P}} |p|, \tag{3.4}$$

where $|p|$ denotes the length of path p.

The *shortest path* between different node pairs in a graph can be of different lengths, where the longest *shortest path* between nodes in graph G is also defined as the *diameter* of the graph.

Definition 3.8 (Graph Diameter) Formally, given a graph G, the *diameter* of graph G can be represented as

$$Diameter(G) = \max_{n_i, n_j \in \mathcal{V}} SP(n_i, n_j). \tag{3.5}$$

Example 3.4 For instance, based on the graph illustrated in Fig. 3.2, the *shortest path* between (1) nodes n_1 and n_2 is "$n_1 \rightarrow n_2$" (of length 1), and (2) nodes n_2 and n_4 is "$n_2 \rightarrow n_5 \rightarrow n_4$" of length 2 (or "$n_2 \rightarrow n_1 \rightarrow n_4$"). For any two nodes selected from the graph, we observe that the *shortest path* length between them are no greater than 2, i.e., the *diameter* of the graph is 2.

3.3 Network Measures

The networks are usually of different structures and will have different properties, which can be indicated by various measures about either the nodes, links, or the overall network structure. In this part, we will introduce a number of measures about the networks, including the *degree* [2, 8] and *centrality* [11] about nodes, *similarity* [67] about node pairs (i.e., the links), and the *transitivity* [19] and *social balance* [20, 57] about the network structures.

3.3.1 Degree

Degree [2, 8] can effectively indicate the number of connections associated with nodes in graphs, which is a very important node measure. In this part, we will introduce the *node degree* concept and the *node degree distribution* [2] in graphs.

3.3.1.1 Node Degree

Given an undirected network $G = (\mathcal{V}, \mathcal{E})$, the node degree denotes the number of edges incident to the nodes, whose formal definition is provided as follows.

Definition 3.9 (Degree) The *degree* of node u in an undirected network $G = (\mathcal{V}, \mathcal{E})$ denotes the number of links incident to it, i.e., $d(u) = |\{(u, v)|v \in \mathcal{V}, (u, v) \in \mathcal{E}\}|$.

In an undirected network, each link will be incident to two nodes, and the total node degree of a network will always be an even number. Furthermore, as to the specific numbers of the degrees, we have the following theorem.

Theorem 3.1 *Given an undirected network $G = (\mathcal{V}, \mathcal{E})$, the total number of node degrees equal to twice the number of links in the network, i.e.,*

$$\sum_{u \in \mathcal{V}} d(u) = 2|\mathcal{E}|. \tag{3.6}$$

Proof In network G, the total node degree can be represented as $\sum_{u \in \mathcal{V}} d(u)$. The removal of link $(u, v) \in \mathcal{E}$, will lower down the degree of nodes u and v by 1, respectively. The total node degree after removing link (u, v) will be equal to $\sum_{u \in \mathcal{V}} d(u) - 2$. After removing all the links (i.e., $|\mathcal{E}|$ links) from the network, the total node degree will be reduced to 0 as all the nodes are isolated without any connections. Therefore, $\sum_{u \in \mathcal{V}} d(u) - 2|\mathcal{E}| = 0$, and we have

$$\sum_{u \in \mathcal{V}} d(u) = 2|\mathcal{E}|. \tag{3.7}$$

In the case that links in the networks are directed, the node degree concept will be further refined into *node in-degree d_{in}* and *node out-degree d_{out}*, which denotes the number of links coming into the nodes and those going out from the nodes, respectively.

Theorem 3.2 *Given a directed network $G = (\mathcal{V}, \mathcal{E})$, the total number of node in-degree and out-degree are both equal to the number of nodes in the network, i.e.,*

$$\sum_{u \in \mathcal{V}} d_{in}(u) = \sum_{u \in \mathcal{V}} d_{out}(u) = |\mathcal{E}|. \tag{3.8}$$

Proof Similarly, we can represent the total node in-degree and out-degree of network G as $\sum_{u \in \mathcal{V}} d_{in}(u)$ and $\sum_{u \in \mathcal{V}} d_{out}(u)$, respectively. From network G, the removal of each link $(u, v) \in \mathcal{E}$ will decrease the out-degree of u and in-degree of v by 1. Therefore, after the removal of link (u, v), the new total node in-degree and out-degree of network G will be $\sum_{u \in \mathcal{V}} d_{in}(u) - 1$ and $\sum_{u \in \mathcal{V}} d_{out}(u) - 1$, respectively. After removing all the links in $\mathcal{E}$, the node in-degree and out-degree will be decreased to 0, and all the nodes will become isolated without any connections. In other words, we have $\sum_{u \in \mathcal{V}} d_{in}(u) - |\mathcal{E}| = \sum_{u \in \mathcal{V}} d_{out}(u) - |\mathcal{E}| = 0$, which implies that

$$\sum_{u \in \mathcal{V}} d_{in}(u) = \sum_{u \in \mathcal{V}} d_{out}(u) = |\mathcal{E}|. \tag{3.9}$$

3.3.1.2 Degree Distribution

Node degree is an important property about the nodes, while the distribution of the node degrees displays an important property of the whole network instead. Given a node degree value d, we can represent the proportion of nodes with degree d as

$$P(d) = \frac{|\{v|v \in \mathcal{V}, d(v) = d\}|}{|\mathcal{V}|}, \tag{3.10}$$

where the numerator denotes the number of nodes with degree d.

All the potential degree values of nodes in the network can be represented as set $\mathcal{D} = \{d(u)|\forall u \in \mathcal{V}\}$. Therefore, the node degrees together with the corresponding proportions will be represented as a tuple set $\{(d, P(d))\}_{d \in \mathcal{D}}$, which can be represented as a distribution plot with degrees as the x axis and the proportions as the y axis.

Example 3.5 For instance, given an undirected network shown in Fig. 3.2, there exist five nodes n_1, n_2, n_3, n_4, n_5 with degrees 3, 3, 3, 2, 3, respectively. Therefore, the node degree and proportion tuples can be represented as $\left\{\left(2, \frac{1}{5}\right), \left(3, \frac{4}{5}\right)\right\}$. We can represent the degree distribution in Fig. 3.3, where majority of the nodes have degree 3 (the largest node degree in the network) and a small proportion of nodes have degree 2 (the smallest node degree in the network).

Such a degree distribution about the toy example shown in Fig. 3.3 is not common in the real-world social networks. In many of the cases, most of the users are regular users with a limited number of friends online (i.e., a small degree), and a small number of celebrities can have a large number of friends (i.e., a large degree).

Example 3.6 In Fig. 3.4, we show the degree distribution plots of two crawled data sets about the Foursquare and Twitter online social networks, where each of them contains about 5000 users. According to the plots, we observe that the user fraction generally drops as the node degree increases in both Foursquare and Twitter. Among all the users, most of the users in both Foursquare and Twitter have a very small degree (less than 10). Compared with Foursquare, users in Twitter have more dense connections and tend to have larger degrees. For instance, the fraction of users with small degrees in Twitter is less than that in Foursquare (i.e., the red dots are below the blue dots for small degrees), while the Twitter user fractions of larger degrees are above those of Foursquare (i.e., the right part of the plot). According to the plot, there also exists one user in the Twitter network with a degree greater

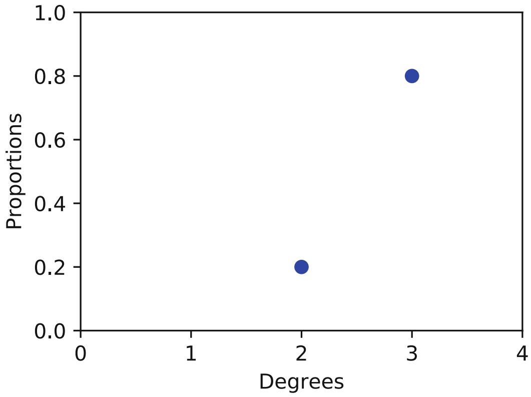


Fig. 3.3 Degree distribution of the example network

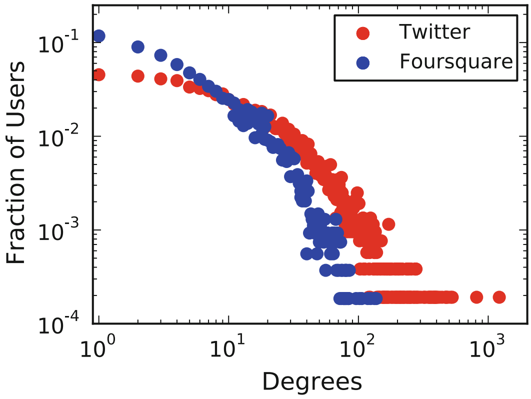


Fig. 3.4 Degree distribution of the Foursquare and Twitter networks

than 1000, i.e., the rightmost red dot, whose is usually a celebrity with a great number of followers in the networks.

3.3.2 Centrality

The concept *centrality* [11] defines how important a node is in the network. To quantify the node importance in the networks, different kinds of metrics can be applied to define the node *centrality*, which will be introduced in this part.

3.3.2.1 Degree Centrality

In the real-world online social networks, the users with lots of connections (i.e., large degrees) tend to be important, as their roles are recognized by other users via the connections with them. Therefore, the node importance can be quantified as the node degrees. Given an undirected network G, the *degree-based centrality* [11, 65] of a node u in the network can be defined as

$$C_d(u) = d(u). \tag{3.11}$$

All the nodes in G can be ordered by their *degree-based centrality*, where the nodes with larger degrees will be more important compared with other nodes with smaller degrees. Meanwhile, given a directed network G, the node centrality can be defined as either their in-degrees, out-degrees, or in-degrees together with out-degrees, which can be formally represented as follows:

$$C_{in}(u) = d_{in}(u), \tag{3.12}$$

$$C_{out}(u) = d_{out}(u), \tag{3.13}$$

$$C_{in/out}(u) = d_{in}(u) + d_{out}(u). \tag{3.14}$$

Example 3.7 For instance in Fig. 3.5, we show a graph with five nodes and five undirected links. Based on the *degree centrality*, among all the nodes in the graph, node n_1 has the largest centrality score, i.e., 3, compared with the remaining nodes. Node n_3 has the smallest centrality score, i.e., 1, and the remaining nodes all have a centrality score of 2.

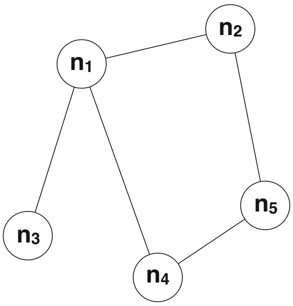


Fig. 3.5 An input graph example

3.3.2.2 Normalized Degree Centrality

Generally, the *degree-based centrality* in different networks is usually of different scale. For example, the Facebook network is of a much larger scale compared with other social networks, like Twitter and Foursquare, and the *degree-based centrality* in Facebook is usually much larger than that in Twitter and Foursquare. To ensure the comparability of the *degree-based centrality* across different networks, one method is to normalize all the centrality measures to a common value interval. Here, different numbers can be used as the denominator for centrality rescaling, e.g., the *maximal degree*, *sum degree*, and *maximum degree*, which will bring about different *normalized degree centrality* measures.

The maximal number of nodes each node can be connected within a network is $|\mathcal{V}| - 1$, which can be applied to rescale the *degree centrality* to the range [0, 1]. It actually helps define the *maximal degree-based normalized degree centrality*:

$$C_{\max}(u) = \frac{C(u)}{|\mathcal{V}| - 1}. \tag{3.15}$$

Another way to do the normalization will be to define the centrality as the ratio of the degrees with regard to the total degree in the networks, i.e., the *sum degree-based normalized degree centrality*:

$$C_{sum}(u) = \frac{C(u)}{\sum_{v \in \mathcal{V}} d(v)} = \frac{C(u)}{2 \times |\mathcal{E}|}. \tag{3.16}$$

Generally, in the online social networks, few nodes can achieve a degree with values $|\mathcal{V}| - 1$ or $2 \times |\mathcal{E}|$. In other words, these two normalized node degree centrality measure values are highly to be concentrated in a very narrow region $[0, \alpha]$ ($\alpha < 1$ and can be a very small number), where the α will also be different for different online social networks and violate the comparability objective. To resolve such a problem, we propose to normalize the measures with maximum node degree instead, which defines the *maximum degree-based normalized degree centrality*:

$$C_{maximum}(u) = \frac{C(u)}{\max_{v \in \mathcal{V}} d(v)}. \tag{3.17}$$

3.3.2.3 Eigen-Centrality

In the *degree centrality* definition, the users having more friends are assumed to be more important by default. However, in the real world, it can be not the case. Instead of having lots of online friends, *users having more important friends will be more important*. In other words, the users' *centrality* is determined by their online friends' *centrality* [46], i.e.,

$$C(u) = \frac{1}{\lambda} \sum_{v \in \Gamma(u)} C(v), \tag{3.18}$$

where set $\Gamma(u) = \{v | v \in \mathcal{V} \wedge (u, v) \in \mathcal{E}\}$ denotes the set of online neighbors of user u in the network G and λ is a constant scalar.

By organizing the social connections among users in the network as the social adjacency matrix $\mathbf{A} \in \{0, 1\}^{|\mathcal{V}| \times |\mathcal{V}|}$, we can rewrite the above equation as follows:

$$\lambda \mathbf{c} = \mathbf{A}^\top \mathbf{c}, \tag{3.19}$$

where vector $\mathbf{c} = [C(u_1), C(u_2), \ldots, C(u_{|\mathcal{V}|})]^\top$ contains all the centrality values of users in the network.

The above equation indicates that the centrality vector is actually a eigenvector of the social adjacency matrix $\mathbf{A}^\top$, whose corresponding eigenvalue is λ. However, given a matrix $\mathbf{A}^\top$, it will have multiple eigenvectors and eigenvalues. Usually, we prefer to use the positive values to define the centrality measure. According to the Perron–Frobenius theorem [42], given a matrix, there always exists a non-negative eigenvector of the matrix, which corresponds to the largest eigenvalue of $\mathbf{A}$. Therefore, we will use the eigenvector corresponding to the largest eigenvalue of matrix $\mathbf{A}^\top$ to define the *eigen-centrality* [11, 46].

Example 3.8 For example, given an undirected graph shown in Fig. 3.6a, we can represent the adjacency matrix of the undirected input graph as $\mathbf{A} = \begin{bmatrix} 0, 1, 1, 1, 0 \\ 1, 0, 1, 0, 1 \\ 1, 1, 0, 0, 1 \\ 1, 0, 0, 0, 1 \\ 0, 1, 1, 1, 0 \end{bmatrix}$. By decomposing the matrix, we can achieve the eigenvalues of matrix $\mathbf{A}$ to be $[2.856, -2.177, 1.429 \times 10^{-16}, 0.322, -1.0]$. Its largest eigenvalue is 2.856, and the corresponding eigenvector can be represented as

$$\mathbf{c} = \begin{bmatrix} 0.456 \\ 0.491 \\ 0.491 \\ 0.319 \\ 0.456 \end{bmatrix}, \tag{3.20}$$

which denotes the *centrality scores* achieved by the nodes in the graph.

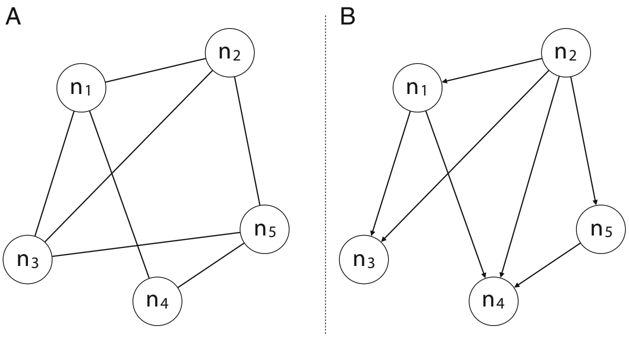


Fig. 3.6 A directed input graph example. (**a**) Undirected graph, (**b**) directed graph

In other words, nodes n_2 and n_3 actually have the largest centrality score among all the nodes in the graph, which is 0.491; the next group will be nodes n_1 and n_5 with a centrality score 0.456; and node n_4 has the lowest centrality score, which is 0.319.

Example 3.9 In Fig. 3.6b, we show an example of a directed input graph with different connections. According to the graph structure, we can represent the graph adjacency matrix as $\mathbf{A} = \begin{bmatrix} 0, 0, 1, 1, 0 \\ 1, 0, 1, 1, 1 \\ 0, 0, 0, 0, 0 \\ 0, 0, 0, 0, 0 \\ 0, 0, 0, 1, 0 \end{bmatrix}$. By decomposing the matrix, we can achieve its eigenvalues to be 0 for all the nodes in the graph, which may make the *eigen-centrality* fail to work in handling the directed graphs.

3.3.2.4 Katz Centrality

As shown in the previous example, when the networks are directed, the *eigen-centrality* measure may suffer from some serious problems. To overcome such a problem, a new centrality measure, the *Katz centrality* [11], has been proposed, which is defined as follows:

$$\mathbf{c} = \alpha \cdot \mathbf{A}^\top \mathbf{c} + \beta \cdot \mathbf{1}, \tag{3.21}$$

where parameters α and β denote the weights of the *eigen-centrality* and the bias term, respectively. In the case that matrix $\mathbf{I} - \alpha \cdot \mathbf{A}^\top$ is invertible, the *Katz centrality* vector can be formally represented as

$$\mathbf{c} = \beta \cdot (\mathbf{I} - \alpha \cdot \mathbf{A}^\top)^{-1} \cdot \mathbf{1}. \tag{3.22}$$

To ensure the invertibility of matrix $\mathbf{I} - \alpha \cdot \mathbf{A}^\top$, the choice of parameter α can be a little bit tricky. Smaller α tends to unify the *Katz centrality* of all the nodes in the network closer to the value of β, while larger α will reduce the effectiveness of the bias term. In practice, $\alpha < \frac{1}{\lambda_{max}}$ is usually selected, where λ_{max} denotes the maximum eigenvalue of matrix $\mathbf{A}^\top$.

