Spatial networks and small worlds

Luca Lombardo

Dicember 2022

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Contents

1 Introduction

Given a social network, which of its nodes are more central? This question has been asked many times in sociology, psychology and computer science, and a whole plethora of centrality measures (a.k.a. centrality indices, or rankings) were proposed to account for the importance of the nodes of a network.

These networks, typically generated directly or indirectly by human activity and interaction (and therefore hereafter dubbed social"), appear in a large variety of contexts and often exhibit a surprisingly similar structure. One of the most important notions that researchers have been trying to capture in such networks is "node centrality": ideally, every node (often representing an individual) has some degree of influence or importance within the social domain under consideration, and one expects such importance to surface in the structure of the social network; centrality is a quantitative measure that aims at revealing the importance of a node.

Among the types of centrality that have been considered in the literature, many have to do with distances between nodes. Take, for instance, a node in an undirected connected network: if the sum of distances to all other nodes is large, the node under consideration is peripheral; this is the starting point to define Bavelas's closeness centrality [1], which is the reciprocal of peripherality (i.e., the reciprocal of the sum of distances to all other nodes).

The role played by shortest paths is justified by one of the most well-known features of complex networks, the so-called small-world phenomenon. A small-world network [2] is a graph where the average distance between nodes is logarithmic in the size of the network, whereas the clustering coefficient is larger (that is, neighborhoods tend to be denser) than in a random Erdős-Rényi graph with the same size and average distance. The fact that social networks (whether electronically mediated or not) exhibit the small-world property is known at least since Milgram's famous experiment [] and is arguably the most popular of all features of complex networks. For instance, the average distance of the Facebook graph was recently established to be just 4.74 [3].

2 Characterization of networks

Before 1960, graph theory mainly dealt with the properties of specific individual graphs. In the 1960s, Paul Erdős and Alfred Rényi initiated a systematic study of random graphs [1]. Two well-studied graph ensembles are $G_{N,M}$, the ensemble of all graphs with N nodes and M edges, and $G_{N,p}$, the ensemble of all graphs with N nodes and probability p of any two nodes being connected.

An important attribute of a graph is the average degree, i.e., the average number of edges connected to each node. We will denote the degree of the ith node by k_i and the average degree by $< r >$. N-vertex graphs with $< k >= O(N^0)$ are called sparse graphs.

The Erdős-Rényi model has traditionally been the dominant subject of study in the field of random graphs. Recently, however, several studies of real-world networks have found that the ER model fails to reproduce many of their observed properties. One of the simplest properties of a network that can be measured directly is the degree distribution, or the fraction P(k) of nodes having k connections (degree k). A well-known result for ER networks is that the degree distribution is Poissonian,

$$P(k) = \frac{e^z z^k}{k!} \tag{1}$$

Where $z =< k >$. is the average degree. Direct measurements of the degree distribution for real networks show that the Poisson law does not apply. Rather, often these nets exhibit a scale-free degree distribution:

$$P(k) = ck^{-\gamma} \quad \text{for} \quad k = m, ..., K \tag{2}$$

Where $c \sim (\gamma - 1)m^{\gamma-1}$ is a normalization factor, and m and K are the lower and upper cutoffs for the degree of a node, respectively. All real-world networks are finite and therefore all their moments are finite. The actual value of the cutoff K plays an important role. It may be approximated by noting that the total probability of nodes with $k > K$ is of order $1/N$

$$\int_K^\infty P(k)dk \sim \frac{1}{N} \tag{3}$$

This yields the result

$$K \sim mN^{1/(\gamma-1)} \tag{4}$$

The degree distribution alone is not enough to characterize the network. There are many other quantities, such as the degree-degree correlation (between connected nodes), the spatial

[1] Random graph theory is, not the study of individual graphs, but the study of a statistical ensemble of graphs (or, as mathematicians prefer to call it, a probability space of graphs). The ensemble is a class consisting of many different graphs, where each graph has a probability attached to it. A property studied is said to exist with probability P if the total probability of a graph in the ensemble possessing that property is P (or the total fraction of graphs in the ensemble that has this property is P). This approach allows the use of probability theory in conjunction with discrete mathematics for studying graph ensembles.

correlations, the clustering coefficient, the betweenness or central-ity distribution, and the self-similarity exponents.

2.1 Random graphs as a model of real networks

Many natural and man-made systems are networks, i.e., they consist of objects and interactions between them. These include computer networks, in particular the Internet, logical networks, such as links between WWW pages, and email networks, where a link represents the presence of a person's address in another person's address book. Social interactions in populations, work relations, etc. can also be modeled by a network structure. Networks can also describe possible actions or movements of a system in a configuration space (a phase space), and the nearest configurations are connected by a link. All the above examples and many others have a graph structure that can be studied. Many of them have some ordered structure, derived from geographical or geometrical considerations, cluster and group formation, or other specific properties. However, most of the above networks are far from regular lattices and are much more complex and random in structure. Therefore, it is plausible that they maintain many properties of the appropriate random graph model.

