

CHAPTER 1

Preliminaries and Basic Definitions in Network Theory

Guido Caldarelli¹ and Alessandro Vespignani^{2,3}

¹*INFN-CNR Centro SMC Dipartimento di Fisica Università di Roma
"La Sapienza" P.le A. Moro 5 00185 Roma, Italy*

²*School of Informatics & Biocomplexity Center,
Indiana University, IN 47406, USA*

³*Laboratoire de Physique Théorique (UMR du CNRS 8627), Batiment 210,
Université de Paris-Sud 91405 Orsay, France*

1.1. Introduction

In very general terms a network can be described as a graph whose nodes (vertices) identify the elements of the system. The set of connecting links (edges) represents the presence of a relation or interaction among these elements. With such a high level of generality it is easy to perceive that a wide array of systems can be approached within the framework of network theory. In this section we provide a basic notation and the definitions needed to describe networks. Not surprisingly, each field concerned with network science introduced its own basic notation and nomenclature. The natural framework for a rigorous mathematical description of networks, however, is found in graph theory and we will stick to it in this Chapter. Note that graph theory consists in an impressive body of work and we are not in the condition to provide a formal and complete presentation of it. Our purpose in this introductory chapter is to provide some notions useful to describe networks and commonly used in the rest of the book. For the interested reader, amongst the various introductory books on graph theory, we suggest to consult those by West,¹ Bollobás,² Diestel,³ and Caldarelli.⁴

1.2. Basic definitions

As in any mathematical abstraction, when we describe a systems as a **graph** we decide to discard many of the specific peculiarities of the real phenom-

ena and focus only on a few features of interest. In particular, a graph is essentially a way to code a relation (physical links, interactions etc.) between the elements of a system. The elements of the system identify the set V (set of **vertices**), and the relations among those the set E (set of **edges**). The graph indicated as $G(V, E)$ can be drawn plotting the vertices as points and the edges as lines between them. It is not important how they are actually drawn. Ultimately the only thing that matters is to know which vertices are connected.

- A graph $G(V, E)$ where V has n elements (n vertices) is said to have **order** n . Analogously, the **size** of a graph is the number m of its edges (the number of elements of the set E).
- When an edge e links vertices v_1, v_2 we have that vertices v_1, v_2 are **incident** with the edge e . Alternatively the edge e *joins* v_1, v_2 that are its *endvertices*.
- Vertices v_1, v_2 joined by edge e are **adjacents** or neighbours.
- A *dominating set* for a graph is a set of vertices whose neighbors, along with themselves, constitute all the vertices in the graph.
- A graph with *order* n cannot have more than m_{max} edges where $m_{max} = n(n-1)/2$ (the *size* is smaller than m_{max}). When all these possible edges are present the graph is **complete** and it is indicated with the symbol K^n . The opposite case happens when there are no edges at all. The graph is then *empty* and it is indicated by the symbol E^n .

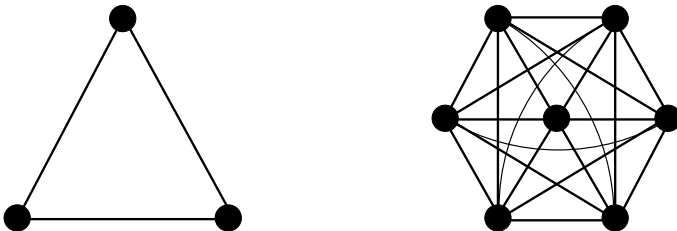


Fig. 1. Two different examples of a complete graph. On the left K^3 and on the right K^7 .