Example 3.10 For instance, given the directed graph shown in Fig. 3.6b, by assigning the parameters $\alpha = \beta = 0.5$, we have the *Katz centrality* vector as follows:

$$\mathbf{c} = \begin{bmatrix} 1.0 \\ 1.875 \\ 0.5 \\ 0.5 \\ 0.75 \end{bmatrix}, \tag{3.23}$$

among which node n_2 has the largest *Katz centrality* (i.e., 1.875) in the input graph.

3.3.2.5 PageRank Centrality

Both *eigen centrality* and *Katz centrality* treat all the neighbor nodes in graphs equally when calculating the centrality scores for the target node. However, in the real world, the impacts of the neighbor nodes are usually different in determining users' centrality score. For example, in online social networks, users like to get connected with celebrities, and these celebrities will be connected with lots of people even though they may not necessary know each other in person. Usually, the celebrities are very important users in online social networks, and they have a large *centrality* score compared against the other users. However, for the users who are connected with these celebrities, we cannot say that they are also important as well. To consider such a phenomenon, a pagerank-based centrality measure has been introduced to provide different neighbors with different weights (determined by their degrees). Formally, the *pagerank centrality* [12] of user u can be defined as

$$C_p(u) = \alpha \cdot \sum_{v \in \Gamma(u)} \frac{C_p(v)}{|\Gamma(v)|} + \beta, \tag{3.24}$$

where the effects from u's neighbors, like $v \in \Gamma(u)$, are weighted by $\frac{1}{|\Gamma(v)|}$. Here, the subscript p denotes the *pagerank*-based *centrality* score.

In other words, for the neighbors with large degrees, their impacts on u will be penalized in the *centrality* score computation, while people with a small degree will have a greater impact on u instead. Formally, the above equation can be rewritten as follows:

$$\mathbf{c} = \alpha \cdot \mathbf{A}^\top \mathbf{D}^{-1} \mathbf{c} + \beta \cdot \mathbf{1}, \tag{3.25}$$

where matrix $\mathbf{D} = diag(d_{out}(u_1), d_{out}(u_2), \ldots, d_{out}(u_{|\mathcal{V}|}))$ is a diagonal matrix with the node out-degrees on its diagonal. In the case that matrices $\mathbf{D}$ and $(\mathbf{I} - \alpha \mathbf{A}^\top \mathbf{D}^{-1})$ are both invertible, we can have the *pagerank centrality* vector to be

$$\mathbf{c} = \beta \cdot (\mathbf{I} - \alpha \cdot \mathbf{A}^\top \mathbf{D}^{-1})^{-1} \cdot \mathbf{1}. \tag{3.26}$$

Parameter α can be selected with similar methods as introduced after Eq. (3.22).

Example 3.11 For example, we can take the directed graph shown in Fig. 3.7 as the input graph, and its *adjacency matrix* together with the *out-degree* diagonal matrix can be represented as

$$\mathbf{A} = \begin{bmatrix} 0, 0, 1, 1, 0 \\ 1, 0, 1, 1, 1 \\ 0, 0, 0, 0, 1 \\ 0, 0, 1, 0, 0 \\ 0, 0, 0, 1, 0 \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 2, 0, 0, 0, 0 \\ 0, 4, 0, 0, 0 \\ 0, 0, 1, 0, 0 \\ 0, 0, 0, 1, 0 \\ 0, 0, 0, 0, 1 \end{bmatrix}. \tag{3.27}$$

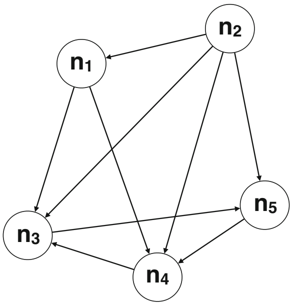


Fig. 3.7 An input graph for pagerank centrality calculation

By assigning $\alpha = \beta = 0.5$, we can compute the *pagerank centrality* scores of nodes in the graph to be

$$\mathbf{c} = \beta \cdot (\mathbf{I} - \alpha \cdot \mathbf{A}^\top \mathbf{D}^{-1})^{-1} \cdot \mathbf{1} = \begin{bmatrix} 0.563 \\ 0.5 \\ 1.406 \\ 1.969 \\ 0.563 \end{bmatrix}. \tag{3.28}$$

Among all the nodes, n_4 has the largest *pagerank centrality* score compared against the other nodes, and n_2 has the lowest *pagerank centrality* score on the other hand.

3.3.2.6 Betweenness Centrality

The centrality measures aforementioned are mostly defined based on the neighborhood information for the nodes. Another way to define the centrality measure is based on their positions connecting nodes in the networks, which is called the node *betweenness centrality* [11, 16] measure. Generally, if a node u effectively joins the connection paths among nodes in the network, then its position will be more important. Formally, the *betweenness centrality* measure of node u can be defined as

$$C_b(u) = \sum_{s,t \in \mathcal{V}, s \neq t \neq v} \frac{|\mathcal{P}_{s,t}(u)|}{|\mathcal{P}_{s,t}|}, \tag{3.29}$$

where $\mathcal{P}_{s,t}(u)$ denotes the set of *shortest paths* between nodes s and t via u in the network, and $\mathcal{P}_{s,t}$ represents the set of all *shortest paths* connecting s and t.

For a node u, it can achieve the maximum *between centrality* if it appears on all the shortest paths $\left(\text{i.e., } \frac{|\mathcal{P}_{s,t}(u)|}{|\mathcal{P}_{s,t}|} = 1\right)$ of all the node pairs in the network, like the central node in the star-structured graph. Formally, in such a case given a network with node set $\mathcal{V}$, the maximum *between centrality* node u achieves can be represented as

$$\begin{aligned} C_b^{\max}(u) &= \sum_{s,t \in \mathcal{V}, s \neq t \neq v} \frac{|\mathcal{P}_{s,t}(u)|}{|\mathcal{P}_{s,t}|} \\ &= \sum_{s,t \in \mathcal{V}, s \neq t \neq v} 1 \\ &= 2\binom{|\mathcal{V}| - 1}{2} \\ &= (|\mathcal{V}| - 1)(|\mathcal{V}| - 2). \end{aligned} \tag{3.30}$$

To ensure the *betweenness closeness* measure in different networks are comparable, one effective way will be to rescale the *betweenness centrality* to range [0, 1] with the maximum *between centrality* in the network.

$$C_{n-b}(u) = \frac{C_{n-b}(u)}{C_b^{\max}(u)} = \frac{\sum_{s,t \in \mathcal{V}, s \neq t \neq v} \frac{|\mathcal{P}_{s,t}(u)|}{|\mathcal{P}_{s,t}|}}{(|\mathcal{V}| - 1)(|\mathcal{V}| - 2)}, \tag{3.31}$$

To compute the shortest path between all pairs of nodes in a graph $G = (\mathcal{V}, \mathcal{E})$, algorithms, like the Floyd–Warshall algorithm, can be used with an $O(|\mathcal{V}|^3)$ time cost. In the exercises, we will have an example about the *betweenness centrality*, and the readers can try to compute the node centrality scores according to the above definitions.

3.3.3 Closeness

Via the connections, nodes in networks will be closely correlated with each other and have different closeness scores with each other. In this part, we will introduce several frequently used *closeness* [67] measures for the node pairs in networks. To illustrate the measures more clearly, we will use the social networks as an example, where the nodes denote the users and links represent the friendship connections.

3.3.3.1 Local Structure-Based Closeness Measures

Many node closeness measures can calculate the proximity among user nodes with the social network local structure information, like the shared common neighbors. In this part, we will introduce a number of local network structure-based user node closeness metrics as follows, which can effectively measure the social proximity scores among the users.

- **Reciprocity**: For the social networks involving directed links among the nodes (i.e., the link denotes the *follow* relationship), given a pair of nodes u and v in the network, there could exist a link between them inside the networks. For example, if user u follows v in the network, there will exist a directed link $u \rightarrow v$ (i.e., (u, v)) pointing from user u to user v. When measuring the closeness between users u and v, the connected user pairs are generally much closer to each other compared against the disconnected ones. Viewed in this perspective, if u follows v (or v follows u), such a link will indicate the strong closeness between these two users. Meanwhile, in the real-world online social networks, most users tend to follow the celebrities. The follow link between regular users and the celebrities may not necessarily denote they are close in the network, like the celebrities may not even know his/her followers.

  One measure that can denote the closeness between two users, e.g., u and v, in the social networks is the *reciprocal links* [22]. Given that user u follows v in the network (i.e., (u, v) exists in the network), if v also follows u back (i.e., (v, u) also exists), then u and v tend to be very close to each other. Here, link (v, u) will be called the *reciprocal link* of (u, v). The *reciprocal links* can also correctly measure the closeness between regular users and celebrities in social networks. For instance, if regular user u follows a celebrity v in a social network, and v also follows u via a *reciprocal link*, it can indicate that u and v tend to know each other and should be close to each other. Such a measure will also work for two regular users or two celebrity users.

  Formally, the *reciprocity closeness* measure between users u and v can be represented as

$$C_R(u, v) = \mathbb{I}((u, v) \in \mathcal{E} \wedge (v, u) \in \mathcal{E}), \tag{3.32}$$

  where $\mathcal{E}$ denotes the link set in the social network and $\mathbb{I}(\cdot)$ returns 1 if the condition can hold.

  Besides measuring the closeness between pairs of user nodes, the *reciprocity* can also be applied to measure the closeness of the whole network G, which can be represented as

$$\begin{aligned} C_R(G) &= \frac{\sum_{u,v \in \mathcal{V}, u \neq v} C_R(u, v)}{|\mathcal{V}|(|\mathcal{V}| - 1)} \\ &= \frac{\sum_{u,v \in \mathcal{V}, u \neq v} \mathbb{I}\left((u, v) \in \mathcal{E} \wedge (v, u) \in \mathcal{E}\right)}{|\mathcal{V}|(|\mathcal{V}| - 1)}. \end{aligned} \tag{3.33}$$

The *reciprocity* of a network denotes among all the potential user pairs in the network, how many percentages of them have the bi-directional follow links. For a network with a larger *reciprocity* score, the connections among users in the network will be stronger, which also indicates closer relationships among the internal nodes.

- **Common Neighbor**: Reciprocity is a closeness measure based on the connections between pairwise user nodes in the network. Actually, besides such pairwise links, via the connections with the other neighbors, many other closeness measures can be defined for user pairs in social networks as well, like the *common neighbor* (CN) [35, 67] closeness measure.

  Given two users $u, v \in \mathcal{V}$ in an undirected social network, if u and v share lots of common friends, it will indicate that they are highly likely to be close friends and may know each other. Formally, according to the introduction provided in the previous sections, we can formally represent the set of online friends whom users u, v have in the network as sets $\Gamma(u)$ and $\Gamma(v)$, respectively. The *common neighbor* closeness measure between users u and v can be formally represented as

  $$C_{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|. \tag{3.34}$$

  For the directed networks, we can define several more refined common neighbor measures, like common in-neighbors (i.e., the common followers), common out-neighbors (i.e., the common followees), common all-neighbors (i.e., the common connected neighbors regardless of the link directions), since the links among users will have a specific direction.

- **Jaccard's Coefficient**: Considering that $CN(u, v)$ can be a very large value merely because the two users both have a lot of neighbors rather than they are strongly related to each other. In other words, the common neighbor measure will have some problems when being used to compute the closeness between certain active users, e.g., the celebrities sharing lots of common fans. Furthermore, the common neighbor measure can neither be used to compare the closeness among the user pairs in different networks, due to the different network scales. One way to overcome these aforementioned problems will be to normalize the common neighbor measures with the users' degrees, which will introduce the following *Jaccard's coefficient* [24, 67] measure.

  Given the two users u and v in an undirected network, we can represent the *Jaccard's coefficient* closeness measure between them as

  $$C_{JC}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}, \tag{3.35}$$

  where the denominator denotes the number of users connected to either u or v. Therefore, for the celebrities, users, or networks with a relatively large scales, the user node closeness will be rescaled by assigning them with a larger penalty.

  In the case that the networks are directed, different other types of directed versions of Jaccard's coefficient measures can be defined, just like the directed *common neighbor* measures we define before. Jaccard's Coefficient can be treated as a weighted version of common neighbor, where each shared neighbor is assigned with an identical weight $\frac{1}{|\Gamma(u) \cup \Gamma(v)|}$. Many other weights can also be applied actually, like $\frac{1}{|\Gamma(u)|+|\Gamma(v)|}$ used in *Sørensen Index* [50], $\frac{1}{\min\{|\Gamma(u)|, |\Gamma(v)|\}}$ used in *Hub Promoted Index* [44], $\frac{1}{\max\{|\Gamma(u)|, |\Gamma(v)|\}}$ used in the *Hub Depressed Index* [80], and $\frac{1}{|\Gamma(u)| \times |\Gamma(v)|}$ in the *Leicht–Holme–Newman Index* [31].

- **Adamic/Adar**: Meanwhile, in measuring the closeness between users, different common users will play a different role and should have a different weight. To achieve such a goal, a closeness measure *Adamic/Adar* (AA) [1,67] index is proposed, which penalizes the shared neighbor nodes with larger degrees. Formally, the AA index between users u and v can be defined as

$$C_{AA}(u,v) = \sum_{w \in (\Gamma(u) \cap \Gamma(v))} \frac{1}{\log |\Gamma(w)|}. \tag{3.36}$$

For each of the common neighbor w shared by u and v, the weight assigned to w is $\frac{1}{\log |\Gamma(w)|}$ in AA. The shared common neighbors with smaller degrees will play an important role in indicating the closeness between the user pair. For the directed networks, by considering the link directions, several directed version of AA can be introduced as well. Besides AA, some other similar measures have been proposed, which assign the shared common neighbors with a different weight, like $\frac{1}{|\Gamma(w)|}$ used in the *Resource Allocation Index* (RA) [80].

3.3.3.2 Global Path-Based Closeness Measure

In addition to the local network structure-based closeness measures, many other closeness measures based on paths throughout the network have also been proposed to measure the proximity among the user nodes.

- **Shortest Path**: Generally, the social closeness among users can be measured by the distance among them in the network structure. Given two users who are far away from each other via all the potential paths connecting them (or they are isolated without any paths), they will have a very low closeness score. On the other hand, for the users who are directly connected via a link or a path of a very short length, they should be closer to each other compared with the isolated users. Based on such an intuition, we can define the closeness measure based on the distance of the *shortest path* [67] connecting users in the network:

$$C_{SP}(u,v) = \min\{|p|\}_{p \in \mathcal{P}_{u,v}}, \tag{3.37}$$

  where $\mathcal{P}_{u,v}$ represents the set of paths connecting users u and v inside the network, and $|p|$ denotes the distance of path p.
- **Katz**: Besides the shortest path, all the potential paths connecting user pairs in the networks can indicate their social closeness. Meanwhile, longer paths will show weaker closeness, and shorter paths denote stronger closeness. The *Katz* closeness measure [25,67] can integrate all these paths together to define the closeness scores among the users in the networks. Formally, the *Katz* closeness between users u and v can be defined as

$$C_{Katz}(u,v) = \sum_{l=1}^{l_{\max}} \beta^l |\mathcal{P}^l_{u,v}|, \tag{3.38}$$

  where $l_{\max}$ denotes the longest path connecting u and v, $\mathcal{P}^l_{u,v}$ denotes the set of paths of length l connecting u and v in the network. Parameter $\beta \in [0,1]$ is a regularizer term. Normally, smaller β favors shorter paths as β^l can decay very quickly as l increases when β is small, in which case the *Katz* measure will behave like the closeness measures based on the local neighbors introduced before.

3.3.3.3 Random Walk-Based Closeness Measure

In addition to the closeness measures that can be calculated from the network structure directly, there also exist another category of closeness measures that can calculate the closeness scores among users based on *random walk* [36, 67]. In this part, we will introduce the concept of *random walk* first, and provide the introduction to several closeness measures based on it, including *hitting time* [37, 67], *commute time* [33, 67], and *cosine similarity* [23, 67].

Formally, given a network $G = (\mathcal{V}, \mathcal{E})$, let matrix $\mathbf{A} \in \{0, 1\}^{|\mathcal{V}| \times |\mathcal{V}|}$ be the adjacency matrix of network G, where entry $A(i, j) = 1$ iff link $(u_i, u_j) \in \mathcal{E}$. The normalized matrix of $\mathbf{A}$ by rows can be represented as $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$, where diagonal matrix $\mathbf{D}$ of $\mathbf{A}$ has value $D(i, i) = \sum_j A(i, j)$ on its diagonal and $P(i, j)$ denotes the probability of stepping on node u_j from node u_i during the walk process. Let vector $\mathbf{x}^{(\tau)}(i)$ denote the probabilities that a random walker is located at user node $u_i \in \mathcal{V}$ at time τ. Then such a probability vector at time $\tau + 1$ will be updated as follows:

$$\mathbf{x}^{(\tau+1)}(i) = \sum_j \mathbf{x}^{(\tau)}(j) P(j, i). \tag{3.39}$$

In other words, the updating equation of vector $\mathbf{x}$ will be as follows, and such an updating process will continue until convergence, i.e.,

$$\text{Updating Equation: } \mathbf{x}^{(\tau+1)} = \mathbf{P}^\top \mathbf{x}^{(\tau)}, \tag{3.40}$$

$$\text{Convergence Equation: } \mathbf{x}^{(\tau+1)} = \mathbf{x}^{(\tau)}, \tag{3.41}$$

which will lead to the final stationary distribution vector $\mathbf{x}$ to be

$$\mathbf{x} = \mathbf{P}^\top \mathbf{x}. \tag{3.42}$$

The above equation denotes that the final stationary probability distribution vector $\mathbf{x}$ of random walk is actually an eigenvector of matrix $\mathbf{P}^\top$ corresponding to eigenvalue 1. Some existing works [15] have pointed out that if a Markov chain is *irreducible* and *aperiodic* then the largest eigenvalue of the transition matrix $\mathbf{P}^\top$ will be equal to 1 and all the other eigenvalues will be strictly less than 1. In addition, in such a condition, there will exist a unique stationary distribution which is vector $\mathbf{x}$ obtained at convergence of the updating equations. Here, we will not cover the proof to the above statement, which will be left as an exercise for the readers at the end of this chapter.

