1 Introduction

Missing information: a general problem. Laying at the heart of the scientific method, data analysis is about using data to validate models, acquire useful information and support decision-making. When the data is incomplete, so are the conclusions that can be drawn from it. Unfortunately, the problem of missing data is a common occurrence, both in science and in many practical situations. Even in the era of Big Data, data can be incomplete for a variety of reasons – such as sheer lack of information, annotation errors, collection problems and privacy concerns. The problem is even more severe when the data has a non-homogeneous structure, because it describes non-trivial, systemic interconnection patterns, such as those characterizing complex networks. To be more concrete, we draw a few examples.

Consider a system biologist looking for the proteins in an organism that have physical, or functional, pairwise interactions. The scientist would need to pick two candidate proteins and set up an experiment to determine whether they interact or not. Blindly considering all possible pairs is unfeasible because experiments can be quite costly: this is why interactions within the proteome are largely unknown, whence the need to pick good candidates for the experiment, using prior information on those interactions that have already been discovered (Redner, 2008; Guimerà and Sales-Pardo, 2009). As another example, consider a social scientist trying to build up a given social network. Two types of problems can arise in this context: i) the available data reports only aggregated statistics on individuals (e.g. the total number of contacts) without disclosing sensible information such as the identities of friends; and ii) the network is extremely large to be explored by crawling algorithms, whence the need to consider subsamples that are representative (Leskovec and Faloutsos, 2006; Liben-Nowell and Kleinberg, 2007).

Farther from the classical typical scientific domain, consider an entrepreneur running an e-commerce platform that sells books. In order to improve sales, it would be a good idea to set up a *recommender system* that shows customers the books they may be interested in buying. The algorithm works well if it is able to predict customer tastes (i.e. possible future purchases) using their buying records (Lü et al., 2012). As a final example we take a regulator working in a central bank. Her job is to run stress tests to determine whether a given bank can withstand a crisis event. Since in a financial system losses and distress propagate through the various financial exposures banks have with each other and with other financial institutions, to accomplish her task properly the regulator should know the detailed network of exposures (who is exposed with whom, and to what extent). Unfortunately this information is confidential, and

the regulator must resort to publicly disclosed information, i.e. the balance sheet of the banks containing only their aggregate exposures (Squartini et al., 2018; Anand et al., 2018).

The common theme of all these situations is that the system at hand is a *network*, namely a system that independently of its nature can be modelled by a complex pattern of interactions (the *links*) between its constituents (the *nodes*). When the network is known only partially, the task is to reconstruct the unknown part. The techniques constituting the field known as *network reconstruction* precisely aim at inferring the (unknown) structure of a network, making an optimal use of the partial knowledge about its properties (Squartini et al., 2018; Lü and Zhou, 2011).

Approaching network reconstruction. Generally speaking, the fundamental assumption at the basis of network reconstruction is *statistical homogeneity*: the empirically observed network structures should be representative of the statistical properties concerning the network as a whole. The validity of such an assumption is the necessary condition for a reconstruction algorithm to work. Clearly, this approach limits the accuracy that can be achieved when reconstructing strongly heterogeneous structures. However, it prevents possible inference biases introduced by arbitrary assumptions not supported by the available information.

In order to deal with the problem of missing information, many different approaches have been attempted so far. Among the most successful ones there are those defined within the framework of information theory (Cover and Thomas, 2006). In a nutshell, these methods prescribe to 1) consider all configurations that are compatible with the available information (an *ensemble*, in the jargon of statistical mechanics) and 2) assign a degree of plausibility to each of them. As it has been proven elsewhere, the least-biased way to do this rests upon the renowned *entropy maximization* prescription (Cover and Thomas, 2006; Jaynes, 1957). Notably, this approach naturally leads to the Exponential Random Graphs (ERG) formalism (Park and Newman, 2004b; Cimini et al., 2019). The importance of ERG models within the network reconstruction field is motivated by three desirable features they possess: *analytical character*, *general applicability* and *versatility*. This is why a large portion of this Element is devoted to discussing the applications of such a powerful formalism.

