Topological modelling of large networks

BY RAÚL J. MONDRAGÓN*

Department of Electronic Engineering, Queen Mary University of London, Mile End Road, London E1 4NS, UK

In a complex network, there is a strong interaction between the network’s topology and its functionality. A good topological network model is a practical tool as it can be used to test ‘what-if’ scenarios and it can provide predictions of the network’s evolution. Modelling the topology structure of a large network is a challenging task, since there is no agreement in the research community on which properties of the network a model should be based, or how to test its accuracy. Here we present recent results on how to model a large network, the autonomous system (AS)-Internet, using a growth model. Based on a nonlinear preferential growth model and the reproduction of the network’s rich club, the model reproduces many of the topological characteristics of the AS-Internet. We also identify a recent method to visualize the network’s topology. This visualization technique is simple and fast and can be used to understand the properties of a large complex network or as a first step to validate a network model.

Keywords: Internet; network models; visualization

1. Introduction

A network can be described as a set of nodes and links. An accurate description of how the nodes are connected via the links is important because form and functionality are closely related. Modelling large complex networks has practical applications. Models can provide realistic network scenarios for simulations and can be used to predict network evolution. As specific networks have different characteristics, we tend to avoid general topological models and select properties to model based on the problem at hand. Here we consider the Internet at the autonomous system (AS) level.

The Internet can be described at the router or AS level. At the router level, the nodes and links of the network represent physical entities. The nodes describe the routers and switches managing the passage of traffic through the network. The links represent the different physical connections between the nodes, for example, optical fibres, copper, wires, etc., and have specific directions between the endpoint nodes. At the router level, a basic topological model of the Internet should include the geographical position of the nodes and links, the capacity of the links and the direction that the Internet traffic follows. For management purposes, the Internet is divided into sub-networks. Each sub-network adheres to common routing conventions, usually the interior gateway

* r.j.mondragon@elec.qmul.ac.uk

One contribution of 16 to a Discussion Meeting Issue ‘Networks: modelling and control’. 
protocol. The management of a sub-network and its routers falls under one administrative entity called an AS that should exhibit to other ASs a coherent interior routing plan and the destinations reachable through the AS. At the AS level, the Internet can be considered in an abstract space where the relevant property is the connectivity between ASs. At this level, we tend to disregard many physical properties of the network like the geographical location of the ASs, which could be in different continents, or the direction of the links and their capacities.

For the Internet, there is considerable data describing it at the router and AS levels (Cooperative Association for Internet Data Analysis (CAIDA) 2007, http://www.caida.org). In §2, we introduce some of the topological properties obtained from the measurements that are used in the development of network models.

There are two general approaches to model a network (Bu & Towsley 2002; Dorogovtsev & Mendes 2003; Costa et al. 2007; Krioukov et al. 2007): static models (Doar 1996; Zegura et al. 1996; Calvert et al. 1997; Winick & Jamin 2002) based on random networks and dynamical models based on network growth models. The latter are considered to be the more promising as, if correct, they can describe the evolution of the network (Willinger et al. 2002; Krioukov et al. 2007). There are two kinds of dynamical models: descriptive models based on matching various topological properties of a network are used to study which topological properties give a good description of a network; and explanatory models that attempt to simulate the core principles and factors responsible for the network’s structure and evolution, in particular the router network (Willinger et al. 2002). A proper validation of all the network models is lacking due to the limited quality of the available measures (Krioukov et al. 2007). In §3, we describe a dynamical descriptive network model that reproduces many properties of the AS-Internet.

Section 4 describes a recent visualization technique that can be used to understand the topological properties of a network or as a first step to validate a network model and conclusions are given in §5.

2. Topological description of a network

The AS-Internet is described as an undirected graph \( G = (N, L) \), where \( N = \{n_i\} \) is a finite set of \( N \) nodes and \( L = \{l_i\} \) is a finite set of \( L \) links. Two nodes are neighbours if there is a link joining them. The connectivity of the nodes is described by the adjacency matrix whose \( a_{ij} \) entry is 1 if node \( n_i \) is adjacent to node \( n_j \) and 0 otherwise. For undirected graphs, \( a_{ij} = a_{ji} \). The degree \( k \) of a node is the number of neighbours that a node has, \( k_i = \sum_j a_{ij} \). The degree is the principal parameter to characterize a network. Two of the simplest properties of a network are its maximum degree \( k_{\text{max}} = \max\{k_i\}, i=1, ..., N \) and its average degree \( \langle k \rangle = \frac{\sum_{j=1}^{N} k_{ij}}{N} \).

A first step to describe and discriminate between different networks is to measure the degree distribution \( P(k) \); the fraction of nodes in the network with degree \( k \). For the AS-Internet, Faloutsos et al. (1999) found that its degree distribution decays as the power law \( P(k) \sim k^{-\gamma}, \gamma=2.1 \). This means that the majority of the nodes have few neighbours and there is a small set of nodes that
have a very large number of neighbours. Networks with a power-law decay in their degree distribution are known as scale free (Barabási & Albert 1999). This kind of decay is also present in other technological, biological and sociological networks (Dorogovtsev & Mendes 2003).