- **Hitting Time**: Let a variable $x^{(\tau)} = u$ denote that a random walker is at node u at step τ, and the *hitting time*-based closeness measure between users u and v can be represented as:

  $$C_{HT}(u, v) = \mathbb{E}(\{\tau | x^{(\tau)} = v \wedge x^{(0)} = u\}), \tag{3.43}$$

  where $\mathbb{E}(\cdot)$ denotes the expectation of the variable.

  Considering a random walker can reach v from u via different paths. The above equation denotes the expected number of steps to reach v from u, which is also called the *average hitting time* [37, 67]. Generally, close friends in the online social networks will have a small *average hitting time*.

Another way to define the *hitting time* between nodes u and v is to count the minimum number of steps needed to reach v from u, which can be represented as:

$$C_{mHT}(u, v) = \min\{\tau | x^{(\tau)} = v \wedge x^{(0)} = u\}, \tag{3.44}$$

which is also called the *minimum hitting time* measure.

- **Commute Time**: According to the above definition of *hitting time*, we can see that the measure is actually asymmetric, i.e., $C_{HT}(u, v) \neq C_{HT}(v, u)$, especially when the networks are directed. Such an asymmetric property will cause some problems when applying the *hitting time* in measuring the closeness among users in the real-world social networks. To overcome such a problem, some new measures, like *Commute Time* [33, 67], have been proposed, which counts the *hitting time* between user pairs from both of the directions, i.e.,

$$C_{CT}(u, v) = C_{HT}(u, v) + C_{HT}(v, u). \tag{3.45}$$

Formally, based on the adjacency matrix $\mathbf{A}$, we can define its corresponding Laplace matrix as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ ($\mathbf{D}$ is a diagonal matrix). The pseudo-inverse matrix of $\mathbf{L}$ can be represented as $\mathbf{L}^{\dagger}$, and the *commute time* for user pairs (u_i, u_j) can be represented as

$$\begin{aligned} C_{CT}(u_i, u_j) = 2|\mathcal{E}| \cdot (L^{\dagger}(i, i) \\ + L^{\dagger}(j, j) - 2L^{\dagger}(i, j)). \end{aligned} \tag{3.46}$$

The proof to the above equation will not be provided here, and more detailed information for the proof is available in [67].

- **Cosine Similarity**: With the pseudo-inverse matrix $\mathbf{L}^{\dagger}$, we can introduce a vector $\mathbf{z}_u = (\mathbf{L}^{\dagger})^{\frac{1}{2}}\mathbf{e}_u$ and vector $\mathbf{e}_u$ is a binary vector of 0s except the entries corresponding to node u which is filled with 1. According to existing works, the closeness between users u and v can be defined based on the *cosine similarity* [23, 67] measure of vectors $\mathbf{z}_u$ and $\mathbf{z}_v$ as follows:

$$C_{CS}(u, v) = \frac{\mathbf{z}_u^{\top}\mathbf{z}_v}{\sqrt{(\mathbf{z}_u^{\top}\mathbf{z}_u)(\mathbf{z}_v^{\top}\mathbf{z}_u)}}. \tag{3.47}$$

Furthermore, based on the pseudo-inverse matrix $\mathbf{L}^{\dagger}$, the above cosine similarity can be represented as

$$C_{CS}(u_i, u_j) = \frac{L^{\dagger}(i, j)}{\sqrt{L^{\dagger}(i, i) \cdot L^{\dagger}(j, j)}}. \tag{3.48}$$

These above closeness measures are all defined based on the regular *random walk* model. Meanwhile, in recent years, several variant *random walk* models have been proposed, which allow the walker to jump back to the starting point with a certain chance. Based on the definition of random walk, if the walker is allowed to return to the starting point with a probability of $1-c$, where $c \in [0, 1]$, then the new random walk method is formally defined as *random walk with restart* (RWR) [41], whose updating equation is shown as follows:

$$\mathbf{x}_u^{(\tau+1)} = c\mathbf{P}^{\top}\mathbf{x}_u^{(\tau)} + (1-c)\mathbf{e}_u, \tag{3.49}$$

where vector $\mathbf{x}_u^{(\tau+1)}$ denotes the probability of the random walker at all the nodes in the network starting from u initially.

By keeping updating the vector $\mathbf{x}_u^{(\tau+1)}$ until convergence, if matrix $(\mathbf{I} - c\mathbf{P}^\top)$ is invertible, we can have the stationary distribution vector of the RWR model to be

$$\mathbf{x}_u = (1-c)(\mathbf{I} - c\mathbf{P}^\top)^{-1}\mathbf{e}_u, \tag{3.50}$$

Furthermore, the closeness measure between user pairs u and v with the RWR model can be represented as

$$C_{RWR}(u, v) = \mathbf{x}_u(v), \tag{3.51}$$

where entry $\mathbf{x}_u(v)$ denotes the stationary probability of walking from u to v based on the RWR model.

3.3.4 Transitivity and Social Balance

The links in online social networks actually create various relationships among users. In this part, we will analyze several important properties about social networks based on the connections, which include *social transitivity* [19], *clustering coefficient* [5], and *social balance* [20, 57], respectively.

3.3.4.1 Social Transitivity

In discrete mathematics, a relation R on the domain $\mathcal{D}$ is a transitive relation iff $\forall u, v, w \in \mathcal{D}$ the following equation can hold:

$$R(u, v) \wedge R(v, w) \rightarrow R(u, w). \tag{3.52}$$

The transitive relation can also be used to describe the social connections among users in online social networks. In the real world, there is a social phenomenon that

> Friends of my friend can also be my friend.

Such a social phenomenon has been adopted in many friend recommender systems in online social networks for either recommendation or candidate pruning. Given three users $u, v, w \in \mathcal{V}$ in an online social network, if users u, v are friends, v, w are friends (i.e., links $(u, v), (v, w) \in \mathcal{E}$), and u, w also happen to be friends in the network, then we can observe a transitive friend relation among the three users. These three users together with the friendship connections among them will form a triangle. Therefore, to measure the transitivity of a network, the number of triangles existing in the network can be an important signal.

3.3.4.2 Clustering Coefficient

For a network with denser connections, there tend to be more triangles formed by the users in the network. We can measure how close a network compared to a complete network (i.e., a network with all node pairs connected) with the *clustering coefficient* concept. Formally, the network *clustering coefficient* [5] denotes among any three user nodes in the network, given that there exist two links connecting them already, how many of them will form triangles.

Formally, let set $\mathcal{P}^2 = \{(u, v, w) | u, v, w \in \mathcal{V} \wedge (u, v) \in \mathcal{E} \wedge (v, w) \in \mathcal{E}\}$ denote the node triples forming paths of length 2, and $\mathcal{T} = \{(u, v, w) | u, v, w \in \mathcal{V} \wedge (u, v) \in \mathcal{E} \wedge (v, w) \in \mathcal{E} \wedge (u, w) \in \mathcal{E}\}$

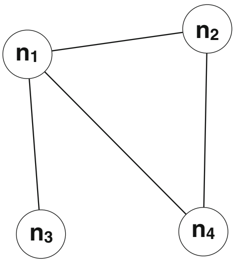


Fig. 3.8 An input graph for network clustering coefficient calculation

represent the set of node triples forming a triangle. We can represent the *clustering coefficient* of the network structure as follows:

$$CC = \frac{|\mathcal{T}|}{|\mathcal{P}^2|}. \tag{3.53}$$

Since in each triangle, there exist six different closed paths of length 2 and 2 different connected node triples in a path of length 2, the above equation can also be rewritten as follows:

$$\begin{aligned} CC &= \frac{\text{Number of triangles} \times 6}{|\mathcal{P}^2|} \\ &= \frac{\text{Number of triangles} \times 6}{\text{Number of connected triples of nodes} \times 2}, \end{aligned} \tag{3.54}$$

which can make the counting works simpler.

Example 3.12 In Fig. 3.8, we show an input graph with four nodes and four links. Among all these nodes, there exists one single triangle structure, i.e., the triangle involving n_1, n_2, and n_4. Meanwhile, there are five different paths of length 2, i.e., $n_1 - n_2 - n_4$, $n_2 - n_4 - n_1$, $n_4 - n_1 - n_2$, $n_3 - n_1 - n_2$, and $n_3 - n_1 - n_4$. Therefore, according to the above definition, we can calculate the *clustering coefficient* score of the network to be $\frac{1 \times 6}{5 \times 2} = \frac{3}{5}$.

3.3.4.3 Social Balance

Another concept strongly correlated with *transitivity* is *social balance* [20,57], which denotes whether a triangle social structure is balanced or not especially in *signed networks* [32,77]. A *signed network* denotes a social network, where the links are associated with polarities (either positive or negative). Depending on the specific network settings, the polarities attached to the links will have different physical meanings, like *trust* vs. *distrust* [63], *friend* vs. *enemy* [59], and *good attitude* vs. *bad attitude* [64].

The *social balance* theory describes the consistency of the signed connections among users. Some informal cases of *social balanced* structures in networks include:

> Friends of my friend can be my friend,
> Friends of my enemy can be my enemy,
> Enemies of my friend can be my enemy,
> Enemies of my enemies can be my friend.

Given three users $u, v, w \in \mathcal{V}$ in a network, we can represent the signs of relationships among them as $s_{u,v}$, $s_{v,w}$ and $s_{u,w}$, respectively. For instance, sign $s_{u,v} = +1$ denotes that users u and v are friends, while sign $s_{u,v} = -1$ denotes that users u and v are enemies. The relationships among these

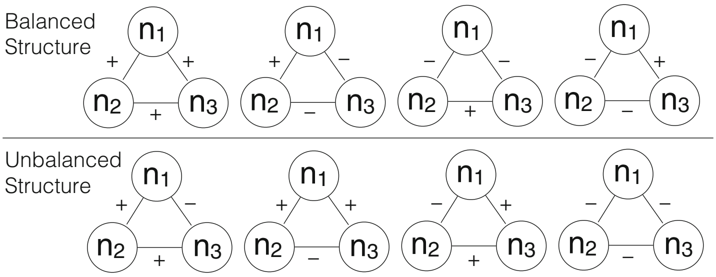


Fig. 3.9 Examples of structures based on the social balance theory

three users in the above four cases will form the *balanced structures*, and all the remaining structures among these three users are all called *unbalanced structure*.

Example 3.13 For instance, in Fig. 3.9, we provide an example about the *balanced* and *unbalanced* social structures formed by three users (i.e., n_1, n_2, and n_3). Among three users in a triangle, there can exist eight different social structure formed by them with signed links, which are shown in Fig. 3.9. In these eight cases, four of them are *balanced* (as shown at the top) and four are *unbalanced* (as shown at the bottom).

Actually, there exists a very simple method to determine whether a social structure is *balanced* or *unbalanced*. Based on the sign notations, the triangle formed by users u, v, w is a *balanced structure*, iff

$$s_{u,v} \cdot s_{v,w} \cdot s_{u,w} \geq 0. \tag{3.55}$$

Otherwise, the structure is said to be *unbalanced*.

3.4 Network Categories

The network concept introduced in the previous section can be used to model various types of network structured datasets available in the real world, including *online social networks* [38], *bibliographical networks* [52], *transportation networks* [6], and *computer networks* [10]. For instance, when we use the concept to define the *online social networks*, those various types of information entities in the social networks can be represented as the nodes, while the connections among the information entities are denoted as the links. Different online social networks are usually of different properties, and the corresponding network representations will have different kinds of characteristics as well.

For example, in some online social networks, the social connections among users can be (1) either *directed* (e.g., the social connections are the uni-directional follow links) or undirected (e.g., the social connections denote the bi-directional friendship links); (2) either *weighted* (e.g., users have different closeness scores with their friends) or *unweighted* (i.e., no closeness information is indicated in defining the social links); and (3) either *signed* (e.g., friendship links have different physical meanings actually and the link polarities denote different social attitudes) or *unsigned* (no social attitude information is provided in defining the social links).

Given a network $G = (\mathcal{V}, \mathcal{E})$, the nodes and links involved in it usually belong to different categories. Formally, we can represent the sets of node and link types involved in the network as

$\mathcal{N}$ and $\mathcal{R}$, respectively. Meanwhile, the corresponding network definition can be updated by adding the mappings indicating the node and link type information.

Definition 3.10 (Network) Formally, a network structured data can be represented as $G = (\mathcal{V}, \mathcal{E}, \phi, \psi)$, where $\mathcal{V}$, $\mathcal{E}$ are the sets of nodes and links in the network, and mappings $\phi : \mathcal{V} \rightarrow \mathcal{N}$, $\psi : \mathcal{E} \rightarrow \mathcal{R}$ project the nodes and links to their specific types, respectively. In many cases, the mappings ϕ, ψ are omitted assuming that the node and link types are known by default.

In this section, depending on the categories of information involved in the networks, we propose to categorize the network data into three groups: *homogeneous networks* [48], *heterogeneous networks* [51], and *multiple aligned heterogeneous networks* [29, 71–73], which will be introduced as follows, respectively.

3.4.1 Homogeneous Network

Definition 3.11 (Homogeneous Network) For a network $G = (\mathcal{V}, \mathcal{E}, \phi, \psi)$, if there exists one single type of nodes and one single type of links in the network (i.e., $|\mathcal{N}| = |\mathcal{R}| = 1$), then the network is called a *homogeneous network*.

Many different types of network structures can be represented as the *homogeneous networks* actually, like online social networks [38] involving users and friendship links only, company internal organizational network [74, 76, 78] involving employees and the management relationships, and computer networks [10] involving PCs and the internet connections. *Homogeneous networks* are one of the simplest network structures, analysis of which can provide many fundamental knowledge about networks with more complex structures. In the following part, we will introduce several common *homogeneous network* structures first.

3.4.1.1 Friendship Networks

Friendship network is one of the most common homogeneous social network structures, and they can be represented as the graph $G = (\mathcal{V}, \mathcal{E})$ defined before, where $\mathcal{V}$ represents the set of individuals while $\mathcal{E}$ denotes the set of social relationships among these individuals. Depending on whether the links in G are directed or undirected, the social links can denote either the *follow* links or *friendship* links among the individuals. Given an individual $u \in \mathcal{V}$ in an undirected friendship social network, the set of individuals connected to u can be represented as the friends of user u in the network G, denoted as $\Gamma(u) \subset \mathcal{V} = \{v|(u, v) \in \mathcal{E}\}$. The number of friends that user u has in the network is also called the degree of node u, i.e., $|\Gamma(u)|$.

Meanwhile, in a directed network G, the set of individuals followed by u (i.e., $\Gamma_{out}(u) = \{v|(u, v) \in \mathcal{E}\}$) are called the followees of u; and the set of individuals that follow u (i.e., $\Gamma_{in}(u) = \{v|(v, u) \in \mathcal{E}\}$) are called the followers of u. The number of users who follow u is called the in-degree of u, and the number of users followed by u is called the out-degree of u in the network. For the users with large out-degrees, they are called the *hubs* [27] in the network; while those with large in-degrees, they are called the *authorities* [27] in the network.

Example 3.14 In Fig. 3.10, we provide two examples of *friendship networks*, where plot (a) involves an undirected network and plot (b) contains a directed network. The links in plot (a) denote the friendship links, while those in plot (b) represent the follow links. Among all the users in plot (b),

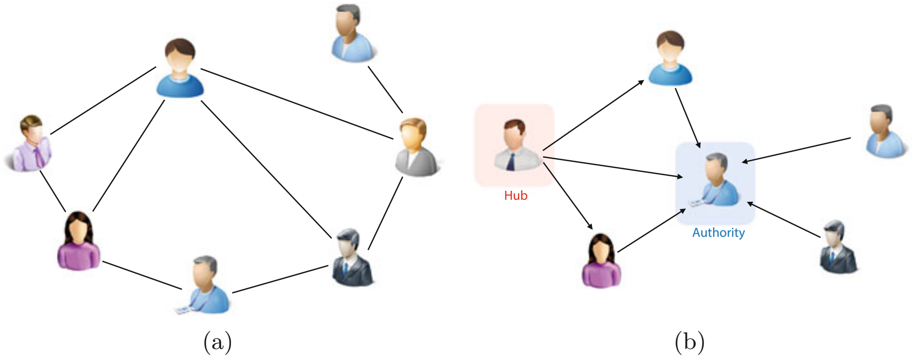


Fig. 3.10 Examples of friendship networks: (**a**) Undirected friendship network, (**b**) directed friendship network

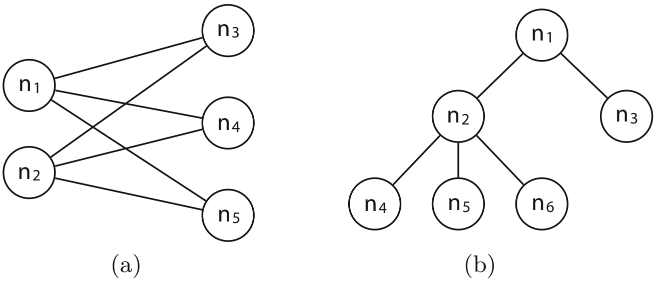


Fig. 3.11 Examples of homogeneous networks: (**a**) Bipartite network, (**b**) tree

we can identify one *authority user*, i.e., the one in blue square box with lots of in-links, and one *hub user*, i.e., the one in red square box with many out-links.

3.4.1.2 Computer Network

For the computer networks, like a local area network (LAN) or a wide area network (WAN), involving a set of computers and the access relationships among the computers, they can also be represented as the homogeneous networks as well. Generally, in a computer web, depending on the roles, the computers in the web network can serve as either the servers or the PCs. The PCs are the regular computers used by the end users, while the servers usually host some websites. The PCs can access the servers by visiting the websites or connecting with them via secure shell (SSH). If we don't consider the access relationships among the PCs and servers, respectively, then the computers together with their access relationships will form a *bipartite network* [65].

Definition 3.12 (Bipartite Computer Network) Formally, a *bipartite network* can be represented as $G = (\mathcal{V}_L \cup \mathcal{V}_R, \mathcal{E})$, where $\mathcal{V}_L$ and $\mathcal{V}_R$ denote the nodes on the left and right sides in the network and $\mathcal{E} \subset \mathcal{V}_L \times \mathcal{V}_R$ represents the access relationships between nodes on the left and right sides.