A quick overview of the Element. The discussion of the network reconstruction problem is divided into three sections, according to the scale of the reconstruction task: *macroscale*, *mesoscale* and *microscale*. This distinction is intended

to provide a wide overview of the reconstruction techniques while presenting detailed results in some specific contexts.

The section **Network reconstruction at the macroscale** focuses on the inference of global features of the network, such as *(dis)assortative* and *hierarchical* patterns. In this case, reconstruction techniques are typically informed on node-specific properties (and possibly on trends that characterize the network as a whole), without considering any specific topological detail (i.e. the occurrence of a particular link). After describing the general ideas and results, particularly in the context of ERG, we will delve into the estimation of *systemic risk* in a partially-accessible network. As already mentioned, this exercise is particularly relevant for financial networks, where the knowledge of the interconnections between financial institutions is required to run stress tests and assess the stability of the system.

The section **Network reconstruction at the mesoscale** instead deals with the detection and reproduction of network patterns like *modular*, *core-periphery* and *bipartite* structures. The topic is of great interest for disciplines as diverse as epidemiology, finance, biology and sociology as it ultimately boils down to identifying some sort of *structural* or *functional* similarity between nodes. The presence of mesoscale patterns then affects a wide range of dynamical processes *on* networks (e.g. information and epidemic spreading, fake-news diffusion, etc.) whence the need to properly account for them. A fundamental point is to understand to what extent *accessible* node properties are informative about the presence of mesoscale structures.

Finally, the section **Network reconstruction at the microscale** is devoted to the topic of *single link* inference, a problem that is better known as *link prediction*. In contrast to the network reconstruction problem at the macro- and at the meso-scale, when considering the micro-scale many details of the network are known (typically a large number of connections) and the goal is to predict those links that are either not known because the source data used to define the network is incomplete, or simply do not exist yet. We will review the link prediction techniques that build on the partial knowledge on the network, and not on any additional information like nodes features.

2 Network Reconstruction at the Macroscale

A network is defined as a set of constituent elements (the *nodes*) and a set of connections (the *links*) among them. Mathematically speaking, a network is a graph with nontrivial topological features. In practice, networks are the natural way to represent and model a large class of very diverse systems, and thus we can speak of technological and information networks, social and economic networks as well as biological and brain networks.

Macroscale Properties: An Overview

Binary Properties

Let us start by introducing the basic notation and the macroscale properties of *binary*, *undirected* (*directed*) *graphs* with N nodes. Graphs of this kind are completely specified by a symmetric (generally asymmetric) $N \times N$ *adjacency matrix* $\mathbf{A}$, whose generic entry is either $a_{ij} = 0$ or $a_{ij} = 1$, respectively indicating the absence or the presence of a connection between nodes i and j (from i to j). As usual, *self-loops*, namely links starting and ending at the same node, will be ignored (in formulas, $a_{ii} = 0, \ \forall \ i$). The description above also applies to *bipartite* graphs, where nodes form two disjoint sets that are not connected internally.

Connectance. The simplest macroscopic characterization of a network is the connectance, or *link density*, defined as

$$\rho(\mathbf{A}) = \frac{2L}{N(N-1)}, \tag{2.1}$$

where $L(\mathbf{A}) = \sum_{i<j} a_{ij} \equiv L$ is the total number of links in the network. Thus $\rho(\mathbf{A})$ is the fraction of node pairs that are connected by a link. For directed networks, $\rho(\mathbf{A}) = \frac{L}{N(N-1)}$ with $L = \sum_{i \neq j} a_{ij}$.

Notably, real-world networks are usually characterized by a very low density of links (i.e. they are sparse). Reproducing the network connectance is a sort of baseline requirement of any reconstruction method. The simplest model satisfying this requirement is the *Erdös-Renyi* (ER) random graph (Erdös and Rényi, 1960; Park and Newman, 2004b). According to this model, the probability p_{ij} of a connection between nodes i and j (that is, the average value of the adjacency matrix element a_{ij} in the model, $\langle a_{ij} \rangle_{\mathrm{ER}}$) reads

$$p_{ij}^{\mathrm{ER}} = p = \frac{2L}{N(N-1)} = \rho, \ \forall \ i < j; \tag{2.2}$$

therefore, any two nodes establish a connection with the same probability p.