1.3. Different kinds of graphs

1.3.1. Weighted, directed and oriented graphs

- Whenever a real number can be attached to an existing edge we have that the edge is characterized by a weight w . Note that in this book the weights are (almost exclusively) positive real numbers (i.e. $w > 0$). The graph in this case is a **weighted** graph.
- A **directed** graph $G(V, E)$ is given by two disjoint sets E and V plus two functions $I(E \rightarrow V)$ and $F(E \rightarrow V)$. The first one assigns to every edge e an initial vertex $I(e)$. The second one assigns to every edge e a final vertex $F(e)$. More simply, every edge e has assigned a direction from one vertex $I(e)$ to another $F(e)$.
- Sometime $I(e)$ and $F(e)$ coincide. In this case e is a **loop**. Moreover, we can have different edges directed between the same two vertices $I(e)$ and $F(e)$. This is the case of **multiple edges**.

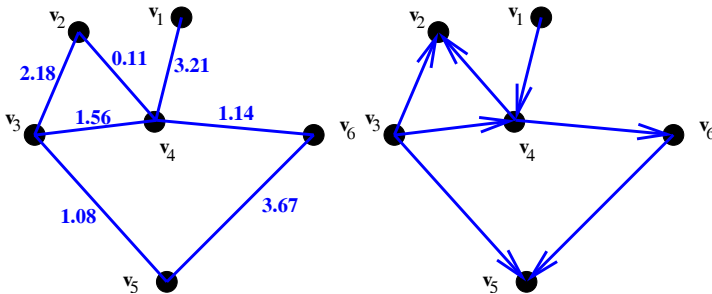


Fig. 2. On the left a realization of a weighted graph. The degree of vertex v_4 is 6.02 (given by $0.11 + 3.21 + 1.14 + 1.56$). On the right an example of an oriented graph. If the weight were the same in this case we would have an in-degree $k^{w,in} = 4.77(3.21 + 1.56)$ and an out-degree $k^{w,out} = 1.25(0.11 + 1.14)$

- Whenever the direction is assigned but neither loops nor multiple edges are present, then the graph is **oriented**. Intuitively oriented graphs are undirected graphs where for every edge one assigns a direction.
- A **multigraph** is a pair of disjoint sets (V, E) together with a map $E \rightarrow V \cup [V]^2$ assigning to every edge either one or two vertices (the ends). A multigraph is then similar to a directed graph, with multiple edges and loops but no direction assigned. A sketch of the various kind of graph is presented in Fig. 3.

- The number of edges of vertex v_i in a graph is called **degree** of vertex v_i and it is indicated here by $k(v_i)$. In the case of an oriented graph the degree can be distinguished in *in-degree* $k^{in}(v_i)$ and the *out-degree* $k^{out}(v_i)$. In the case of weighted graphs we will consider the *weighted-degree* $k^w(v_i)$ of a vertex v_i as the sum of the weight of the edges on v_i .
- If the set V in graph $G(V, E)$ is composed by vertices v_1, v_2, \dots, v_n then the series $k(v_1), k(v_2), \dots, k(v_n)$ is a *degree sequence* of $G(V, E)$. Particular importance in this book is devoted to the statistical properties of such degree sequence.

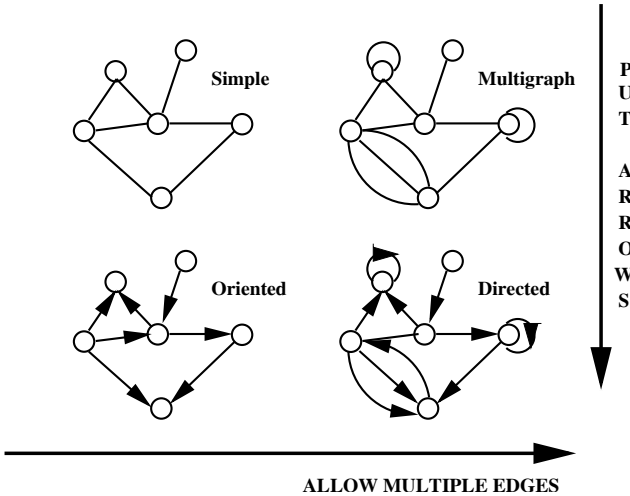


Fig. 3. The distinction between the various types of graphs.