Example 3.15 An example of a *bipartite computer network* is shown in Fig. 3.11a, which involve five different nodes (two on the left and three on the right) and six links, where all the nodes on the left side are connected with the nodes on the right side. According to the above definition, the *bipartite network* can be formally represented as $G = (\{n_1, n_2\} \cup \{n_3, n_4, n_5\}, \{n_1, n_2\} \times \{n_3, n_4, n_5\})$.

3.4.1.3 Company Organizational Network

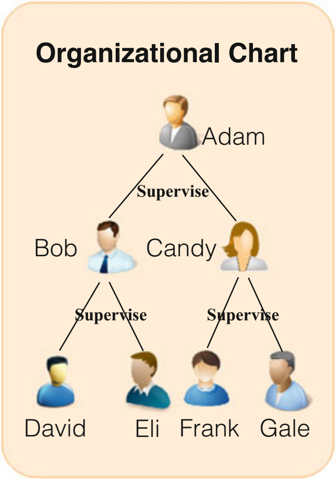
In many cases, the network structure or the sub-network structure of interest is a tree-structured diagram. Formally, in mathematics and computer science, *tree* is a special type of connected graph with no cycles formed by the nodes. As shown in Fig. 3.11b, for the nodes in *trees*, those with degree 1 are called the *leaf nodes* (i.e., the ones at the bottom) and the remaining ones are called *internal nodes*. The *tree* structured networks have several important properties, like *every tree has at least one edge and at least two nodes*, and *every tree with n nodes has exactly* $n - 1$ *links*. We will not provide the formal proof of these statements here. Tree is an important concept in networks representations, and many important network structures can be represented as a *tree* formally, like the *company organizational chart* as discussed in [74, 76, 78].

Definition 3.13 (Company Organizational Chart) Formally, a company management structure can be represented as a *rooted tree* $T = (\mathcal{V}, \mathcal{E}, root)$, where $\mathcal{V}$ and $\mathcal{E}$ denote the employees and management relationships among the employees in the company. Node $root \in \mathcal{V}$ usually denotes the CEO of the company.

Example 3.16 An example of the *company organizational chart* is shown in Fig. 3.12. As shown in the figure, in the *company organizational chart*, all the employees will have their managers except the CEO (i.e., Adam in the plot). The employees who are not in a management position (i.e., the leaf nodes) are named as the *base employees*. Different from the regular social networks, there generally exist no cycles in terms of management relationships in the *company organizational chart*. It is very important for companies, as a clear outline of the positions and responsibilities of the employees can avoid management confusion and chaos. What's more, in the *company organizational chart*, employees at higher levels can be connected to multiple lower-level employees, i.e., the *subordinates*, at the same time. Meanwhile, each employee at lower levels will be connected to one single employee at higher level, i.e., the *manager*. In other words, managers can manage multiple employees simultaneously, while each employee reports to one single manager.

Besides the *company organizational network*, many other networks can also be represented as tree structured diagram, like *ontologies* [18] outlining the relationships among different *categories of beings*, and the *cascades* [26] in information diffusion indicating how information propagates from the source users to the other users in the network.

Fig. 3.12 An example of company organizational chart



3.4.2 Heterogeneous Network

Definition 3.14 (Heterogeneous Network) For a network $G = (\mathcal{V}, \mathcal{E}, \phi, \psi)$, if there exist multiple types of nodes or links in the network (i.e., $|\mathcal{N}| > 1$, or $|\mathcal{R}| > 1$), then the network is called a *heterogeneous network*.

Most of the network structured data in the real world may contain very complex information involving multiple types of nodes and connections, which can be represented as the *heterogeneous networks* [51] formally. Representative examples include *heterogeneous social networks* [29, 54, 73] involving users, posts, check-ins, words, and timestamps, as well as the friendship links, write links, and the other links among these nodes; *bibliographic networks* [52] including authors, papers, conferences, and the write, cite, and publish-in links among them; and *movie knowledge libraries* [34] containing movies, casts, reviewers, review comments, as well as the complex links among these nodes. Many of the concepts introduced before for the *homogeneous networks* can also be applied to the *heterogeneous networks* as well.

3.4.2.1 Online Social Networks

The *online social networks* [29, 54, 73] usually allow the users to perform different social activities, like *make friends with other users*, *write posts online*, and *check-in at some places*, which will generate different kinds of information entities and very complex connections among these information entities. Formally, an *online social network* involving these diverse information entities and complex links is called a *heterogeneous social network*.

Example 3.17 In Fig. 3.13, we illustrate an example of a *heterogeneous social network*. Formally, according to the heterogeneous network definition, it can be represented as $G = (\mathcal{V}, \mathcal{E})$ (the mappings are not provided), where the node set $\mathcal{V}$ can be divided into several subsets $\mathcal{V} = \mathcal{U} \cup \mathcal{P} \cup \mathcal{L} \cup \mathcal{T}$ representing the *user*, *post*, *location*, and *timestamp* nodes, respectively. Meanwhile, depending on the node types that the links are connected to, the links in $\mathcal{E}$ can also have different physical meanings and can be further divided into subsets $\mathcal{E} = \mathcal{E}_{u,u} \cup \mathcal{E}_{u,p} \cup \mathcal{E}_{u,l} \cup \mathcal{E}_{p,t}$, which correspond to the friendship links among users, and the links between users and posts, locations and timestamps, respectively.

In the *heterogeneous social networks*, each node can be connected with a set of nodes belonging to different categories via various type of connections. For example, given a user $u \in \mathcal{U}$, the set of user node incident to u via the friendship links can be represented as the online friends of u, i.e., set $\{v | v \in \mathcal{U}, (u, v) \in \mathcal{E}_{u,u}\}$; the set of post node incident to u via the write links can be represented as the posts written by u, i.e., set $\{w | w \in \mathcal{P}, (u, w) \in \mathcal{E}_{u,p}\}$. It is very similar for the location and timestamp nodes as well, from which we can achieve the set of locations visited by users and the collection of timestamps that the users perform the social actions.

Many interesting research problems have been studied based on the *online social networks*, like *friend recommendation* [56, 61, 71, 72], *social community detection* [62, 68], *social information diffusion* [26, 75, 76, 79] via the connections among users. *Friend recommendation* problems aim at recommending online friends for users in the social networks, which can be formulated either as a ranking problem or as a link prediction problem. *Community detection* problem focuses on dividing the users into different social groups, where users who frequently interact with each other tend to appear in the same group. *Information diffusion* problems aim at modeling how information propagates within the online social networks, and when the users can be activated by certain information propagated from their friends. These problems mentioned here will also be covered in the following Chaps. 7–11 in great detail.

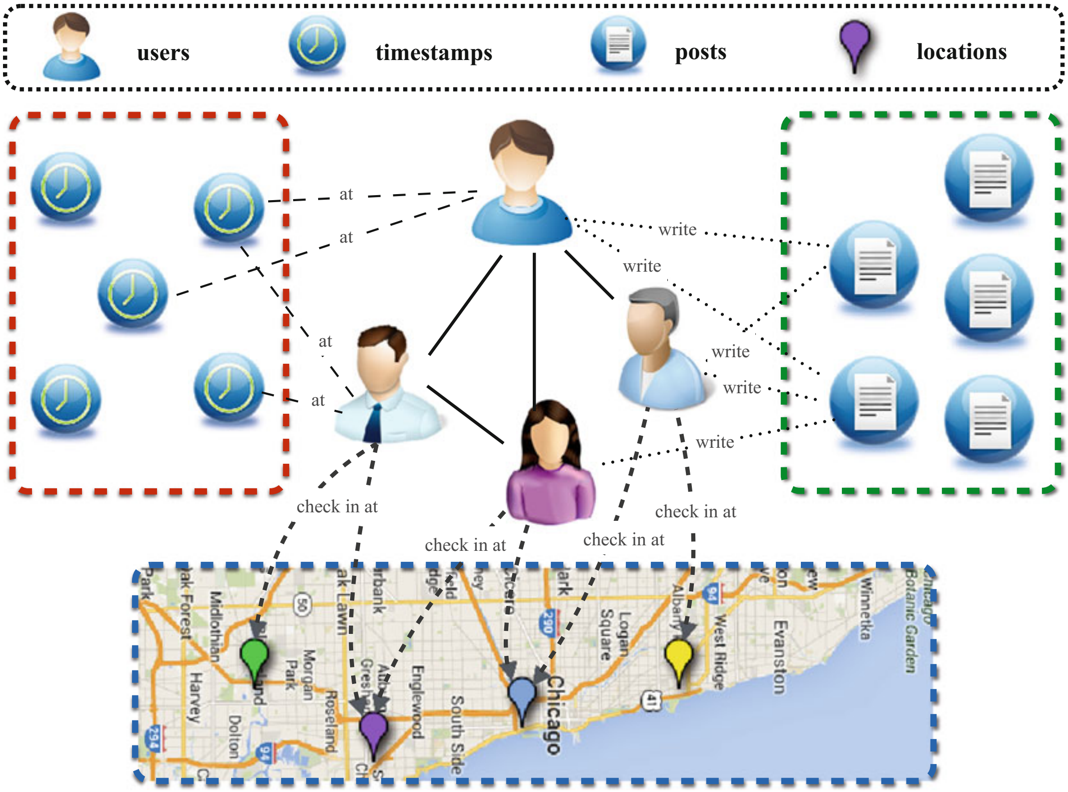


Fig. 3.13 An example of heterogeneous online social network

3.4.2.2 Bibliographic Networks

Another type of heterogeneous network well studied in research is called the *bibliographic networks* [52], which denote the academic networks depicting the paper authorship, paper citation, and paper publishing venues. Generally, the *bibliographic networks* may involve multiple types of information entities, like authors, papers, conferences/journals, and very complex connections among these information entities, which can be represented as a heterogeneous network as well.

Example 3.18 As shown in Fig. 3.14, a *bibliographic network* can be represented as graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{A} \cup \mathcal{P} \cup \mathcal{V}$ containing the authors, papers, and venues (i.e., conferences or journals), and $\mathcal{E} = \mathcal{E}_{a,p} \cup \mathcal{E}_{p,p} \cup \mathcal{E}_{p,v}$ involving the authorship links between authors and papers, citation links among the papers, and publishing links between papers and venues. In the example, MLI [73], MNA [29], and MFC [79] are the model names proposed by the authors in their papers.

In many cases, the information entities in a *bibliographic network* may also be associated with a set of attributes indicating their properties, like expertise/skills about the authors, the title, abstract, keywords, categories of papers, and the year, categories (like data mining, machine learning) of the publication venues. Via the papers, the authors can get correlated with each other. For instance, given a paper $p \in \mathcal{P}$, we can obtain the set of authors who are involved in writing p as $\{a | a \in \mathcal{A}, (a, p) \in \mathcal{E}_{a,p}\}$. For any author pairs $a_1, a_2 \in \{a | a \in \mathcal{A}, (a, p) \in \mathcal{E}_{a,p}\}$, they will be the co-authors on paper p. Similarly, from the *bibliographic network*, we can also obtain the set of papers published at certain

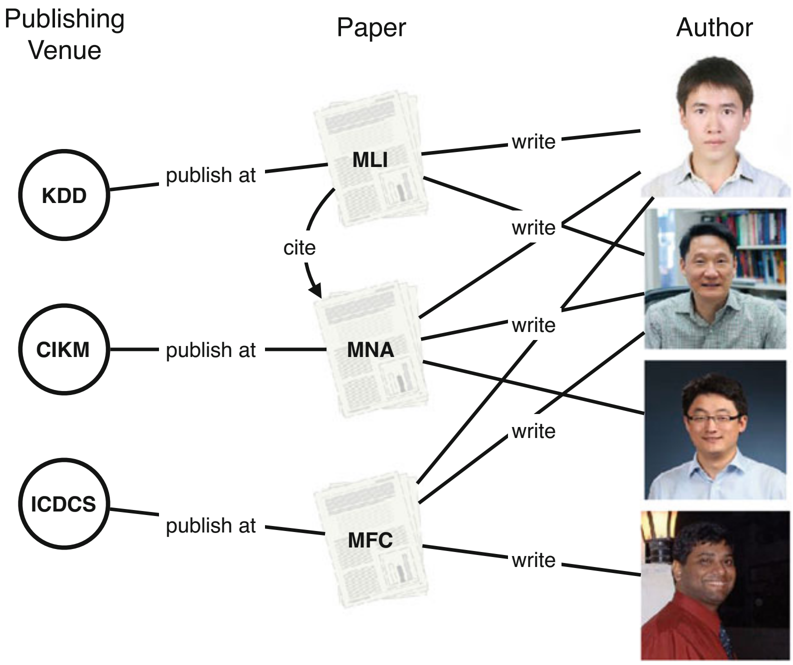


Fig. 3.14 An example of heterogeneous bibliographic network

venue $v \in \mathcal{V}$ as $\{p | p \in \mathcal{P}, (p, v) \in \mathcal{E}_{p,v}\}$. Many other interesting information, like authors who have ever published at a similar conference and conferences frequently participated in by certain authors, can be analyzed with the meta path concept to be introduced later as well.

Many interesting problems can be studied in the *bibliographic networks*, like co-author recommendation [52], rankings of authors, papers and venues [53], project team formation [30, 78]. Co-author recommendation is an important problem for academia, as it will help researchers find their collaborators to carry out the projects. The researchers, papers, and publishing venues are usually of different quality, some of which are highly ranked but some are of lower ranks. An effective ranking of the researchers, papers, and venues will make it easier for people to find qualified collaborators, related works, and publishing venues. Meanwhile, in the real world, great researchers tend to write innovative research papers and get them published at top-tier publishing venues. The ranking problems of the authors, papers, and venues are usually strongly correlated. To finish certain research projects, the project leader may need to build a team of researchers with different kinds of required skills. Team formation problems aim at identifying the team members for projects. In this book, we will mainly focus on the *social networks*, and these aforementioned problems for bibliographic networks will not be covered in this book. However, the readers are recommended to read the referred articles, if you are interested in these problems.

3.4.2.3 Movie Knowledge Libraries

For the online movie review sites, like IMDB[1] and Douban,[2] they involve very complex information and can be represented as heterogeneous networks as well [34]. Generally, in these sites, users can post review comments and ratings for the movies to express their favor regarding some movies. Meanwhile, for the movies, we can obtain the cast involved in producing the movies, like the writers, directors,

[1] http://www.imdb.com/.

[2] https://www.douban.com/.

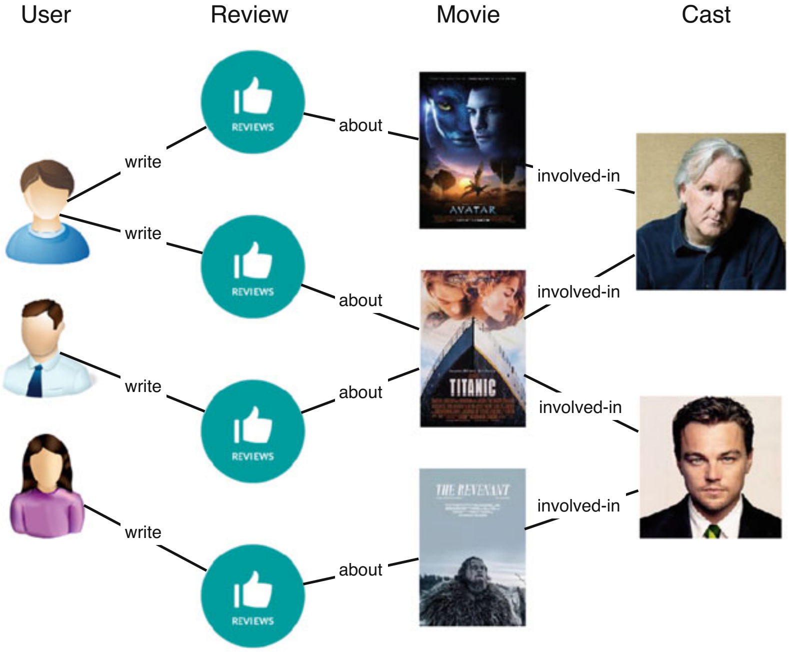


Fig. 3.15 An example of heterogeneous movie network

actors, and actress. A set of attributes can be obtained for the movies and casts as well, like movie title, story outline, movie genres, and cast profile information.

Example 3.19 In Fig. 3.15, we illustrate an example of the *heterogeneous movie knowledge library*, which can be represented as a graph $G = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V} = \mathcal{M} \cup \mathcal{C} \cup \mathcal{U} \cup \mathcal{P}$ involves the movie nodes, cast nodes, user nodes and review post nodes, and link set $\mathcal{E} = \mathcal{E}_{m,c} \cup \mathcal{E}_{u,p} \cup \mathcal{E}_{p,m}$ contains the links between movies and casts, users, and review posts, and those between the review posts and movies.

Via the heterogeneous links in these online *movie knowledge libraries* [34], the nodes are extensively connected with each other and lots of interesting knowledge can be discovered from the online *online knowledge libraries*. For example, given a movie $m \in \mathcal{M}$, we can obtain the set of reviews posted for it as set $\{p | p \in \mathcal{P}, (p, m) \in \mathcal{E}_{p,m}\}$, based on the review comments and ratings contained by these review posts, we can analyze the sentiment and favor of audiences about the movie.

Based on the online *movie knowledge libraries*, research problems like *movie recommendation* [7], *movie box-office analysis* [3, 34], and *movie planning problem* [34, 45] can be studied. The movie recommendation problems aim at recommending movies for users based on their movie rating historical records, and inferring their potential ratings for the recommended movies. From the investors' perspective, they generally want to invest their money on promising movies that can achieve a good box-office, while the movie box-office depends on various factors, like movie genre, movie storyline, and movie cast. Given a movie basic profile information, inferring the potential box-office can be obtained by them is an important problem. The movie planning problem is studying the correlation between movie profile information and box-office in a reverse direction, which aims at designing the optimal movie configurations within the provided budget to achieve the largest movie box-office.