Degrees. The *degree* of a node counts the number of its neighbors, or equivalently the number of its incident connections. In formulas, $k_i(\mathbf{A}) = \sum_{j \neq i} a_{ij}, \forall i$. An important and ubiquitous characterization of real-world networks is the *heavy-tailed* shape of the degree distribution, with a few *hub* nodes that are highly connected ($k_{hub} = O(N)$) and the vast majority of other nodes (of the order $O(N)$) with a small degree. Although the mathematical nature of these heavy-tailed distributions is still debated, often they have been found to be *scale-free* (Caldarelli, 2007; Barabási, 2009):

$$P(k) \sim k^{-\gamma},\ 2 < \gamma < 3, \tag{2.3}$$

for which the "typical" degree is simply missing. In any event, the strong heterogeneity of the degree distribution is the basic feature that makes networks different from homogeneous systems and regular lattices. Therefore, any good reconstruction algorithm should be able to reproduce it.[1] Notice that such a requirement rules out the ER model as a potentially good reconstruction model: in fact, although it ensures that the link density is reproduced, it fails to preserve the degree heterogeneity, since the model average $\langle k_i \rangle_{\rm ER} = \sum_{j \neq i} \langle a_{ij} \rangle_{\rm ER} = \sum_{j \neq i} p_{ij}^{\rm ER} = 2L/N,\ \forall\ i$. Such evidence has motivated the definition of the Chung-Lu (CL) model (Chung and Lu, 2002), according to which

$$p_{ij}^{\rm CL} = \frac{k_i k_j}{2L},\ \forall\ i \neq j; \tag{2.4}$$

by definition, then, $\langle k_i \rangle_{\rm CL} = \sum_{j \neq i} \langle a_{ij} \rangle_{\rm CL} = \sum_{j \neq i} p_{ij}^{\rm CL} \simeq k_i,\ \forall\ i$.

In the directed case, there are two kinds of degree: the total number of links outgoing from a node (the *out-degree* $k_i^{out}(\mathbf{A}) = \sum_{j \neq i} a_{ij},\ \forall\ i$) and the total number of links incoming to a node (the *in-degree* $k_i^{in}(\mathbf{A}) = \sum_{j \neq i} a_{ji},\ \forall\ i$). The directed extension of the Chung-Lu model (DCL) reads

$$p_{ij}^{\rm DCL} = \frac{k_i^{out} k_j^{in}}{L},\ \forall\ i \neq j. \tag{2.5}$$

Assortativity. Generally speaking, this term indicates the tendency of nodes to establish connections with other nodes having either similar (*positive assortativity*) or different (*negative assortativity* or *disassortativity*) characteristics. Particularly relevant in the study of complex networks is the assortativity *by degree*. In this case, assortativity can be studied by considering the *average nearest neighbor degree* (ANND), which for generic node i is defined as

$$k_i^{nn}(\mathbf{A}) = \frac{\sum_{j \neq i} a_{ij} k_j}{k_i},\ \forall\ i. \tag{2.6}$$

ANND is a quadratic function of the adjacency matrix and thus is a *second-order* network property. Plotting k_i^{nn} versus k_i reveals the two-point correlation structure of the network: an increasing trend corresponds to an assortative pattern (poorly connected nodes are connected to other poorly connected nodes, highly connected nodes are connected to other highly connected nodes), while a decreasing trend corresponds to the opposite disassortative pattern (poorly connected nodes are connected to highly connected nodes and vice versa). Notice that assortativity is typically observed in social networks (where it is also

[1] Moreover, preserving the degrees automatically ensures that the link density is preserved.

known by the term *homophily*), whereas economic and technological networks are usually disassortative (Newman, 2002).

Assortativity acts as the test bench for the CL model. Since $\langle k_i^{nn}\rangle_{\rm CL} \simeq \frac{\sum_{j\neq i} p_{ij}^{\rm CL} k_j}{k_i} = \frac{\sum_{j\neq i} k_j^2}{2L}$, $\forall\ i$, in this model k_i^{nn} is weakly dependent on node i – basically, the ANND is the same for all nodes (Squartini and Garlaschelli, 2011). As a consequence, the CL model is not capable of reproducing any (dis)assortativity, thus it lacks one of the characteristic features of real-world networks. The solution lies in the definition of a more refined model, the *Configuration Model* (CM – see below).