1.3.2. Subgraphs

- Consider two graphs $G(V, E)$ and $G'(V', E')$. We can define a new graph indicated by $G \cap G'$ whose vertices are in the set $V \cap V'$ and the edges in the set $E \cap E'$. If $V \cap V' = \emptyset$ the two graphs are **disjoint**. On the other hand if $V' \subseteq V$ and $E' \subseteq E$ then $G'(V', E')$ is an *induced subgraph* of $G(V, E)$ and we indicate this by writing $G'(V', E') \subseteq G(V, E)$. Finally, if $G'(V', E') \subseteq G(V, E)$ and $V' = V$, $G'(V', E')$ is **spanning** of $G(V, E)$

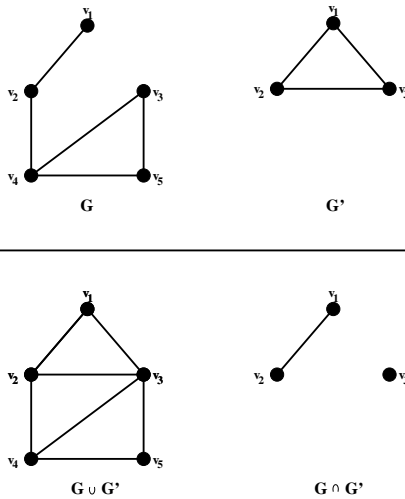


Fig. 4. The operations of union and intersection of two graphs

- Two graphs are *isomorphic* if you can re-draw one of them so that it looks exactly like the other. An open problem is to determine in a short time whether two graphs are isomorphic or not.

1.3.3. Partited graphs

- Let $r \geq 2$ be an integer, a graph $G(V, E)$ is called **r-partite** if it can be divided in r classes such that every edge has its ends in different classes. This means that vertices in the same class cannot be adjacent. If $r = 2$ the graph is also called **bipartite**
- A complete *bipartite clique* $K_{i,j}$ is a graph where every one of i nodes has an edge directed to each of the j nodes.
- A *bipartite core* $C_{i,j}$ is a graph on $i + j$ nodes that contains at least one $K_{i,j}$ as a subgraph.

1.4. Paths and cycles

- A **path** is a (not empty) graph $G'(V', E')$ of the form $V' = v_0, v_1, \dots, v_n$, $E' = e_1, \dots, e_n$ where v_0, v_1, \dots, v_n a set of vertices for which e_i is an edge joining vertices v_{i-1} and v_i . Less formally we can say that a series of consecutive edges forms a **path**. The number of edges in a path is called the *length* of the path.

- if $P = e_1 + e_2 + \dots + e_n$ is a path then if $n \geq 3$ and we add an edge e_0 joining vertices v_n and v_0 , we obtain a **circuit**. Put in other words a **circuit** is a path whose endvertices coincide. If in the circuit all the vertices are distinct each other the circuit is a **cycle**. A cycle of length k is indicated as C^k . Note that **a cycle is different from a loop**
 - A *Hamiltonian path* is a path passing once through all the vertices (not necessarily through all the edges) in the graph. A Hamiltonian circuit is a Hamiltonian path which begin and ends in the same vertex. By construction this circuit is also a cycle.
 - An *Eulerian path* is a path that passes once through all the edges (not necessarily once through all the vertices) in the graph. An Eulerian circuit is an Eulerian path which begins and ends in the same edge. If the vertices in the circuit are all different then the circuit is a cycle.
- When a path exists between any couple of vertices v_i, v_j in a graph, the graph is **connected**. This property is called **connectivity**.