3.4.3 Aligned Heterogeneous Networks

In the real world, about the same information entities, e.g., social media users, researchers in academia, and the imported foreign movies, a large amount of information can actually be collected from various sources. These sources are usually of different varieties, like Facebook and Twitter, data mining and machine learning research areas, the USA and China online movie libraries. Generally, these multiple information sources sharing some common information entities can be modeled as *multiple aligned heterogeneous networks* [29, 68, 72, 73].

Definition 3.15 (Multiple Aligned Heterogeneous Networks) Formally, the *multiple aligned heterogeneous networks* involving n networks can be defined as $\mathcal{G} = ((G^{(1)}, G^{(2)}, \ldots, G^{(n)}), (\mathcal{A}^{(1,2)}, \mathcal{A}^{(1,3)}, \ldots, \mathcal{A}^{(n-1,n)}))$, where the networks $G^{(1)}, G^{(2)}, \ldots, G^{(n)}$ denote these n heterogeneous networks and $\mathcal{A}^{(1,2)}, \mathcal{A}^{(1,3)}, \ldots, \mathcal{A}^{(n-1,n)}$ represent the sets of undirected *anchor links* aligning these networks.

In the above definition, the *anchor links* [29,73] refer to the mappings of information entities across different sources, which actually correspond to the same information entity in the real world, e.g., users shared between online social networks, authors involved in multiple bibliographic networks, and the common movies shared in different movie libraries. As proposed in [29, 69, 70], *anchor links* are usually subject to the *one-to-one* cardinality constraint, which can be formally defined as follows.

Definition 3.16 (Anchor Link) Given two heterogeneous networks $G^{(i)}$ and $G^{(j)}$ which share some common information entities, the set of *anchor links* connecting $G^{(i)}$ and $G^{(j)}$ can be represented as set $\mathcal{A}^{(i,j)} = \{(u_m^{(i)}, u_n^{(j)}) | u_m^{(i)} \in \mathcal{V}^{(i)} \wedge u_n^{(j)} \in \mathcal{V}^{(j)} \wedge u_m^{(i)}, u_n^{(j)}$denote the same information entity$\}$.

Example 3.20 In Fig. 3.16, we provide an example of *multiple aligned heterogeneous social networks*, which involve two heterogeneous networks Foursquare and Twitter, respectively. Both Foursquare and Twitter have very complex information, which can both be represented as the heterogeneous networks. Between these two networks, they share five common users, who are connected by the red dashed anchor links across networks.

Anchor links mainly exist between pairwise networks, when it comes to multiple (more than 2) aligned networks, there will exist a specific set of anchor links between any network pairs. The *anchor links* depict a transitive relationship among the information entities across different networks. Given three information entities $u_m^{(i)}, u_n^{(j)}, u_o^{(k)}$ from networks $G^{(i)}, G^{(j)}$, and $G^{(k)}$ respectively, if $u_m^{(i)}, u_n^{(j)}$ are connected by an anchor link and $u_n^{(j)}, u_o^{(k)}$ are connected by an anchor link, then the user pair $u_m^{(i)}$, $u_o^{(k)}$ will be connected by an anchor link by default.

For the information entities which are connected by the anchor links, they are named as the *anchor information entities*, like *anchor users* [73] in social networks, *anchor authors* in bibliographic networks, *anchor movies* in movie knowledge libraries. Meanwhile the remaining information entities are called the *non-anchor information entities*.

Definition 3.17 (Anchor Information Entities) Given a pair of heterogeneous networks $G^{(i)}$ and $G^{(j)}$, the anchor links $\mathcal{A}^{(i,j)}$ aligning them, and the information entity sets $\mathcal{V}^{(i)}$ and $\mathcal{V}^{(j)}$ involved in them, respectively, the set of *anchor information entities* in $G^{(i)}$ can be represented as $\mathcal{V}_a^{(i),(i,j)} = \{u_m^{(i)} | u_m^{(i)} \in \mathcal{V}^{(i)}, \exists u_n^{(j)} \in \mathcal{V}^{(j)}, (u_m^{(i)}, u_n^{(j)}) \in \mathcal{A}^{(i,j)}\}$. Similarly, we can also represent the set of anchor information entities in $G^{(j)}$ as $\mathcal{V}_a^{(j),(i,j)} \subset \mathcal{V}^{(j)}$.

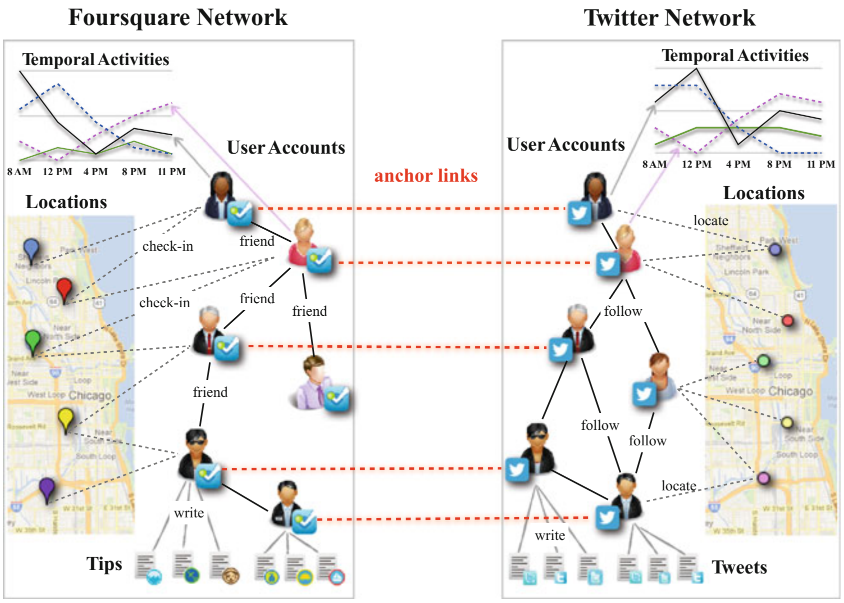


Fig. 3.16 An example of multiple aligned heterogeneous social networks

Definition 3.18 (Non-anchor Information Entities) Given a pair of heterogeneous networks $G^{(i)}$ and $G^{(j)}$, the anchor links $\mathcal{A}^{(i,j)}$ aligning them, and the information entity sets $\mathcal{V}^{(i)}$ and $\mathcal{V}^{(j)}$ involved in them, respectively, the set of *non-anchor information entities* in $G^{(i)}$ can be represented as $\mathcal{V}^{(i),(i,j)}_{non\text{-}a} = \{u^{(i)}_m | u^{(i)}_m \in \mathcal{V}^{(i)}, \forall u^{(j)}_n \in \mathcal{V}^{(j)}, (u^{(i)}_m, u^{(j)}_n) \notin \mathcal{A}^{(i,j)}\} = \mathcal{V}^{(i)} \setminus \mathcal{V}^{(i),(i,j)}_a$. In a similar way, we can represent the set of non-anchor information entities in network $G^{(j)}$ as well, which can be denoted as $\mathcal{V}^{(j),(i,j)}_{non\text{-}a}$.

The *anchor information entities* and *non-anchor information entities* concepts are defined based on the provided network pairs, which will be different (e.g., those in network $G^{(i)}$) as the network pair changes. For instance, the set of *anchor information entities* and *non-anchor information entities* in $G^{(i)}$ between network pairs $G^{(i)}$ and $G^{(j)}$ will be different from those in $G^{(i)}$ between network pairs $G^{(i)}$ and $G^{(k)}$. Furthermore, depending on the availability of *anchor information entities* and *non-anchor information entities*, the networks can be either *fully aligned*, *partially aligned*, and *non-aligned*, respectively.

Definition 3.19 (Full Alignment) Given a pair of heterogeneous networks $G^{(i)}$ and $G^{(j)}$ with non-anchor information entity sets $\mathcal{V}^{(i),(i,j)}_{non\text{-}a}$ and $\mathcal{V}^{(j),(i,j)}_{non\text{-}a}$, respectively. $G^{(i)}$ is said to be fully aligned with $G^{(j)}$ iff $\mathcal{V}^{(i),(i,j)}_{non\text{-}a} = \emptyset$, and $G^{(j)}$ is said to be fully aligned with $G^{(i)}$ iff $\mathcal{V}^{(j),(i,j)}_{non\text{-}a} = \emptyset$. $G^{(i)}$ and $G^{(j)}$ are said to be mutually fully aligned iff $\mathcal{V}^{(i),(i,j)}_{non\text{-}a} = \emptyset \wedge \mathcal{V}^{(j),(i,j)}_{non\text{-}a} = \emptyset$.

Network $G^{(i)}$ is said to be fully aligned with network $G^{(j)}$ if the information entities involved in $G^{(i)}$ are a subset of those involved in $G^{(j)}$, and vice versa. Networks $G^{(i)}$ and $G^{(j)}$ are mutually fully aligned if the information entities in $G^{(i)}$ and $G^{(j)}$ are actually identical. Fully aligned networks may exist in the real world, but a much common scenario will be *partial alignment* of networks instead.

Definition 3.20 (Partial Alignment) Given a pair of heterogeneous networks $G^{(i)}$ and $G^{(j)}$ with information entity sets $\mathcal{V}^{(i)}$ and $\mathcal{V}^{(j)}$ and anchor information entity sets $\mathcal{V}_a^{(i),(i,j)}$ and $\mathcal{V}_a^{(j),(i,j)}$, respectively. Network $G^{(i)}$ is partially aligned with network $G^{(j)}$ iff $\mathcal{V}_a^{(i),(i,j)} \neq \emptyset \wedge \mathcal{V}^{(i)} \neq \mathcal{V}_a^{(i),(i,j)}$, and vice versa. Networks $G^{(i)}$ and $G^{(j)}$ are said to be mutually partially aligned iff $\mathcal{V}_a^{(i),(i,j)} \neq \emptyset \wedge \mathcal{V}^{(i)} \neq \mathcal{V}_a^{(i),(i,j)}$ and $\mathcal{V}_a^{(j),(i,j)} \neq \emptyset \wedge \mathcal{V}^{(j)} \neq \mathcal{V}_a^{(j),(i,j)}$.

Network $G^{(i)}$ is said to be partially aligned with network $G^{(j)}$ if one part of the information entities in $G^{(i)}$ are involved in $G^{(j)}$. Both *full alignment* and *partial alignment* are not symmetric relationships. In the case that all the information entities in $G^{(i)}$ are also involved in $G^{(j)}$ while many information entities in $G^{(j)}$ are not involved in $G^{(i)}$, network $G^{(i)}$ will be fully aligned with $G^{(j)}$ but $G^{(i)}$ will be partially aligned with $G^{(i)}$ instead.

Definition 3.21 (Non-alignment) Given a pair of heterogeneous networks $G^{(i)}$ and $G^{(j)}$ with anchor information entity sets $\mathcal{V}_a^{(i),(i,j)}$ and $\mathcal{V}_a^{(j),(i,j)}$, respectively. Networks $G^{(i)}$ and $G^{(j)}$ are said to be non-aligned iff the information entities involved in two networks $G^{(i)}$ and $G^{(j)}$ are totally different, i.e., $\mathcal{V}_a^{(i),(i,j)} = \emptyset$ and $\mathcal{V}_a^{(j),(i,j)} = \emptyset$.

Different from *full alignment* and *partial alignment*, the *non-alignment* is a bi-directional relationships. In other words, if $G^{(i)}$ is non-aligned with $G^{(j)}$, then $G^{(j)}$ will be non-aligned with $G^{(i)}$ as well.

Lots of real-world network structures can actually share some common information entities, and can be represented as the *multiple aligned heterogeneous networks*. We will provide several examples as follows.

3.4.3.1 Multiple Aligned Heterogeneous Online Social Networks

To enjoy different kinds of social network services at the same time, users nowadays are usually involved in multiple online social networks simultaneously, e.g., Facebook, Twitter, Foursquare, and Google+. For the online social networks sharing common users, they can be represented as the *multiple aligned heterogeneous online social networks*.

Example 3.21 In Fig. 3.16, we have provided an example of two partially aligned heterogeneous online social networks: Foursquare and Twitter. Both Foursquare and Twitter can provide the users with different kinds of social network services, like make online friends with other users, write/like/comment on posts, check-in at some locations, and their online social activities are also associated with timestamps as well. Many users tend to join in Foursquare and Twitter at the same time, who are connected by the anchor links in the example.

In each of these two *aligned heterogeneous social networks*, we can have more data about the common users, which provides researchers and practitioners the opportunity to study users' social behaviors within these two networks. Moreover, the multiple aligned networks setting also allows the researchers to carry out a comparative study of users' social behaviors in different networks, which will provide a more comprehensive understanding about their social preferences and personal social behaviors.

3.4.3.2 Multiple Aligned Heterogeneous Bibliographic Networks

In the academia, the researchers are usually involved in various interdisciplinary projects and may collaborate with many researchers from other areas. For instance, the researchers of bioinformatics

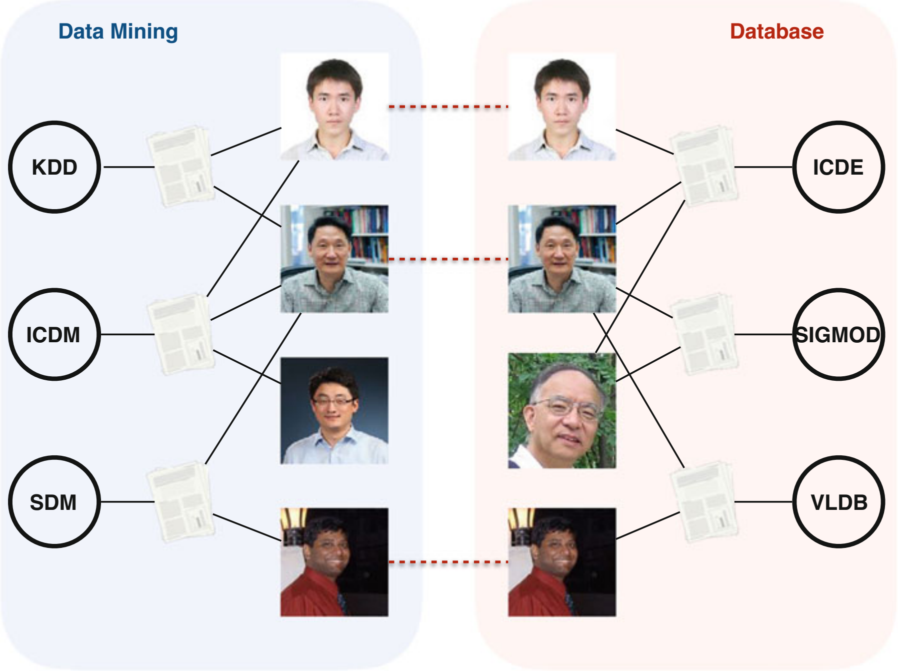


Fig. 3.17 An example of multiple aligned heterogeneous bibliographic networks

tend to have background in either computer science or biology; people working on data mining can publish works in either machine learning, data mining, or database; and the researchers working on neural networks can be experts on machine learning or neural science. Viewed in such a perspective, various closely related research areas may share lots of common researchers, and each researcher can also publish their works in different areas as well. Such complex relationships can be effectively modeled as the *multiple aligned heterogeneous bibliographic networks* formally.

In Fig. 3.17, we show an example of two partially aligned heterogeneous bibliographic networks in data mining and database. Between these two networks, there exist a large number of shared researchers, like Jiawei Zhang, Philip S. Yu, and Charu C. Aggarwal, who are active in both of these two areas and have published lots of academic papers in data mining and database conferences, like KDD, ICDM, SDM and ICDE, SIGMOD, VLDB. These two areas have different focuses in research actually, where data mining emphasizes more on knowledge discovery, while database is interested in data storage and management instead. Therefore, the researchers involved in these two areas are not exactly identical, and the shared researchers are indicated with the anchor links between them.

The multiple aligned heterogeneous bibliographic network setting allows us to study many interesting problems. In each of the networks, we can analyze the researchers' personal research interests, their preferred paper topics, frequently published conferences, which will be helpful to divide them into different research groups. Meanwhile, across the aligned bibliographic network, we can obtain their activities in different research areas. With the data about them across these different research domains, we can know their interdisciplinary research interest and activities, and it will provide extra information for us when studying researchers' personal cross-domain research interest shift, as well as their research progress in different domains.

3.4.3.3 Multiple Aligned Heterogeneous Online Movie Knowledge Libraries

To provide the movie related services in many different countries, lots of *online movie knowledge library* [34] exist on the web, like IMDB launched in the USA, Douban launched in China. Nowadays, to achieve more box-office, the movie import is a common practice between the movie markets in different countries. A movie can be on show in the USA first, and then get imported to show in China. Therefore, the IMDB and Douban *online movie knowledge library* can share lots of common movies, and can be modeled as the *multiple aligned heterogeneous online movie knowledge libraries* formally.

Example 3.22 In Fig. 3.18, we show an example of two partially aligned heterogeneous movie knowledge libraries in the USA and China: IMDB and Douban. Both IMDB and Douban have a very large collection of movies either native or imported from other countries. Lots of movies are very welcome and popular in both the USA and China, and are included in both IMDB and Douban, which act as the bridges aligning these different libraries together. For instance, in the example, these three provided movies, i.e., Avatar, Titanic, and The Revenant, exist in both Douban and IMDB, which make these two movie libraries fully aligned.

Generally, the common movie tend to have identical profile information in different movie knowledge libraries (can be in different languages), while they can receive the review comments and rating from the audience in different countries. These review comments and rating data obtained from different online movie knowledge libraries provide the opportunity to study the preferences of audiences from different countries about the shared movies. Moreover, many movies will be on show in the native countries first, and then get imported by other countries. Before these movies entering a new market, some prior knowledge about the movies in the original native country is available already, which will be very useful in scheduling the screenings in other countries, so as to maximize the revenue for theaters.