When considering directed networks, the ANND can be generalized in five different ways (see Squartini et al., 2011a for further details).

Hierarchy. Assortativity and ANND account for second-order interactions, that is, interactions between nodes along patterns of length two. Third-order interactions (i.e. three-point correlations) are instead typically measured through the *clustering coefficient*, which for any node i is defined as the percentage of pairs of neighbors of i that are also neighbors of each other:

$$c_i(\mathbf{A}) = \frac{\sum_{j\neq i}\sum_{k\neq i,j} a_{ij}a_{ik}a_{jk}}{k_i(k_i-1)},\ \forall\ i; \tag{2.7}$$

otherwise stated, c_i measures the fraction of potential triangles attached to i (and defined by the product $a_{ij}a_{ik}$) that are actually realized (i.e. closed by the third link, a_{jk}). A decreasing trend of c_i as a function of k_i indicates that neighbors of highly connected nodes are poorly interconnected, whereas neighbors of poorly connected nodes are highly interconnected. This behavior characterizes a *hierarchical* network (i.e., a network of densely connected subgraphs that are poorly interconnected). In real-world networks, a scale-free degree distribution often coexists with a large value of the clustering coefficient (Albert and Barabási, 2002).

As for the assortativity, the CL model predicts a value for the clustering coefficient that is only weakly dependent on i, thus calling for a more refined model to reproduce empirical patterns of real-world networks.

Generalizations to directed networks also exist for third-order quantities. Besides five different definitions of the clustering coefficient (Squartini et al., 2011a), there are thirteen possible patterns involving three nodes and all possible connections between them: these quantities are called *motifs* and, as discussed in Section 3, have been proven to play a fundamental role in the self-organization of biological, ecological, and cellular networks. Certain *structures* have been, in fact, suggested to promote specific *functions* (Milo et al., 2002).

Higher-Order Patterns. The presence of higher-order patterns can be inspected using the powers of the network adjacency matrix $\mathbf{A}$. Indeed, the entry indexed by i and j of $\mathbf{A}^n$ (i.e., the nth power of $\mathbf{A}$) counts the number of paths of length n existing between i and j (or from i to j).

A very popular higher-order pattern is given by the *shortest path length*, a concept entering into the definition of the well-known *small-world effect* (Watts and Strogatz, 1998). Small-worldness refers to the evidence that, in many real-world networks, two (apparently) competing features coexist: a large clustering coefficient and a small average shortest path length. More quantitatively, the small-world phenomenon is characterized by an average shortest path length typical of random graphs

$$\bar{d} \simeq d_{\text{random}} \propto \ln N \tag{2.8}$$

(i.e., growing "slowly" with the size of the system) and by an average clustering coefficient typical of regular lattices (i.e. independent of the system size), much larger than that of a random graph

$$\bar{c} \gg \bar{c}_{\text{random}} \propto N^{-1} \tag{2.9}$$

where, in both expressions, the term "random" refers to the ER model.

Nestedness. A pattern that has recently attracted much attention is the *nestedness*. It quantifies how much the *biadjacency* matrix of a bipartite network can be rearranged to let a triangular structure emerge (Johnson et al., 2013; Mariani et al., 2019). Several measures have been defined to quantify the nestedness, among which the NODF (an acronym for "Nestedness metric based on Overlap and Decreasing Fill") that quantifies a matrix "triangularity" by measuring the overlap between rows and between columns (Almeida-Neto et al., 2008). Nestedness has been observed in ecological and economic systems alike. The classical example of nested ecological systems is given by the interactions between plants and pollinators, where nestedness emerges due to the presence of generalist pollinators (being attracted by all species of plants) coexisting with specialist pollinators (being attracted by only a small number of species of plants). Such a structure has been argued to promote the stability of the ecosystem (Bascompte et al., 2003). For what concerns economic systems, nestedness is observed in the structure of countries' exports: a few very diversified countries have a large export basket, while others only export some simple products. Interestingly, this pattern contradicts classical economic theories, according to which countries should specialize and export only those products in which they have a competitive advantage, and implying a block-diagonal biadjacency matrix instead of a nested one (Tacchella et al., 2012).