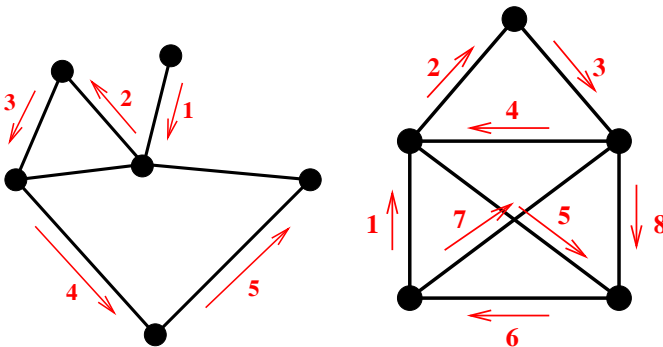


Fig. 5. Left an Hamiltonian path and right an Eulerian circuit

1.4.1. Trees

- A **tree** is a connected graph that also does not contain cycles (also acyclic graph). If the graph is not connected but still acyclic then it is composed by different trees and assume the natural name of **forest**.

- Vertices of degree 1 in a tree are called **leaves**. In any non trivial tree there are at least two leaves.
- It is convenient, in some cases, to consider one vertex of the tree as a special one. This vertex is called **root**. A tree with a fixed root is a *rooted tree*.

1.5. Statistics on graphs

One of the elements that has fostered the recent development of network science can be found in the recent possibility for the systematic gathering and handling of data sets on several large scale networks. The large number of elements comprised in these networks prompts us to the use of a statistical analysis as the proper tool for a useful mathematical characterization. Indeed, in large systems, asymptotic regularities cannot be found by looking at local elements or properties. This consideration has led many researchers, particularly in physics and computer science, to use a large scale statistical characterization. This allows to take into account the aggregate properties of the many interacting units that compose large scale networks. In the recent literature on large scale networks the statistical analysis has been initially focused on three main features, namely the small-world, the clustering and the degree distribution properties.

1.5.1. Small world properties

The small-world property refers to the the fact that in many large scale networks the average distance between vertices is very small compared to the size of the graphs. The distance between two vertices in a graph is measured as the shortest path length $\langle \ell \rangle$ among them. A global statistical measure of the distance among vertices can then be expressed as the average shortest path length among all possible couples of vertices in the network. The small-world concept describes in simple words a simple fact. It is possible to go from one vertex to any other in the system passing through a very small number of intermediate vertices. To be more precise, the small-world property is present when $\langle \ell \rangle$ scales logarithmically (or slower) with the number of vertices.

The small-world effect has been popularized in the sociological context where it is sometimes referred as “six degrees of separation”.⁵ A short number of acquaintances (on the average six) is enough to create a connection between any two people chosen at random. Since then, the small-world effect has been observed in many natural networks⁶ and appears to characterize

several infrastructure networks. As we see in the next chapters, the small-world property can be simply explained by the presence of randomness in the evolution of networks. It finds an elegant mathematical treatment in the celebrated graph model of Erdős and Rényi.

1.5.2. Clustering coefficient

The small-world property alone is not the signature of a special organizing principle. More interesting is the fact that, in close analogy to many social and technological networks,⁶ the small-world effect goes along with a high level of clustering. The concept of clustering of a graph, also called *transitivity* in the context of sociology,⁷ refers to the tendency observed in many natural networks to form cliques in the neighborhood of any given vertex. In this sense, a high clustering implies that, if the vertex i is connected to j , and j is connected to l , then very likely i is also connected to l (the friends of my friends are also friends of mine). The clustering of an undirected graph can be quantitatively measured by means of the *clustering coefficient*. Let us consider the vertex i , with degree k_i , and let us denote by e_i the number of edges existing between the k_i neighbors of i . The clustering coefficient, c_i , of i is defined as the ratio between the actual number of edges among its neighbors, e_i , and its maximum possible value, $k_i(k_i - 1)/2$, i.e.

$$c_i = \frac{2e_i}{k_i(k_i - 1)}. \quad (1)$$

Thus, the clustering coefficient c_i measures the average probability that two neighbors of the vertex i are also connected between them. Note that this measure of clustering has only meaning for $k_i > 1$. For $k_i \leq 1$ we define $c_i \equiv 0$. The finding of many clustered networks with small-world properties raises a very interesting issue: Random graphs feature the small-world effect but are not clustered, while regular grids tend to be clustered but are not small-world. We therefore need to identify the different organizing principles (on their turn related to both hierarchical and geographical factors) that allow the development of both properties at the same time.