For large γ (usually, for $\gamma > 4$) the properties of scale-free networks, such as distances, optimal paths, and percolation, are the same as in ER networks. In contrast, for $\gamma < 4$, these properties are very different and can be regarded as anomalous.

2.2 Properties of real-world networks

`Una piccola introduzione`

2.2.1 Degree distribution

The degree of a node is the number of links connected to it. In directed networks, one can distinguish between the in-degree, out-degree, and the total degree (which is the sum of the two). The degree distribution, $P(k)$, is the fraction of sites having degree k. As can be seen above, many real networks do not exhibit a Poisson degree distribution, as predicted in the ER model. In fact, many of them exhibit a distribution with a long, power-law, tail, $P(k) \sim k^{-\gamma}$ with some γ, usually between 2 and 3.

2.2.2 Distances and optimal paths

Since many networks are not embedded in real space, the geometrical distance between nodes is meaningless. The most important distance measure in such net- works is the minimal number of hops (or chemical distance). That is, the distance between two nodes in the network is defined as the number of edges in the shortest path between them. If the edges are assumed to be weighted, the lowest total weight path, called the optimal path, may also be used. The usual mathematical definition of the diameter of the network is the length of the path between the farthest nodes in the network.

2.2.3 Clustering

The clustering coefficient is usually related to a community represented by local structures. The usual definition of clustering (sometimes also referred to as transi- tivity) is related to the number of triangles in the network. The clustering is high if two nodes sharing a neighbor have a high probability of being connected to each other. There are two common definitions of clustering. The first is global,

$$C = \frac{3 \times \text{the number of triangles in the network}}{\text{the number of connected triples of vertices}} \tag{5}$$

where a "connected triple" means a single vertex with edges running to an unordered pair of other vertices.

A second definition of clustering is based on the average of the clustering for single nodes. The clustering for a single node is the fraction of pairs of its linked neighbors out of the total number of pairs of its neighbors:

$$C_i = \frac{\text{the number of triangles connected to vertex } i}{\text{the number of triples centered on vertex } i} \tag{6}$$

For vertices with degree 0 or 1, for which both numerator and denominator are zero, we use $C_i = 0$. Then the clustering coefficient for the whole network is the average

$$C = \frac{1}{n} \sum_i C_i \tag{7}$$

In both cases the clustering is in the range $0 \leq C \leq 1$. In random graph models such as the ER model and the configuration model, the clustering coefficient is low and decreases to 0 as the system size increases. This is also the situation in many growing network models. However, in many real-world networks the clustering coefficient is rather high and remains constant for large network sizes. This observation led to the introduction of the small-world model, which offers a combination of a regular lattice with high clustering and a random graph.

2.2.4 Betweenness centrality

Path-based measures exploit not only the existence of shortest paths but actually take into examination all shortest paths (or all paths) coming into a node. We remark that in-degree can be considered a path-based measure, as it is the equivalent to the number of incoming paths of length one.

Betweenness centrality was introduced for edges, and then rephrased. The idea is to measure the probability that a random shortest path passes through a given node: if σ_{yz} is the number of shortest paths going from y to z, and $\sigma yz(x)$ is the number of such paths that pass through x, we define the betweenness of x as

$$\beta(x) = \sum_{y \neq x \neq z} \frac{\sigma_{yz}(x)}{\sigma_{yz}}. \tag{8}$$

The intuition behind betweenness is that if a large fraction of shortest paths passes through x, then x is an important junction point of the network. Indeed, removing nodes in betweenness order causes a very quick disruption of the network.

3 The Small-World Phenomenon

The Aim of the project is to study the small-world phenomenon in location-based (social) networks. As test cases, we consider three real-world datasets: Brightkite, Gowalla and Foursquare. In the next sections, we will describe the datasets and the methodology we used to extract the networks from them.

Many real-world networks have many properties that cannot be explained by the ER model. One such property is the high clustering observed in many real-world networks. This led Watts and Strogatz to develop an alternative model, called the "small-world" model [WS98]. Their idea was to begin with an ordered lattice, such as the k-ring (a ring where each site is connected to its $2k$ nearest neighbors - k from each side) or the two-dimensional lattice. A variant of this process is to add links rather than rewire, which simplifies the analysis without considerably affecting the results. The obtained network has the desirable properties of both an ordered lattice (large clustering) and a random network (small world), as we will discuss below.