Centrality. The concept of *centrality* aims at quantifying the "importance" of a node in a network (Newman, 2018a). Besides *degree centrality*, the centrality given by the degree, other well-known measures are the *closeness centrality*, defined as

$$C_i(\mathbf{A}) = \frac{1}{\bar{d}_i}, \ \forall\, i \tag{2.10}$$

(i.e., as the reciprocal of the average topological distance of a node from the others), and the *betweenness centrality*, defined as

$$B_i(\mathbf{A}) = \sum_{j\neq i}\sum_{k\neq i,j}\frac{\sigma_{jk}(i)}{\sigma_{jk}}, \ \forall\, i, \tag{2.11}$$

where σ_{jk} is the total number of shortest paths from node j to node k, and $\sigma_{jk}(i)$ is the number of these paths passing through i.

Most of the proposed centrality measures are computable only on undirected networks. A notable exception is the *PageRank centrality* (Page et al., 1999), which can be computed by solving the iterative equation

$$P_i(\mathbf{A}) = \frac{1-\alpha}{N} + \alpha\sum_{j\neq i}\left(\frac{a_{ji}}{k_j^{out}}\right)P_j(\mathbf{A}), \ \forall\, i. \tag{2.12}$$

In general, it is very difficult to reconstruct the patterns of centrality of a network, unless these are strongly correlated with the degree centrality (Barucca et al., 2018).

Reciprocity. In the specific case of directed networks, it is of particular interest to measure the percentage of links having a counterpart pointing in the opposite direction. This quantity is known as *reciprocity* and reads

$$r(\mathbf{A}) = \frac{L^{\leftrightarrow}}{L} = \frac{\sum_i\sum_{j\neq i}a_{ij}a_{ji}}{\sum_i\sum_{j\neq i}\sum_i a_{ij}}; \tag{2.13}$$

remarkably, different classes of real-world networks are characterized by different values of reciprocity (Garlaschelli and Loffredo, 2004b). For instance, reciprocity is a distinguishing feature of financial networks, being associated with the level of "trust" between banks (Squartini et al., 2013a).

Spectral Properties. This term refers to the features of eigenvalues and eigenvectors of both the adjacency matrix $\mathbf{A}$ and the *Laplacian matrix* $\mathbf{L} = \mathbf{D} - \mathbf{A}$ of the network.[2] (Here $\mathbf{D}$ is the diagonal matrix whose generic entry reads

[2] The focus on undirected binary networks is justified by the ease of treating symmetric matrices, a characteristic ensuring that eigenvalues are real, for example.

$d_{ii} = k_i, \ \forall \ i$.) While Laplacian spectral properties provide information on macroscale network properties like the number of connected components (that matches the multiplicity of the zero eigenvalue of $\mathbf{L}$), spectral properties of $\mathbf{A}$ provide information on higher-order patterns like cycles (Estrada and Knight, 2015) as well as on dynamical properties of spreading processes (Bardoscia et al., 2017). Notice that the reconstruction of spectral properties of empirical networks is still a largely underexplored topic, although a first result in this sense is provided by the Silverstein theorem (Silverstein, 1994).

Weighted Properties

While binary networks are characterized by an adjacency matrix whose entries assume only the values 0 and 1, *weighted*, *undirected* (*directed*) graphs are specified by a symmetric (generally asymmetric) $N \times N$ matrix $\mathbf{W}$ whose generic entry $w_{ij} > = 0$ quantifies the intensity of the link connecting nodes i and j: In the most general case, w_{ij} is a real number; however, in many cases w_{ij} assumes integer values. Naturally, $\mathbf{A}$ and $\mathbf{W}$ are related by the position $a_{ij} = \Theta[w_{ij}], \ \forall \ i,j$, simply stating that any positive weight between i and j carries the information that i and j are indeed connected.

Weight Distribution. When links are characterized by "magnitudes," the first step is to inspect the distribution of these magnitudes. When considering real-world networks, weight distributions are often found to be *fat-tailed*.