1.5.3. Degree distribution

The most basic statistical characterization of a graph is given by the sequence of degrees k_i of its vertices i or, (on average) the relative probability distribution of degrees $P(k)$. This degree distribution $P(k)$ for an undirected graph is defined as the probability that any randomly chosen

vertex has degree k . In the case of directed graphs, one has to consider two different distributions, the in-degree $P(k_{in})$ and out-degree $P(k_{out})$ distributions, defined as the probability that a randomly chosen vertex has in-degree k_{in} and out-degree k_{out} , respectively. The functional forms generally considered to describe the degree distribution of real networks define two broad network classes. The first one refers to the so-called homogeneous networks. In this case the degree distribution have functional form with light tail such as Poisson's or Gaussian distributions. The second class concerns networks with heterogeneous connectivity pattern usually corresponding to heavy tailed degree distribution. A typical example is the case of scale-free networks with power-law degree distribution^b that behaves as $P(k) = Ak^{-\gamma}$. The origin of the discrimination of homogeneous

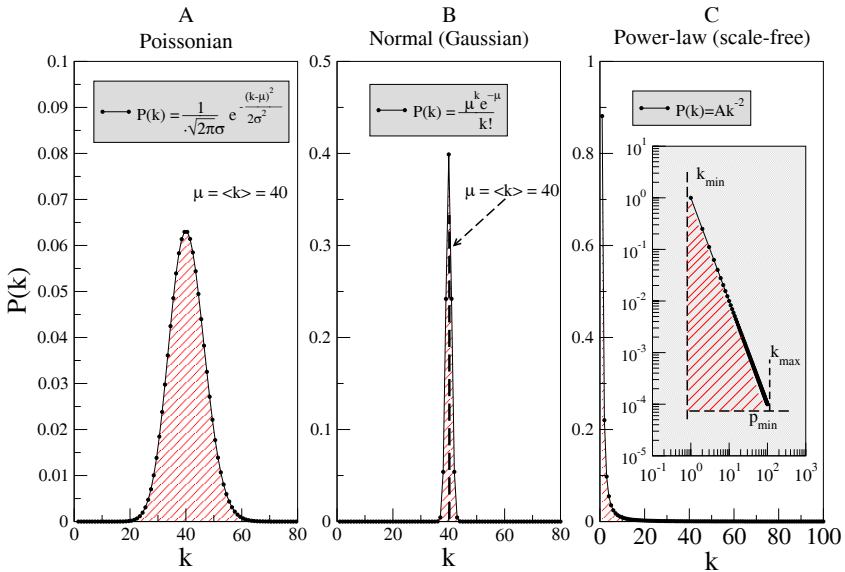


Fig. 6. (A) The plot of a Gaussian Distribution. (B) The plot of a Poisson Distribution. (C) The plot of a Power-law distribution. In the inset of (C) the same plot on a logarithmic scale.

and heterogeneous networks can be understood by looking at the first two moments of the degree distribution. For example we can compute what is the typical value that the degree assumes in the graph. This value will be

^bPower-law distributions are in many cases referred to as Pareto distributions.

indicated by $\langle k \rangle$, where the symbol $\langle \dots \rangle$ indicates an average over all the possible outcomes. A measure of the typical error we make if we assume that every vertex has degree $\langle k \rangle$ (thereby neglecting values fluctuations in our system) is given by the standard deviation σ^2 . By definition these quantities are expressed as