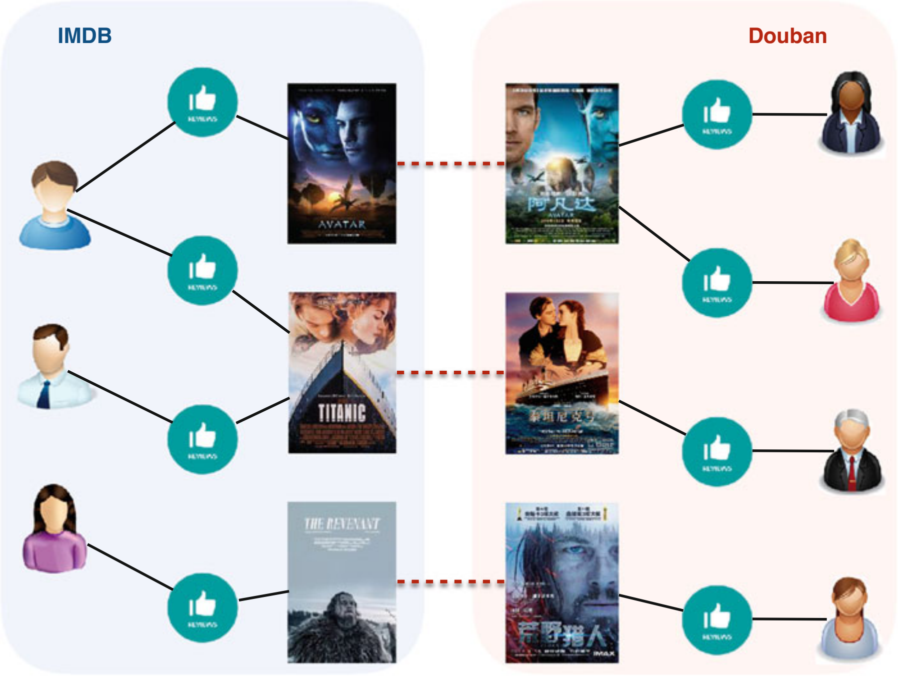


Fig. 3.18 An example of multiple aligned heterogeneous movie libraries

3.5 Meta Path

To deal with the social networks, especially those with heterogeneous information, a useful technique is the *meta path* [52,53,73]. *Meta path* is a concept defined based on the network schema, outlining the connections among nodes belonging to different categories. For the nodes which are not directly connected, their relationships can be depicted with the meta path concept. In this part, we will first introduce the meta path concept, and then talk about a set of meta paths within, as well as across real-world heterogeneous social networks.

3.5.1 Network Schema

Given a network $G = (\mathcal{V}, \mathcal{E})$, we can define its corresponding *network schema* [52,53,73] to describe the categories of nodes and links involved in G.

Definition 3.22 (Network Schema) Formally, the network schema of network $G = (\mathcal{V}, \mathcal{E}, \phi, \psi)$ can be represented as $S_G = (\mathcal{N}, \mathcal{R})$, where $\mathcal{N}$ and $\mathcal{R}$ denote the node type set and link type set of network G, respectively.

Network schema provides a meta level description of the network. Meanwhile, if a network G can be outlined by the network schema S_G, G is also called a *network instance* of the network schema. For a given node $u \in \mathcal{V}$, we can represent its corresponding node type as $\phi(u) = N \in \mathcal{N}$, and call u as an instance of node type N, which can also be represented as $u \in N$ for simplicity. Similarly, for a link (u, v), we can denote its link type as $\psi((u, v)) = R \in \mathcal{R}$. To represent that link (u, v) is an instance of the link type R, we can use the notations like $(u, v) \in R$, or $(u, v) \in S \xrightarrow{R} T$ for simplicity, where $\phi(u) = S \in \mathcal{N}$ and $\phi(v) = T \in \mathcal{N}$. The inverse relation type R^{-1} holds naturally for $T \xrightarrow{R^{-1}} S$, and R is generally not equal to R^{-1}, unless R is symmetric.

Example 3.23 In Fig. 3.19, we show the network schema of the heterogeneous social network on the left. According to the network structure, there exist four different node types, i.e., user, post, time, location, and four link types, i.e., follow, write, at, check-in at, in the network. These node types and link types together define the input network schema.

Meanwhile, in Figs. 3.20 and 3.21, we provide the network schemas of the input heterogeneous bibliographical network and the heterogeneous movie knowledge library, respectively. According to the bibliographical network structure, there exist three different node types and three link types, respectively. The movie knowledge library has a more complex structure, involving five different node types and four link types.

3.5.2 Meta Path in Heterogeneous Social Networks

Meta path [52,53,73] is a concept defined based on the network schema denoting the correlation of nodes based on the heterogeneous information (i.e., different types of nodes and links) in the networks.

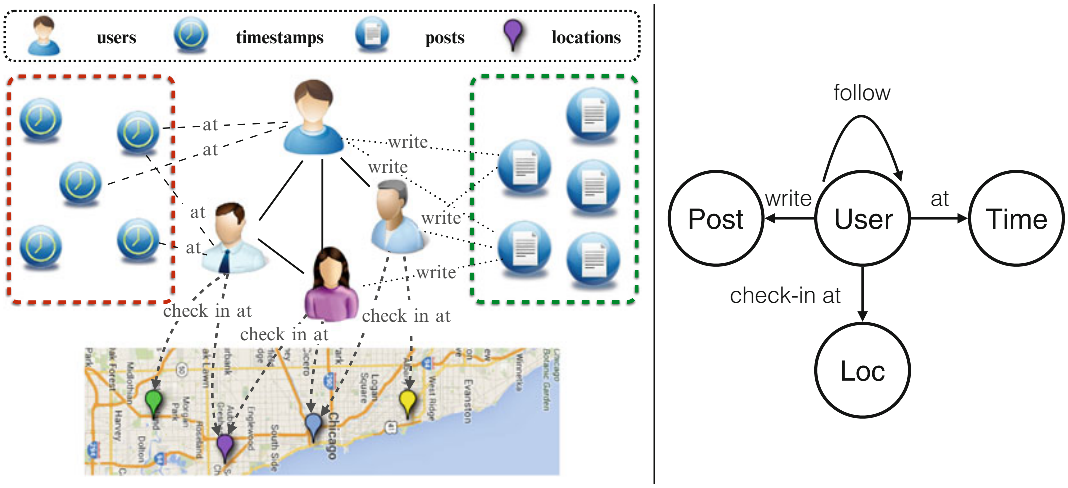


Fig. 3.19 An example of heterogeneous social network schema

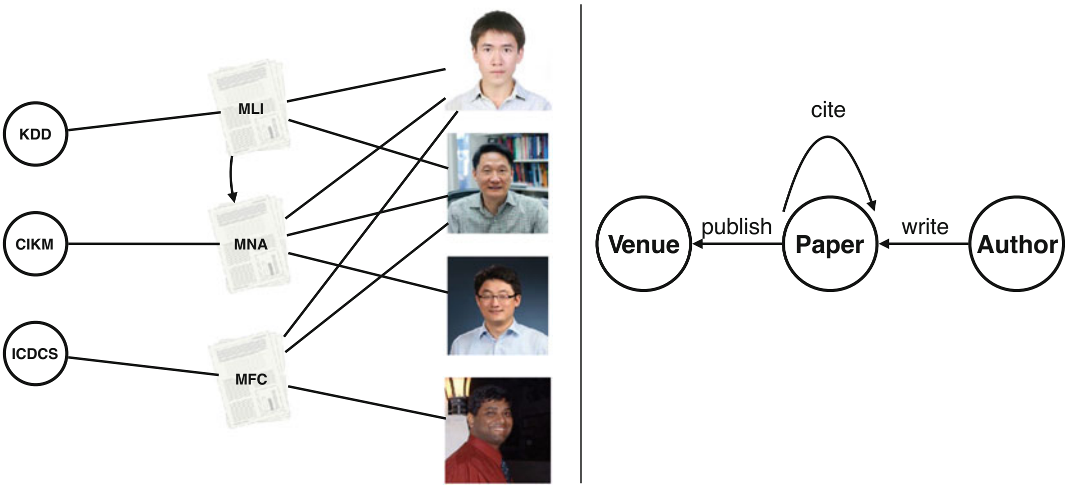


Fig. 3.20 An example of heterogeneous bibliographical network schema

Definition 3.23 (Meta Path) A meta path P defined based on the network schema $S_G = (\mathcal{N}, \mathcal{R})$ can be represented as $P = N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots N_{k-1} \xrightarrow{R_{k-1}} N_k$, where $N_i \in \mathcal{N}, i \in \{1, 2, \ldots, k\}$ and $R_i \in \mathcal{R}, i \in \{1, 2, \ldots, k-1\}$.

Furthermore, depending on the categories of node and link types involved in the meta path, we can specify the meta path concept into two refined groups, like *homogeneous meta path* [73] and *heterogeneous meta path* [73].

Definition 3.24 (Homogeneous/Heterogeneous Meta Path) Let $P = N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots N_{k-1} \xrightarrow{R_{k-1}} N_k$ denote a meta path defined based on the network schema $S_G = (\mathcal{N}, \mathcal{R})$. If

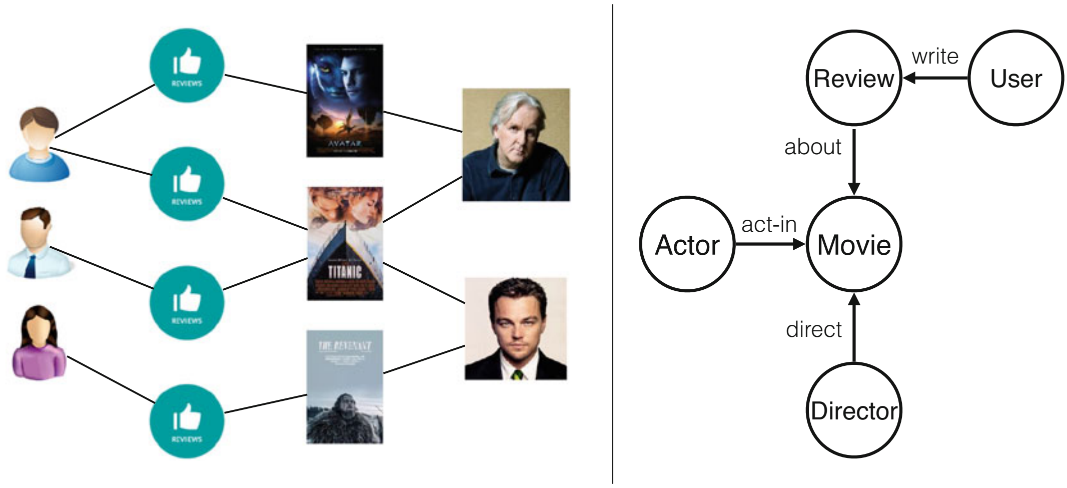


Fig. 3.21 An example of heterogeneous movie library schema

all the node types and link types involved in P are of the same category, P is called a *homogeneous meta path*; otherwise, P is called a *heterogeneous meta path*.

The meta paths connect any kinds of node type pairs, and specifically, for the meta paths starting and ending with the user node types within the same network, such a meta path is called the *social meta paths* [73].

Definition 3.25 (Social Meta Path) Let $P = N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots N_{k-1} \xrightarrow{R_{k-1}} N_k$ denote a meta path defined based on the network schema $S_G = (\mathcal{N}, \mathcal{R})$. If the starting and ending node types N_1 and N_k are both the user node type, P is called a *social meta path*.

Users are usually the main focus in social network studies, and the *social meta paths* connecting the user node type will be frequently used in both research and real-world applications and services. If all the node types in the meta paths are user node type and the link types are also of an identical category, then the meta path is called the *homogeneous social meta path*. The number of path segments in the meta path is called the meta path length. For instance, the length of meta path $P = N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots N_{k-1} \xrightarrow{R_{k-1}} N_k$ is $k - 1$. Meta paths can also been concatenated together with the *meta path composition operator* [52, 53, 73].

Definition 3.26 (Meta Path Composition) Meta paths $P^1 = N_1^1 \xrightarrow{R_1^1} N_2^1 \xrightarrow{R_2^1} \cdots N_{k-1}^1 \xrightarrow{R_{k-1}^1} N_k^1$, and $P^2 = N_1^2 \xrightarrow{R_1^2} N_2^2 \xrightarrow{R_2^2} \cdots N_{l-1}^2 \xrightarrow{R_{l-1}^2} N_l^2$ can be concatenated together to form a longer meta path $P = P^1 \circ P^2 = N_1^1 \xrightarrow{R_1^1} \cdots \xrightarrow{R_{k-1}^1} N_k^1(\text{or } N_1^2) \xrightarrow{R_1^2} N_2^2 \xrightarrow{R_2^2} \cdots N_{l-1}^2 \xrightarrow{R_{l-1}^2} N_l^1$, if the ending node type of P^1 is the same as the starting node type of P^2, i.e., $N_k^1 = N_1^2$. The new composed meta path will be of length $k + l - 2$.

Meta path $P = N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots N_{k-1} \xrightarrow{R_{k-1}} N_k$ can also been treated as the concatenation of simple meta paths $N_1 \xrightarrow{R_1} N_2$, $N_2 \xrightarrow{R_2} N_3$, …, $N_{k-1} \xrightarrow{R_{k-1}} N_k$, which can be represented as $P = R_1 \circ R_2 \circ \cdots \circ R_{k-1} \circ R_k$. Here, we use the link type to denote the simplest meta paths of length 1.

Example 3.24 For instance, based on the network schemas shown in Figs. 3.19, 3.20, and 3.21, a group of meta paths can be defined. Here, we can provide a group of them as follows, which mainly connect the user/author/movie pairs specifically.

1. *Heterogeneous Social Network*
   - User $\xrightarrow{follow}$ User (or $U \rightarrow U$), which denotes a simple *follow* meta path.
   - User $\xleftarrow{follow}$ User $\xrightarrow{follow}$ User (or $U \leftarrow U \rightarrow U$), which denotes a *common follower* meta path.
   - User $\xrightarrow{follow}$ User $\xleftarrow{follow}$ User (or $U \rightarrow U \leftarrow U$), which denotes a *common followee* meta path.
   - User $\xrightarrow{check\text{-}in\ at}$ Location $\xleftarrow{check\text{-}in\ at}$ User (or $U \rightarrow L \leftarrow U$), which denotes a *common location check-in* meta path.
2. *Heterogeneous Bibliographic Network*
   - Author $\xrightarrow{write}$ Paper $\xleftarrow{write}$ User (or $A \rightarrow P \leftarrow A$), which denotes a *co-author* meta path.
   - Author $\xrightarrow{write}$ Paper $\xrightarrow{publish\ at}$ Venue $\xleftarrow{publish\ at}$ Paper $\xleftarrow{write}$ Author (or $A \rightarrow P \rightarrow V \leftarrow P \leftarrow A$), which denotes a *common publishing venue* meta path.
   - Author $\xrightarrow{write}$ Paper $\xrightarrow{cite}$ Paper $\xleftarrow{write}$ Author (or $A \rightarrow P \rightarrow P \leftarrow A$), which denotes a *citation* meta path.
3. *Heterogeneous Movie Library*
   - Movie $\xleftarrow{about}$ Review $\xleftarrow{write}$ User $\xrightarrow{write}$ Review $\xrightarrow{about}$ Movie (or $M \leftarrow R \leftarrow U \rightarrow R \rightarrow M$), which denotes a *shared review author* meta path.
   - Movie $\xleftarrow{direct}$ Director $\xleftarrow{direct}$ Movie (or $M \leftarrow D \rightarrow M$), which denotes a *shared director* meta path.
   - Movie $\xleftarrow{act\text{-}in}$ Actor $\xleftarrow{act\text{-}in}$ Movie (or $M \leftarrow A \rightarrow M$), which denotes a *shared actor* meta path.

   Besides these meta paths shown above, many other meta paths can also be defined based on the network schema structures, which will not be provided here and the readers can try to define some other useful meta paths on your own.

3.5.3 Meta Path Across Aligned Heterogeneous Social Networks

Besides the meta paths within a network, the meta paths can also be defined across multiple aligned heterogeneous networks via the *anchor meta path* [73] (or the anchor link type).

Definition 3.27 (Anchor Meta Path) Let $G^{(1)}$ and $G^{(2)}$ be two aligned heterogeneous networks sharing the common anchor information entity of types $N^{(1)} \in \mathcal{N}^{(1)}$ and $N^{(2)} \in \mathcal{N}^{(2)}$, respectively. The anchor meta path between the schemas of networks $G^{(1)}$ and $G^{(2)}$ can be represented as meta path $\Phi = N^{(1)} \xleftrightarrow{Anchor} N^{(2)}$ of length 1.

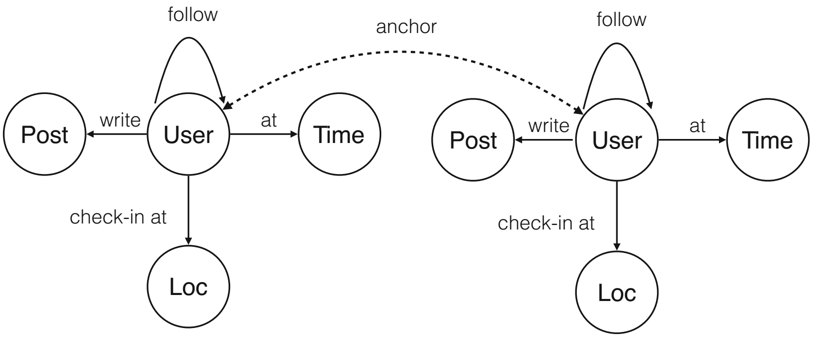


Fig. 3.22 An example of aligned heterogeneous social network schema

Formally, via the *anchor meta path*, given one pair of input *aligned heterogeneous social network* as shown in Fig. 3.16, we can formally represent the network schemas in Fig. 3.22. The *anchor meta path* is the simplest meta path across aligned networks, and a set of *inter-network meta paths* [73] can be defined based on the intra-network meta paths and the anchor meta path.

Definition 3.28 (Inter-Network Meta Path) Given a meta path $\Psi = N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots N_{k-1} \xrightarrow{R_{k-1}} N_k$, Ψ is an *inter-network meta path* between networks $G^{(1)}$ and $G^{(2)}$ iff $\exists m \in \{1, 2, \ldots, k\}, R_m = Anchor$.