3.1 Clustering in a small-world network

The simplest way to treat clustering analytically in a small-world network is to use the link addition, rather than the rewiring model. In the limit of large network size, $N \to \infty$, and for a fixed fraction of shortcuts ϕ, it is clear that the probability of forming triangle vanishes as we approach $1/N$, so the contribution of the shortcuts to the clustering is negligible. Therefore, the clustering of a small-world network is determined by its underlying ordered lattice. For example, consider a ring where each node is connected to its k closest neighbors from each side. A node's number of neighbors is therefore 2k, and thus it has 2k(2k - 1)/2 = k(2k - 1) pairs of neighbors. Consider a node, i. All of the k nearest nodes on i's left are connected to each other, and the same is true for the nodes on i's right. This amounts to 2k(k - 1)/2 = k(k - 1) pairs. Now consider a node located d places to the left of k. It is also connected to its k nearest neighbors from each side. Therefore, it will be connected to k - d neighbors on i's right side. The total number of connected neighbor pairs is

$$k(k-1) + \sum_{d=1}^{k} (k-d) = k(k-1) + \frac{k(k-1)}{2} = \frac{32}{k}(k-1) \tag{9}$$

and the clustering coefficient is:

$$C = \frac{3(k-1)}{2(2k-1)} \tag{10}$$

For every k > 1, this results in a constant larger than 0, indicating that the clustering of a small-world network does not vanish for large networks. For large values of k, the clustering

coefficient approaches 3/4, that is, the clustering is very high. Note that for a regular two-dimensional grid, the clustering by definition is zero, since no triangles exist. However, it is clear that the grid has a neighborhood structure.

3.2 Distances in a small-world network

The second important property of small-world networks is their small diameter, i.e., the small distance between nodes in the network. The distance in the underlying lattice behaves as the linear length of the lattice, L. Since $N \sim L^d$ where d is the lattice dimension, it follows that the distance between nodes behaves as:

$$l \sim L \sim N^{1/d} \tag{11}$$

Therefore, the underlying lattice has a finite dimension, and the distances on it behave as a power law of the number of nodes, i.e., the distance between nodes is large. However, when adding even a small fraction of shortcuts to the network, this behavior changes dramatically.

Let's try to deduce the behavior of the average distance between nodes. Consider a small-world network, with dimension d and connecting distance k (i.e., every node is connected to any other node whose distance from it in every linear dimension is at most k). Now, consider the nodes reachable from a source node with at most r steps. When r is small, these are just the *r-th* nearest neighbors of the source in the underlying lattice. We term the set of these neighbors a "patch". the radius of which is kr , and the number of nodes it contains is approximately $n(r) = (2kr)d$.

We now want to find the distance r for which such a patch will contain about one shortcut. This will allow us to consider this patch as if it was a single node in a randomly connected network. Assume that the probability for a single node to have a shortcut is Φ. To find the length for which approximately one shortcut is encountered, we need to solve for r the following equation: $(2kr)^d\Phi = 1$. The correlation length ξ defined as the distance (or linear size of a patch) for which a shortcut will be encountered with high probability is therefore,

$$\xi = \frac{1}{k\Phi^{1/d}} \tag{12}$$

Note that we have omitted the factor 2, since we are interested in the order of magnitude. Let us denote by $V(r)$ the total number of nodes reachable from a node by at most r steps, and by $a(r)$, the number of nodes added to a patch in the *r-th* step. That is, $a(r) = n(r) - n(r-1)$. Thus,

$$a(r) \sim \frac{\mathrm{d}n(r)}{\mathrm{d}r} = 2kd(2kr)^{d-1} \tag{13}$$

When a shortcut is encountered at the r step from a node, it leads to a new patch. This new patch occurs after r steps, and therefore the number of nodes reachable from its origin is $V(r-r')$. Thus, we obtain the recursive relation

$$V(r) = \sum_{r'=0}^{r} a(r')[1 + \xi^{-d}V(r-r')] \tag{14}$$

where the first term stands for the size of the original patch, and the second term is derived from the probability of hitting a shortcut, which is approximately $\xi - d$ for every new node encountered. To simplify the solution of 14, it can be approximated by a differential equation. The sum can be approximated by an integral, and then the equation can be differentiated with respect to r . For simplicity, we will concentrate here on the solution for the one-dimensional case, with $k = 1$, where $a(r) = 2$. Thus, one obtains

$$\frac{\mathrm{d}V(r)}{\mathrm{d}r} = 2[1 + V(r)/\xi] \tag{15}$$

the solution of which is:

$$V(r) = \xi\left(e^{2r/\xi} - 1\right) \tag{16}$$

For $r \ll \xi$, the exponent can be expanded in a power series, and one obtains $V(r) \sim 2r = n(r)$, as expected, since usually no shortcut is encountered. For $r\ gg\xi$, $V(r)$. An approximation for the average distance between nodes can be obtained by equating $V(r)$ from 16 to the total number of nodes, $V(r) = N$. This results in

$$r \sim \frac{\xi}{2}\ln\frac{N}{\xi} \tag{17}$$

As apparent from 17, the average distance in a small-world network behaves as the distance in a random graph with patches of size ξ behaving as the nodes of the random graph.

References

[1] Alex Bavelas. A mathematical model for group structures. *Applied Anthropology*, 7(3):16–30, 1948.

[2] Reuven Cohen and Shlomo Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, 2010.

[3] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[4] Dingqi Yang, Daqing Zhang, Vincent. W. Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015.

[5] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation, 2011.