$$\langle k \rangle = \int kP(k)dk \quad (2)$$

$$\sigma^2 = \int (k - \langle k \rangle)^2 P(k)dk \quad (3)$$

The peculiar fact about a distribution with a heavy tail is that there is a finite probability of finding vertices with degree much larger than the average $\langle k \rangle$. In other words, the consequence of heavy tails is that the average behavior of the system is not typical. The characteristic degree is the one that, picking up a vertex at random, should be encountered most of the times. In power-law distributions most of the times vertices will have a small degree, but there is an appreciable probability of finding vertices with large degree values. Yet all intermediate values are probable and the average degree does not represent any special value for the distribution. We are in presence of very heterogeneous networks. This is clearly opposite to bell-shaped distributions with fast decaying tails, in which the average value is very close to the maximum of the distribution and represents the most probable value in the system. In more mathematical terms the heavy-tail property translates in a very large level of degree fluctuations. In the case of distributions with a power-law tail with exponent $2 \leq \gamma \leq 3$ we have that fluctuations are therefore unbounded and depend only on the system size. The absence of any intrinsic scale for the fluctuations implies that the average value is not a characteristic scale for the system. In other words, we are in presence of a *scale-free* network for what concerns the statistical properties of the vertices' degree. This reasoning can be extended to values of $\gamma \leq 2$, since in this case even the first moment is unbounded. The power-law behavior and the relative exponent thus represent a quantitative measure of the level of *heterogeneity* of the network's degree.

1.6. Complexity

While the extreme heterogeneity of networks is a well defined mathematical property, the definition of complex networks implies the distinction of what is "complex" and what is the merely complicated. This distinction is a critical one because the characteristic features and the behavior of complex

systems differ significantly from those of merely complicated systems. A general and accepted definition of complexity does not exist. Authors in different context provide different definitions which are often tailored on specific systems or areas of interest. Without entering in the details of such a discussion, however, a minimal definition of complexity may involve two main features: i) the system exhibits complications and heterogeneity that extend virtually on all scales allowed by the physical size of the system; ii) these features are the spontaneous outcome of the interactions among the many constituent units of the system, i.e. we are in the presence of an emergent phenomenon. It is easy to realize that the WWW, the Internet, the airport network are all systems which grow in time by following complicate dynamical rules and without a global supervision or blueprint. The same can be said for many social and biological networks. All these networks are self-organizing systems, which at the end of the evolution show an emergent architecture with unexpected properties and regularities. At the same time, heavy tails and heterogeneity appear to be common to a large number of these networks, along with other complex topological features such the presence of communities, motifs, hierarchies and modular ordering. We are thus in the presence of structures whose fluctuations and complications are unbounded and extend over all possible scales allowed by the physical size of the systems, therefore defining the class of complex networks.

1.7. *What is next*

The characterization of large complex networks goes far beyond the basic properties discussed in the previous sections. Real networks comprises systems of a very different nature that show several other complex structural properties that might differ from case to case. The increasing evidence in networks for the presence of communities, motifs, hierarchies and modular ordering opens a series of important questions and at the same time defines different classes of complex networks. In this perspective, it becomes particularly relevant to develop specific tools for the characterization and analysis of large scale networks as well as a theoretical understanding that might uncover the very general principles underlying the networks formation. In the next chapters the reader will find an extensive review of recent studies concerning the structural analysis of complex networks and the applications of these concepts and models to several real world systems. Chapter 2 presents the basic modeling paradigms for static and evolving networks. Chapter 3 offers a discussion on the empirical analysis and modeling of cor-

relations and clustering in complex networks. In Chapter 4 the reader finds an introduction to weighted networks and their relevance in the modeling of real world networks. Chapter 5 is devoted to the analysis of communities and motifs in complex networks and the methods for their detection. The last methodological Chapter 6 addresses the questions arising in the visualization of large networks and presents recent tools and developments in this area. The remaining chapters are devoted to specific domain applications. Namely Chapter 7 deals with the world-wide web graph; Chapter 8 discusses the analysis of the real Internet; Chapter 9 reviews the use of networks in the ecological domain; Chapter 10 presents some recent applications of network analysis to large scale socio-economical networks. We are confident that the following part might provide a “state-of-the-art” discussion of the complex networks research and its application in many real world instances.