The *inter-network meta paths* can be viewed as a composition of *intra-network meta paths* and the *anchor meta path*. An *inter-network meta path* can be a meta path starting with an *anchor meta path* followed by the *intra-network meta paths*, or those with *anchor meta paths* in the middle and starting/ending with the *intra-network meta paths*. Here, we would like to introduce several categories *inter-network meta paths* involving the anchor meta paths at different positions as follows [73]:

- $\Psi(G^{(1)}, G^{(2)}) = \Phi(G^{(1)}, G^{(2)})$, which denotes the set of simplest *inter-network meta paths* composed of the anchor meta path only between networks $G^{(1)}$ and $G^{(2)}$.
- $\Psi(G^{(1)}, G^{(2)}) = \Phi(G^{(1)}, G^{(2)}) \circ P(G^{(2)})$, which denotes the set of *inter-network meta paths* starting with anchor meta path and followed by the intra-network meta path in network $G^{(2)}$ connected by an anchor meta path between networks $G^{(1)}$ and $G^{(2)}$.
- $\Psi(G^{(1)}, G^{(2)}) = P(G^{(1)}) \circ \Phi(G^{(1)}, G^{(2)})$, which denotes the set of *inter-network meta paths* starting with the intra-network meta path in network $G^{(1)}$ followed by an anchor meta path between networks $G^{(1)}$ and $G^{(2)}$.
- $\Psi(G^{(1)}, G^{(2)}) = P(G^{(1)}) \circ \Phi(G^{(1)}, G^{(2)}) \circ P(G^{(2)})$, which denotes the set of *inter-network meta paths* starting and ending with the intra-network meta path in networks $G^{(1)}$ and $G^{(2)}$, respectively, connected by an anchor meta path between networks $G^{(1)}$ and $G^{(2)}$.
- $\Psi(G^{(1)}, G^{(2)}) = P(G^{(1)}) \circ \Phi(G^{(1)}, G^{(2)}) \circ P(G^{(2)}) \circ \Phi(G^{(2)}, G^{(1)})$, which denotes the set of *inter-network meta paths* starting and ending with node types in network $G^{(1)}$ and traverse across the networks twice via the anchor meta path.
- $\Psi(G^{(1)}, G^{(2)}) = P(G^{(1)}) \circ \Phi(G^{(1)}, G^{(2)}) \circ P(G^{(2)}) \circ \Phi(G^{(2)}, G^{(1)}) \circ P(G^{(1)})$, which denotes the set of *inter-network meta paths* starting and ending with the *intra-network meta paths* in network $G^{(1)}$ and traverse across the networks twice via the anchor meta path between them.

Example 3.25 Based on the above descriptions, we can also represent several examples of *inter-network meta paths* across social networks as follows:

- $\text{User}^{(1)} \xleftrightarrow{anchor} \text{User}^{(2)} \xrightarrow{follow} \text{User}^{(2)}$ (or $U^{(1)} \leftrightarrow U^{(2)} \rightarrow U^{(2)}$).
- $\text{User}^{(1)} \xrightarrow{follow} \text{User}^{(1)} \xleftrightarrow{anchor} \text{User}^{(2)} \xleftarrow{follow} \text{User}^{(2)}$ (or $U^{(1)} \rightarrow U^{(1)} \leftrightarrow U^{(2)} \leftarrow U^{(2)}$).
- $\text{User}^{(1)} \xleftarrow{follow} \text{User}^{(1)} \xleftrightarrow{anchor} \text{User}^{(2)} \xrightarrow{follow} \text{User}^{(2)}$ (or $U^{(1)} \leftarrow U^{(1)} \leftrightarrow U^{(2)} \rightarrow U^{(2)}$).
- $\text{User}^{(1)} \xrightarrow{check\text{-}in\ at} \text{Location}^{(1)} \xleftarrow{check\text{-}in\ at} \text{User}^{(1)} \xleftrightarrow{anchor} \text{User}^{(2)} \xrightarrow{follow} \text{User}^{(2)}$ (or $U^{(1)} \rightarrow L^{(1)} \leftarrow U^{(1)} \leftrightarrow U^{(2)} \rightarrow U^{(2)}$).

Generally, shorter meta paths may convey more concrete physical meanings compared with the long meta paths. Due to the extensive connections among nodes in networks, extremely long meta paths will may not be useful, since almost all the node pairs in the network can be connected by such meta path instances. In the following parts, we will introduce several meta path-based network measures about node degree, node centrality, and node pair closeness, respectively.

3.5.4 Meta Path-Based Network Measures

The meta path concept introduced above provides a meta level description of information available within and across networks, and they can be used to compute various node and link measures based on the heterogeneous social networks. All the degree, centrality, and closeness measures introduced in the previous subsections are mainly based on the direct social links among users in homogeneous networks. In this part, we will extend these measures to the multiple aligned heterogeneous networks scenario based on the meta path concept specifically.

3.5.4.1 Meta Path-Based Node Degree

Via the meta paths, nodes in the networks which are not directly connected can be extensively correlated with each other. In this part, we will take the user node as an example, and try to study how the users are connected with each other via the meta paths. Let $\mathcal{U}^{(1)}$ be a user set in network $G^{(1)}$, and $\mathcal{P}$ be the set of various meta paths starting and ending with the user node type in network $G^{(1)}$ (which can be either intra-network or inter-network meta paths).

For each user pair in network $G^{(1)}$, e.g., $u, v \in \mathcal{U}^{(1)}$, based on one specific meta path $P_k \in \mathcal{P}$, we can denote the set of concrete meta path instances connecting u and v as set $P_k(u, v)$. The number of user nodes that u is connected with, i.e., its degree, based on meta path $P_k \in \mathcal{P}$ can be denoted as

$$D_{P_k}(u) = \sum_{v \in \mathcal{U}^{(1)}} |P_k(u, v)|. \tag{3.56}$$

Furthermore, for all these meta paths in set $\mathcal{P}$, we can represent the degree vector of user u as a $|\mathcal{P}|$-dimensional degree distribution vector

$$\mathbf{D}_{\mathcal{P}}(u) = [D_{P_1}(u), D_{P_2}(u), \ldots, D_{P_{|\mathcal{P}|}}(u)]^\top. \tag{3.57}$$

Generally, for these different meta paths in the set $\mathcal{P}$, they are usually of different weights. For instance, in some scenarios, shorter meta paths can denote stronger connections among users than longer meta paths; meta paths among the distinguishable node types (i.e., those only a small number of node types will be connected with them) will represent a more effective correlation than those composed of indistinguishable ones (i.e., those all the node types can be connected with them). By taking the meta path differences into consideration, we can represent the weighted meta path-based node degree as

$$D(u) = \sum_{P_i \in \mathcal{P}} w_{P_i} \cdot D_{P_i}(u), \tag{3.58}$$

where vector $[w_{P_1}, w_{P_2}, \ldots, w_{P_{|\mathcal{P}|}}]^\top$ $\left(\sum_{P_i \in \mathcal{P}} w_{P_i} = 1\right)$ represents the weight parameters corresponding to the different meta paths.

3.5.4.2 Meta Path-Based Node Centrality and Closeness

Given the user node type and the set of meta paths $\mathcal{P}$ in $G^{(1)}$, based on each of the meta path $P_i \in \mathcal{P}$, the connections among users can be organized as homogeneous (weighted) graph $G_{P_i} = (\mathcal{U}, \mathcal{E}_{P_i}, w_{P_i})$, where the link set $\mathcal{E}_{P_i} = \{(u, v)|u, v \in \mathcal{U}, P_i(u, v) \neq \emptyset\}$. Mapping $w_{P_i} : \mathcal{E}_{P_i} \to \mathbb{R}$ denotes the weight of links in $\mathcal{E}_{P_i}$, where $w_{P_i}((u, v)) = |P_i(u, v)|$ represents the number of meta path instances of P_i connecting u and v. If we don't care about the link weights, the weight mapping is optional in the graph definition and can be discarded. In other words, based on the meta path concept, we can transform a *heterogeneous social network* into a group of *homogeneous networks* instead, where the edge weight equals to the meta path instance number.

Based on graph G_{P_i}, we can define the *centrality* measure of user $u \in \mathcal{U}$ as $C_{P_i}(u)$, which denotes either degree centrality, eigen-centrality, Katz centrality, pagerank centrality, or betweenness centrality that we have introduced before. Similar to the meta path-based degree concept introduced before, different meta path can play a different role in defining the users' centrality. One way to define the centrality measure of user u based on all the meta paths can be represented as

$$C(u) = \sum_{P_i \in \mathcal{P}} w_{P_i} \cdot C_{P_i}(u), \tag{3.59}$$

where w_{P_i} denotes the weight of the centrality measure based on meta path P_i.

In a similar way, the closeness measure among the user node pairs in the networks can be represented as

$$C(u, v) = \sum_{P_i \in \mathcal{P}} w_{P_i} \cdot C_{P_i}(u, v), \tag{3.60}$$

where $C_{P_i}(u, v)$ represents the closeness, e.g., *common neighbor* or *Jaccard's coefficient*, between users u and v in the network computed with meta path P_i.

3.6 Network Models

We have covered the basic knowledge about graphs, network measures, network category, and meta path already in this chapter. Before we end this chapter, we would like to introduce several models proposed for networks specifically. To model the link formation process in online social networks, several different models have been introduced, which can simulate how these networks are formed about the users. In this part, we will discuss several well-known network models, and analyze the properties, like degree distribution, clustering coefficient, and average path length, of networks generated by these models.

3.6.1 Random Graph Model

In the random graph model [14], the links among the nodes are assumed to be formed randomly, and each link will form with an equal chance. Based on such a simple assumption, the random graph model greatly simplifies the process of link formation in the real-world networks. Several different random graph models have been proposed already, and in this part, we will use the random graph model proposed by Gilbert [17] and Solomonoff and Rapoport [49] as an example.

In the random graph model, given a fixed number of nodes, e.g., n, the links among these nodes are formed independently with probability p. Formally, we denote the graph formed by following such a process as $G(n, p)$.

Theorem 3.3 *In graph $G(n, p)$, the number of links is not certain and the expected link number is $\frac{1}{2}\binom{n}{2}p$ (if the links are undirected).*

Proof We can represent the link number in the formed graph as m, and we have

$$\begin{aligned} m &= \frac{1}{2} \sum_{u,v \in \mathcal{V}, u \neq v} p((u,v)) \times 1 + (1 - p((u,v))) \cdot 0 \\ &= \frac{1}{2} \sum_{u,v \in \mathcal{V}, u \neq v} p((u,v)) \\ &= \frac{1}{2} \binom{n}{2} p. \end{aligned} \tag{3.61}$$

Meanwhile, given a graph $G(n, p)$, we can also infer the probability of forming m links in $G(n, p)$ according to the following theorem.

Theorem 3.4 *In graph $G(n, p)$, the probability of forming m links is $\binom{\frac{\binom{n}{2}}{2}}{2} p^m (1-p)^{\frac{\binom{n}{2}}{2} - m}$.*

Proof In graph $G(n, p)$, there exist $\frac{\binom{n}{2}}{2}$ potential links to be formed among these n nodes. Among these potential links, the probabilities that $0 \leq m \leq \frac{\binom{n}{2}}{2}$ of them are formed and the remaining are not formed can be denoted as p^m and $(1-p)^{\frac{\binom{n}{2}}{2} - m}$, respectively. Therefore, the final probability that m out of these $\frac{\binom{n}{2}}{2}$ potential links are formed can be represented as

$$P(m) = \binom{\frac{\binom{n}{2}}{2}}{2} p^m (1-p)^{\frac{\binom{n}{2}}{2} - m}. \tag{3.62}$$

Theorem 3.5 *In graph $G(n, p)$, the expected degree of nodes is $(n-1)p$.*

Proof For a node u in graph $G(n, p)$, it can be connected with the remaining $n - 1$ nodes all with probability p. Therefore, the expected degree of the node u in $G(n, p)$ can be represented as

$$\begin{aligned} \mathbb{E}(D(u)) &= \sum_{v \in \mathcal{V} \setminus \{u\}} p \cdot 1 + (1-p) \cdot 0 \\ &= (n-1)p. \end{aligned} \tag{3.63}$$

Theorem 3.6 *In graph $G(n, p)$, the probability that a node has a degree of d is $\binom{n-1}{d} p^d (1-p)^{n-1-d}$.*

Proof Among these $n - 1$ potential neighbors of a given node, e.g., u, the node has $\binom{n-1}{d}$ different choices to select d neighbors for u to get connected with. Meanwhile, the probability of merely forming links with these selected d neighbors is $p^d (1 - p)^{n-1-d}$. In other words, the probability for a node u to have degree d will be

$$P(D(u) = d) = \binom{n-1}{d} p^d (1 - p)^{n-1-d}. \tag{3.64}$$

Theorem 3.7 *The global clustering coefficient of a random graph $G(n, p)$ is p.*

Proof According to the definition of clustering coefficient, we have

$$C(G(n, p)) = \frac{|\mathcal{T}|}{|\mathcal{P}|}, \tag{3.65}$$

where set $\mathcal{T}$ denotes the node triples which form triangles and set $\mathcal{P}$ denotes the node triples forming a path of length 2.

In the random graph $G(n, p)$, given three nodes $u, v, w \in \mathcal{V}$, the probability that they will form a path $u - v - w$ (where u and w can be either connected or unconnected) is p^2. Meanwhile, the probability that these three nodes will form a triangle will be p^3. Therefore, we have the sizes of sets $\mathcal{T}$ and $\mathcal{P}$ will be $\sum_{u,v,w \in \mathcal{V}, u \neq v \neq w} p^3$ and $\sum_{u,v,w \in \mathcal{V}, u \neq v \neq w} p^2$, respectively, and the global clustering coefficient is

$$C(G(n, p)) = \frac{\sum_{u,v,w \in \mathcal{V}, u \neq v \neq w} p^3}{\sum_{u,v,w \in \mathcal{V}, u \neq v \neq w} p^2} = p. \tag{3.66}$$

Theorem 3.8 *In graph $G(n, p)$, given two nodes $u, v \in \mathcal{V}$, the probability that there exists a path of length k connecting u and v is $\binom{n-2}{k-1} p^k$.*

Proof Between u and v, if there exists a path of length k connecting them, we can denote such a path as $P = u \rightarrow u_1 \rightarrow \cdots, u_{k-1} \rightarrow v$, where the intermediate nodes $u_1, u_2, \ldots, u_{k-1} \in \mathcal{V} \backslash \{u, v\}$. There exist $\binom{n-2}{k-1}$ different choices of these $k - 1$ nodes. Meanwhile, among these $k - 1$ selected nodes, the probability that they will form a path connecting u and v will be p^k. Therefore, we have the probability that there exists a path of length k connecting nodes u and v can be represented as

$$P(u, v, k) = \binom{n-2}{k-1} p^k. \tag{3.67}$$

Given the node number n, some properties of the random graph $G(n, p)$ will change as parameter p increases from 0 to 1. In the case that $p = 0$, the random graph $G(n, 0)$ will only involve n isolated nodes without any connections. In such a graph, the graph diameter will be 0 and the size of the largest connected component will contain merely 1 node and the average path length is 0 as no path exists among these nodes. As p increases, some links will be formed among the nodes, and the diameter of the graph $G(n, p)$ will increase which can also be greater than 1. At the same time, the size of the largest component increases, while the average path length will also increase and can be greater than 1. Meanwhile, in the case that $p = 1$, the graph will be a complete graph involving n nodes and $\frac{n(n-1)}{2}$ links with diameter 1, where all the nodes will be incorporated into one single connected component.

All the nodes will be connected, and the average path length in $G(n, 1)$ will be 1. Formally, the point where the diameter increases first and starts to shrink is called the *phase transition* [9] point.

Theorem 3.9 *The phase transition happens at* $p = \frac{1}{n-1}$ *in the random graph model.*

Proof The proof of the above theorem is left as an exercise for the readers.

In the random graph model, the formation of all the links is assumed to be independent with identical probabilities. However, in the real-world social networks, such an assumption cannot hold. For instance, in the socialization among users, people tend to form a small community involving connections with a very limited number of people, like friends, family members, and colleagues. Many other models, like the *small-world model* [28, 39, 58] can be used to model the formation process of such a phenomenon better.

3.6.2 Preferential Attachment Model

When making friends, generally the people with a large neighborhood can attract the connections more easily. For instance, in the real-world online social networks, the celebrities, like the politicians and super stars, are well known and they are usually among the top candidates that we choose to follow. A well-established method to model such an observation in network formation is called the *preferential attachment* model [4].

In the *preferential attachment* model, at the very beginning, there exist n_0 node in the network and new nodes will be added to form connections with these existing nodes. The new node will connected $n \leq n_0$ other existing nodes. Formally, we can represent the degrees of nodes in the existing graph, e.g., u, as $D(u)$, and new nodes are more likely to establish connections with the active nodes, i.e., those with a large degree. The probability for a new node to get connected with u can be represented as $P(u) = \frac{D(u)}{\sum_{v \in \mathcal{V}} D(v)}$.

Theorem 3.10 *The degree distribution of the graph generated by the preferential attachment model follows the power-law distribution with an exponent* $b = 3$.

Proof According to the introduction, the probability for the newly added node to connected with an existing node u is

$$P(u) = \frac{D(u)}{\sum_{v \in \mathcal{V}} D(v)}. \tag{3.68}$$

Meanwhile, at each step t, the expected increase of u's degree is proportional to $D(u)$, which can be modeled with a mean-field setting,

$$\begin{aligned} \frac{\mathrm{d}D(u)}{\mathrm{d}t} &= nP(u) \\ &= \frac{nD(u)}{\sum_{v \in \mathcal{V}} D(v)} \\ &= \frac{nD(u)}{2nt} \\ &= \frac{D(u)}{2t}. \end{aligned} \tag{3.69}$$

In each step, n links will be added, and after t steps, the total node degree will be equal to $\sum_{v \in \mathcal{V}} D(v) = 2nt$. By solving such a partial differential equation, we can get

$$D(u) = n \left(\frac{t}{t_u} \right)^{\frac{1}{2}}, \tag{3.70}$$

where t_u denotes the step that u is added into the network.