Strengths. The weighted analogue of the degree is the so-called *strength*. It is defined as $s_i(\mathbf{W}) = \sum_{j \neq i} w_{ij}, \ \forall \, i$ (i.e., as the sum of the weight of the links connected to node i). Similar to the case of degrees, strength distributions are often found to be *fat-tailed*. When directed networks are considered, one speaks of *out-strength* and *in-strength*, respectively defined as $s_i^{out}(\mathbf{W}) = \sum_{j \neq i} w_{ij}, \ \forall \ i$ and $s_i^{in}(\mathbf{W}) = \sum_{j \neq i} w_{ji}, \ \forall \ i$.

From a network reconstruction perspective, strengths play an important role, since they often represent the only kind of information available for the system under consideration. The typical example is that of financial networks, where only the total *assets* and *liabilities* of each bank (respectively the out- and in-strengths of the respective node) are accessible. This has motivated the definition of the weighted analogue of the Chung-Lu model, also known as the *MaxEnt* (ME) recipe. Its directed version, reading

$$\hat{w}_{ij}^{\mathrm{ME}} = \frac{s_i^{out} s_j^{in}}{W}, \ \forall \ i \neq j \tag{2.14}$$

(with $W = \sum_i s_i^{out} = \sum_i s_i^{in}$), is extensively used to estimate the magnitude of links in economic and financial networks (Mistrulli, 2011; Upper, 2011; Squartini et al., 2018).

Weighted Assortativity. The concept of assortativity can be easily extended to the weighted case. The weighted counterpart of the average nearest neighbors degree of node i is the *average nearest neighbor strength* (ANNS):

$$s_i^{nn}(\mathbf{W}) = \frac{\sum_{j\neq i} a_{ij}s_j}{k_i}, \ \forall \ i. \tag{2.15}$$

Analogous to the binary case, the correlation between strengths can be inspected by plotting s_i^{nn} versus s_i. Note that since $\langle a_{ij}\rangle_{\mathrm{ME}} = p_{ij}^{\mathrm{ME}} = \Theta[\hat{w}_{ij}^{\mathrm{ME}}]$, the weighted version of the CL model always generates a very densely connected network and, as a consequence, a value for the ANNS of node i that is weakly dependent on i itself, that is, $\langle s_i^{nn}\rangle_{\mathrm{ME}} \simeq \frac{\sum_{j\neq i} p_{ij}^{\mathrm{ME}} s_j}{\langle k_i\rangle_{\mathrm{ME}}} \simeq \frac{\sum_{j\neq i} s_j}{N-1} \simeq \frac{2W}{N-1}$, $\forall\, i$ (Squartini and Garlaschelli, 2011). The weighted CL model thus suffers from the same limitations affecting the binary CL model.[3]

Weighted Hierarchy. A *weighted clustering coefficient* (WCC) can be defined to capture the "intensity" of the triangles in which node i participates (Squartini et al., 2011b):

$$c_i^w(\mathbf{W}) = \frac{\sum_{j\neq i}\sum_{k\neq i,j}(w_{ij}w_{jk}w_{ki})^{1/3}}{k_i(k_i-1)}. \tag{2.16}$$

Contrary to what is observed for the vast majority of binary networks, plotting c_i^w versus s_i reveals an increasing trend for many real-world networks, indicating that nodes with larger total activity participate in more "intense" triangles.

For extensions of ANNS and WCC to directed networks, see Squartini et al. (2011b).

Weighted Reciprocity. A weighted version of link reciprocity can be defined as

$$r^w(\mathbf{W}) = \frac{W^{\leftrightarrow}}{W} = \frac{\sum_i\sum_{j\neq i}\min[w_{ij}, w_{ji}]}{\sum_i\sum_{j\neq i} w_{ij}}, \tag{2.17}$$

a quantity whose numerator accounts for the "minimum exchange" between any two nodes (Squartini et al., 2013b).

[3] As we will see in what follows, the weighted counterpart of the CM, namely the *Weighted Configuration Model* (WCM), does not represent the solution to this problem. We will need to consider degrees and strengths together as in the *Enhanced Configuration Model* (ECM).