The probability that $D(u)$ is less than d can be represented as

$$P(D(u) < d) = P\left(t_u > \frac{n^2 t}{d^2}\right) = 1 - P\left(t_u \leq \frac{n^2 t}{d^2}\right). \tag{3.71}$$

If we assume that $t_u \sim \text{Uniform}(0, t)$, we have

$$P(D(u) < d) = 1 - P\left(t_u \leq \frac{n^2 t}{d^2}\right) = 1 - \frac{n^2}{d^2} \frac{1}{n_0 + t}. \tag{3.72}$$

Let the node degree distribution density function to be $P(D)$, we have

$$P(d) = \frac{\partial P(D(u) < d)}{\partial d} = \frac{2n^2 t}{d^3 (t + n_0)} \approx_{t \to \infty} \frac{2n^2}{d^3}. \tag{3.73}$$

Theorem 3.11 *Based on the preferential attachment model, by using the mean-field analysis, the expected clustering coefficient of the generated network is*

$$C = \frac{n_0 - 1}{8} \frac{(\ln t)^2}{t}. \tag{3.74}$$

Theorem 3.12 *Based on the preferential attachment model, the average path length of nodes in the generated network is*

$$l \approx \frac{\ln |\mathcal{V}|}{\ln(\ln(|\mathcal{V}|))}. \tag{3.75}$$

The proofs of the above two theorems are out of the scope of this book, which will not be introduced here. For the readers who are interested in the proof, please refer to [65].

3.7 Summary

In this chapter, we provided an overview about the essential knowledge of online social networks, which can generally be represented as graphs involving nodes and connections among the nodes. Some basic information about graphs were provided at the beginning of this chapter, covering the different graph representation methods, e.g., adjacency matrix and adjacency list, and graph connectivity concepts, e.g., adjacent neighbors, incident links, walk, trail, tour, path, cycle, as well as node reachability and connect component.

We introduced the various measures for networks in this chapter, including degree, centrality, closeness, transitivity, and social balance. We talked about the node degree concept as well as the node degree distribution, which provide the basic information about the network connectivity structures. To denote the importance of node roles in the network, several different node centrality measures were introduced. The closeness between the node pairs in the networks can be computed with various closeness measures based on the local network structures, global paths, and random walks. We introduced the concepts of social transitivity, clustering coefficient, and social balance to analyze various social connection-based network properties.

Depending on the network structures and the involved information, the networks could be divided into various categories, e.g., homogeneous network, heterogeneous network, and aligned heterogeneous networks. The representative examples of homogeneous networks include the friendship network, computer network, and company organizational chart; the examples of heterogeneous networks cover the online social network, bibliographic network, and movie knowledge library, while the aligned heterogeneous networks concept provides the opportunity to model the information across multi-platforms.

To depict the diverse information inside the networks, meta path can be a very useful methods, which can outline the potential connections among the nodes. In the provided definition of the meta path concept based on the network schema, meta paths can be represented as the sequences of node types connected by the link types. Besides the meta paths within one single heterogeneous network, we also introduced the meta path across heterogeneous networks via the anchor meta path. Various network measures, e.g., degree, centrality, and closeness, were defined based on the meta path concept as well.

We concluded this chapter with several network models, including the random graph model and the preferential attachment model. A brief introduction and analysis about these two models were provided, which can also be applied to study various social network learning problems to be introduced in the following chapters as well.

3.8 Bibliography Notes

Studying online social networks and other related network structured data have been one of the most important research topics in the academia of machine learning and data mining in recent years, since lots of real-world data can be modeled as the networks [40]. There exist some survey articles on social networks [21], heterogeneous information networks [48, 51], and aligned social networks [66] published in recent years already, which can serve as the road map to study these related areas for the readers.

If the readers are interested in learning more knowledge about graph theory, you are very recommended to read the textbook "*Graph Theory and Complex Networks: An Introduction*" [55], which is well-written and well-organized book and covers a very broad topic about graphs. The recent "*Social Media Mining: An Introduction*" textbook [65] also provides a brief introduction to the graph related essential background knowledge, and the readers can take a look at that book as well.

Node degree distribution usually follows the power-law distribution [13], where the majority of the nodes only have a very small degree, while a very small number of the nodes can have a very large degree instead. Node centrality metric can measure the importance of nodes based on their positions inside the network, and a systematic overview of existing centrality measures is available in [11]. As to the node closeness, the readers can take a look at the recent survey article [67], which introduces various closeness measures as potential link predictors. The network transitivity, clustering coefficient, and social balance concepts are covered in [19, 20, 57], respectively.

A comprehensive survey about the network categories and existing network mining problems has been provided in [66], which also covers one section on network fusion and learning specifically. For the heterogeneous information network research works, the readers are suggested to read the lecture synthesis book [51], which covers the ranking, search, classification, and clustering problems on heterogeneous information networks. About the aligned heterogeneous social network alignment and mining problems, the readers are suggested to read the latest survey paper [48], which covers the alignment, link prediction, clustering, information diffusion, and embedding problems.

The meta path concept was initially introduced by Sun et al. in [53], and lately extended by Zhang et al. to the cross-network scenario in [73], which serves as an important tool for handling the heterogeneous network structures. Based on the assumptions that networks are generated randomly, the random graph models [14,17,49] can depict the generation process of graphs and certain properties that these generated graphs can have. Meanwhile, the preferential attachment model can depict the addition of new nodes into graphs, whose detailed description is available in [4].

3.9 Exercises

1. (Easy) Please compute the *diameter* of the graph shown in Fig. 3.23, and provide the maximum *shortest path*.
2. (Easy) Please compute the *betweenness centrality* and the *normalized betweenness centrality* of all the nodes in the input graph shown in Fig. 3.23.
3. (Easy) Please draw the *degree distribution* plot for the graph shown in Fig. 3.23.
4. (Easy) Please compute the *closeness* scores for all potential node pairs in Fig. 3.23 based on *common neighbor*, *Jaccard's coefficient*, and *Adamic/Adar*, respectively.
5. (Medium) Besides the heterogeneous network examples provided in this chapter, please think about some other data in the real world, which can be represented as a *heterogeneous network*. Please also provide its *network schema*, and list some *meta path* examples based on the schema.
6. (Medium) Based on the network schema, we can define a large number of meta paths. However, in many applications, extremely long meta paths (e.g., longer than 10) are not very useful. Please think why and write down the potential reasons.
7. (Medium) In Sect. 3.3.4.2, we show that the *clustering coefficient* equals to

$$CC = \frac{\text{Number of triangles} \times 6}{|\mathcal{P}^2|}. \tag{3.76}$$

Please also prove that the following equation also holds for computing the network *clustering coefficient*:

$$CC = \frac{\text{Number of triangles} \times 6}{\text{Number of connected triples of nodes}}. \tag{3.77}$$

8. (Hard) Please try to prove Theorem 3.9 regarding the *phase transition point* in the *random graph model*.

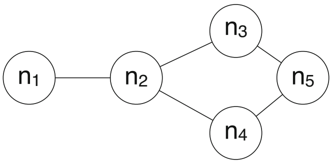


Fig. 3.23 An input graph example

9. (Hard) Please prove that if a Markov chain is *irreducible* and *aperiodic* then the largest eigenvalue of the transition matrix **P** will be equal to 1 and all the other eigenvalues will be strictly less than 1, as introduced in Sect. 3.3.3.3.
10. (Hard) Please try to prove Theorems 3.11 and 3.12 about the *preferential attachment model*.

References

1. L. Adamic, E. Adar, Friends and neighbors on the Web. Soc. Netw. **25**(3), 211–230 (2003)
2. L. Adamic, R. Lukose, A. Puniyani, B. Huberman, Search in power-law networks. Phys. Rev. E **64**, 046135 (2001)
3. M. Bagella, L. Becchetti, The determinants of motion picture box office performance: evidence from movies produced in Italy. J. Cult. Econ. **23**(4), 237–256 (1999)
4. A. Barabasi, R. Albert, Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
5. A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks. Proc. Natl. Acad. Sci. **101**(11), 3747–3752 (2004)
6. H. Bast, D. Delling, A. Goldberg, M. Müller-Hannemann, T. Pajor, P. Sanders, D. Wagner, R. Werneck, Route planning in transportation networks (2015). arXiv:1504.05140
7. M. Berry, S. Dumais, G. O'Brien, Using linear algebra for intelligent information retrieval. SIAM Rev. **37**(4), 573–595 (1995)
8. C. Bettstetter, On the minimum node degree and connectivity of a wireless multihop network, in *Proceedings of the 3rd ACM International Symposium on Mobile Ad Hoc Networking and Computing* (ACM, New York, 2002)
9. B. Bollobás, S. Janson, O. Riordan, The phase transition in inhomogeneous random graphs. Random Struct. Algoritm. **31**(1), 3–122 (2007)
10. O. Bonaventure, *Computer Networking: Principles, Protocols, and Practice* (The Saylor Foundation, Washington, 2011)
11. S. Borgatti, M. Everett, A graph-theoretic perspective on centrality. Soc. Netw. **28**(4), 466–484 (2006)
12. U. Brandes, T. Erlebach, *Network Analysis: Methodological Foundations*. Lecture Notes in Computer Science (Springer, Berlin, 2005)
13. A. Clauset, C. Shalizi, M. Newman, Power-law distributions in empirical data. SIAM Rev. **51**(4), 661–703 (2009)
14. P. Erdos, A. Renyi, On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. **5**(1), 17–60 (1960)
15. F. Fouss, A. Pirotte, J. Renders, M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Trans. Knowl. Data Eng. **19**, 355–369 (2007)
16. L. Freeman, A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977)
17. E. Gilbert, Random graphs. Ann. Math. Stat. **30**(4), 1141–1144 (1959)
18. T. Gruber, Toward principles for the design of ontologies used for knowledge sharing. Int. J. Hum. Comput. Stud. **43**(5–6), 907–928 (1995)
19. F. Harary, H. Kommel, Matrix measures for transitivity and balance. J. Math. Sociol. **6**(2), 199–210 (1979)
20. J. Harmon, The psychology of interpersonal relations. Soc. Forces **37**(3), 272–273 (1959)
21. J. Heidemann, M. Klier, F. Probst, Online social networks: a survey of a global phenomenon. Comput. Netw. **56**(18), 3866–3878 (2012)
22. D. Horton, R. Wohl, Mass communication and para-social interaction. Psychiatry **19**(3), 215–229 (1956)
23. A. Huang, Similarity measures for text document clustering, in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)* (Christchurch, 2008), pp. 49–56
24. P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull. Soc. Vaud. Sci. Nat. **37**(142), 547–579 (1901)
25. L. Katz, A new status index derived from sociometric analysis. Psychometrika **18**, 39–43 (1953)
26. D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2003), pp. 137–146
27. J. Kleinberg, Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
28. J. Kleinberg, The small-world phenomenon: an algorithmic perspective, in *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing* (ACM, New York, 2000), pp. 163–170
29. X. Kong, J. Zhang, P. Yu, Inferring anchor links across multiple heterogeneous social networks, in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (ACM, New York, 2013), pp. 179–188
30. T. Lappas, K. Liu, E. Terzi, Finding a team of experts in social networks, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2009), pp. 467–476
31. E. Leicht, P. Holme, M. Newman, Vertex similarity in networks. Phys. Rev. E **73**, 026120 (2006)

32. J. Leskovec, D. Huttenlocher, J. Kleinberg, Signed networks in social media, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM, New York, 2010), pp. 1361–1370
33. D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
34. Y. Liu, J. Zhang, C. Zhang, P. Yu, Data-driven blockbuster planning on online movie knowledge library, in *2018 IEEE International Conference on Big Data* (IEEE, Piscataway, 2018)
35. F. Lorrain, H. White, Structural equivalence of individuals in social networks. J. Math. Sociol. **1**, 49–80 (1971)
36. L. Lovász, Random walks on graphs: a survey, in *Combinatorics, Paul Erdős is Eighty* (1996)
37. Q. Mei, D. Zhou, K. Church, Query suggestion using hitting time, in *Proceedings of the 17th ACM conference on Information and Knowledge Management* (ACM, New York, 2008)
38. A. Mislove, M. Marcon, K. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and analysis of online social networks, in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (ACM, New York, 2007), pp. 29–42
39. M. Newman, Models of the small world. J. Stat. Phys. **101**(3–4), 819–841 (2000)
40. M. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010)
41. J. Pan, H. Yang, C. Faloutsos, P. Duygulu, Automatic multimedia cross-modal correlation discovery, in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2004), pp. 653–658
42. S. Pillai, T. Suel, S. Cha, The Perron-Frobenius theorem: some of its applications. IEEE Signal Process. Mag. **22**(2), 62–75 (2005)
43. A. Pothen, H. Simon, K. Liou, Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. **11**(3), 430–452 (1990)
44. E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, A. Barabási, Hierarchical organization of modularity in metabolic networks. Science **297**(5586), 1551–1555 (2002)
45. M. Riedl, M. Young, Narrative planning: balancing plot and character. J. Artif. Int. Res. **39**(1), 217–267 (2010)
46. B. Ruhnau, Eigenvector-centrality a node-centrality? Soc. Netw. **22**(4), 357–365 (2000)
47. P. Savalle, E. Richard, N. Vayatis, Estimation of simultaneously sparse and low rank matrices, in *Proceedings of the 29th International Conference on Machine Learning* (2012)
48. C. Shi, Y. Li, J. Zhang, Y. Sun, P. S. Yu, A survey of heterogeneous information network analysis. IEEE Trans. Knowl. Data Eng. **29**, 17–37 (2017)
49. R. Solomonoff, A. Rapoport, Connectivity of random nets. Bull. Math. Biol. **13**, 107–117 (1951)
50. T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. **5**, 1–34 (1948)
51. Y. Sun, J. Han, *Mining Heterogeneous Information Networks: Principles and Methodologies* (Morgan & Claypool Publishers, 2012)
52. Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, J. Han, Co-author relationship prediction in heterogeneous bibliographic networks, in *2011 International Conference on Advances in Social Networks Analysis and Mining* (IEEE, Piscataway, 2011)
53. Y. Sun, J. Han, X. Yan, P. Yu, T. Wu, Pathsim: meta path-based top-k similarity search in heterogeneous information networks. Proc. VLDB Endowment **4**(11), 992–1003 (2011)
54. J. Tang, T. Lou, J. Kleinberg, Inferring social ties across heterogeneous networks, in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (ACM, New York, 2012), pp. 743–752
55. M. van Steen, *Graph Theory and Complex Networks: An Introduction* (Maarten van Steen, Lexington, 2010)
56. Z. Wang, J. Liao, Q. Cao, H. Qi, Z. Wang, Friendbook: a semantic-based friend recommendation system for social networks. IEEE Trans. Mob. Comput. **14**(3), 538–551 (2015)
57. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994)
58. D. Watts, S. Strogatz, Collective dynamics of small-world networks. Nature **393**, 440–442 (1998)
59. K. Wilcox, A.T. Stephen, Are close friends the enemy? Online social networks, self-esteem, and self-control. J. Consum. Res. **40**(1), 90–103 (2012)
60. R. Wilson, *Introduction to Graph Theory* (Wiley, London, 1986)
61. X. Xie, Potential friend recommendation in online social network, in *2010 IEEE/ACM International Conference on Green Computing and Communications & International Conference on Cyber, Physical and Social Computing* (IEEE, Piscataway, 2010). https://ieeexplore.ieee.org/abstract/document/5724926
62. J. Yang, J. McAuley, J. Leskovec, Community detection in networks with node attributes, in *2013 IEEE 13th International Conference on Data Mining (ICDM)* (IEEE, Piscataway, 2013)
63. Y. Yao, H. Tong, X. Yan, F. Xu, J. Lu, Matri: a multi-aspect and transitive trust inference model, in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, New York, 2013), pp. 1467–1476

64. J. Ye, H. Cheng, Z. Zhu, M. Chen, Predicting positive and negative links in signed social networks by transfer learning, in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, New York, 2013), pp. 1477–1488
65. R. Zafarani, M. Abbasi, H. Liu, *Social Media Mining: An Introduction* (Cambridge University Press, New York, 2014)
66. J. Zhang, Social network fusion and mining: a survey (2018). arXiv preprint. arXiv:1804.09874
67. J. Zhang, P. Yu, *Link Prediction Across Heterogeneous Social Networks: A Survey* (University of Illinois, Chicago, 2014)
68. J. Zhang, P. Yu, Community detection for emerging networks, in *Proceedings of the 2015 SIAM International Conference on Data Mining* (Society for Industrial and Applied Mathematics, Philadelphia, 2015), pp. 127–135
69. J. Zhang, P. Yu, Multiple anonymized social networks alignment, in *2015 IEEE International Conference on Data Mining* (IEEE, Piscataway, 2015)
70. J. Zhang, P. Yu, PCT: partial co-alignment of social networks, in *Proceedings of the 25th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee, Geneva, 2016), pp. 749–759
71. J. Zhang, X. Kong, P. Yu, Predicting social links for new users across aligned heterogeneous social networks (2013). arXiv preprint. arXiv:1310.3492
72. J. Zhang, X. Kong, P. Yu, Transferring heterogeneous links across location-based social networks, in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (ACM, New York, 2014), pp. 303–312
73. J. Zhang, P. Yu, Z. Zhou, Meta-path based multi-network collective link prediction, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2014), pp. 1286–1295
74. J. Zhang, P. Yu, Y. Lv, Organizational chart inference, in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2015), pp. 1435–1444
75. J. Zhang, S. Wang, Q. Zhan, P. Yu, Intertwined viral marketing in social networks, in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE, Piscataway, 2016)
76. J. Zhang, P. Yu, Y. Lv, Q. Zhan, Information diffusion at workplace, in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (ACM, New York, 2016), pp. 1673–1682
77. J. Zhang, Q. Zhan, L. He, C. Aggarwal, P. Yu, Trust hole identification in signed networks, in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, Berlin, 2016), pp. 697–713
78. J. Zhang, P. Yu, Y. Lv, Enterprise employee training via project team formation, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (ACM, New York, 2017), pp. 3–12
79. J. Zhang, C. Aggarwal, P. Yu, Rumor initiator detection in infected signed networks, in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)* (IEEE, Piscataway, 2017)
80. T. Zhou, L. Lü, Y. Zhang, Predicting missing links via local information. Eur. Phys. J. B **71**(4), 623–630 (2009)