The degree distribution gives only partial information about the network structure. A better description can be obtained from the correlations between the degrees of different nodes (Pastor-Satorras et al. 2001; Newman 2002). In a finite network, this correlation is defined by the degree–degree distribution

$$P(k, k') = \frac{1}{N^2} \left( \sum_{i,j=1}^{N} \delta_{k_i, k} a_{ij} \delta_{k_j, k'} \right);$$

(2.1)

the probability that an arbitrary link connects a node of degree $k$ with a node of degree $k'$. In the equation, $\delta_{ij}$ is the Kronecker delta. The degree–degree correlation can also be described using the conditional probability that a node with degree $k$ has a neighbour with degree $k'$

$$P(k' | k) = \frac{\langle k \rangle P(k, k')}{kP(k)};$$

(2.2)

where $\sum_{k'} P(k' | k) = 1$. In scale-free networks due to the small number of nodes with high degree and the finite size of the network, it is not possible from the network’s measurements to evaluate accurately the degree–degree distribution. Hence, the structure of the network is characterized using different projections of the degree–degree correlation.

One way to characterize a network is by comparing it with a random network. To obtain a meaningful comparison, the random network is restricted such that its degree distribution $P(k)$ is the same as the network under study (Maslov & Sneppen 2002). The random network is obtained by randomly reshuffling link pairs of the original network with the restriction that the reshuffling process should not change the degree distribution. By comparing a network with its randomized version, scale-free networks can be classified into assortative, disassortative and neutral networks (Newman 2002). Social networks tend to be assortative, in which high-degree nodes prefer to attach to other high-degree nodes and low-degree nodes to low-degree nodes. Information networks (e.g. the World Wide Web and the AS-Internet) and biological networks have been classified as disassortative networks, in which high-degree nodes tend to connect with low-degree ones.

A projection of the degree–degree correlation used to describe the structure of the network is the average degree of the nearest neighbours. If $k$ is the degree of a node, then

$$k_{nn}(k) = \sum_{k' = 1}^{k_{\text{max}}} k' P(k' | k)$$

(2.3)

is its average degree of nearest neighbours (Pastor-Satorras et al. 2001). If $k_{nn}(k)$ is an increasing function of $k$ the network is assortative; if $k_{nn}(k)$ is a decreasing function of $k$ the network is disassortative.

While the AS-Internet is disassortative (Pastor-Satorras et al. 2001; Vázquez et al. 2002), this property does not describe the explicit connectivity between the high-degree nodes. The high-degree nodes are also referred to as ‘rich nodes’. If the rich nodes share many connections the set containing these nodes is known...
as the ‘rich club’ (Zhou & Mondragón 2004b). The AS-Internet has a densely interconnected rich club that plays a dominant role in the functionality of the network. Realistic models of the Internet should reproduce this hub structure. If the rich nodes are the \( r \) best-connected nodes in the network then their connection density is measured by the rich-club coefficient (Zhou & Mondragón 2004b) \( \Phi(r) = 2E_r/(r(r - 1)) \), where \( E_r \) is the number of links among the top \( r \) richest nodes and \( r(r-1)/2 \) is the maximum number of links that these nodes can share. If \( \Phi(r)=0 \) there is no club at all; if \( \Phi(r)=1 \) the members of the club form a clique. As a function of the node’s degree, the rich-club coefficient is \( \phi(k) = 2E_k/(N_k(N_k - 1)) \) (Colizza et al. 2006), where \( N_k \) is the number of nodes with degree equal to or higher than \( k \) and \( E_k \) is the number of links among these \( N_k \) nodes. The rich-club coefficient is another projection of the degree–degree distribution as its satisfies (Colizza et al. 2006)

\[
\phi(k) = \frac{2N(k) \sum_{i=k}^{k_{\text{max}}} \sum_{j=k}^{k_{\text{max}}} P(i,j)}{\left( N \sum_{i=k}^{k_{\text{max}}} P(i) \right) \left( N \sum_{i=k}^{k_{\text{max}}} P(i) - 1 \right)}.
\]

(2.4)

Alternatively, from the definition of the rich-club coefficient,

\[
\Delta L_k = \frac{1}{2} (\phi(k + 1)N_{k+1}(N_{k+1} - 1) - \phi(k)N_k(N_k - 1))
\]

(2.5)

is the number of links that have at one end a node with degree \( k \) and at the other end a node with degree \( k' \), with \( k' \geq k \). In terms of the conditional probability

\[
\Delta L_k = \left( N \sum_{i=k}^{k+1} P(i) \right) \left( \sum_{k'=k}^{k_{\text{max}}} P(k'|k) - \frac{1}{2} P(k + 1|k) \right) .
\]

(2.6)

3. Models

The main property that any model tends to reproduce is the degree distribution \( P(k) \). \( P(k) \) apart, there is no agreement on which other statistical properties an Internet model should be based (Tangmunarunkit et al. 2002; Alvarez-Hamelin et al. 2005; Rodrigues et al. 2007). In a dynamical network model, starting from a small seed network, new links and nodes are added to the network until it evolves to the specified size. The difficulty when developing a dynamical model is that these should evolve networks to match specific topological characteristics. A starting point in generating networks with power-law decay in their degree distribution is to use the Barabási–Albert (BA) model (1999). The model grows a network using a preferential growth mechanism: starting with a small random network, new links and nodes are added to the network until it evolves to the specified size. The attachment is preferential because the probability that a new node will connect to node \( i \) with degree \( k_i \) is

\[
P(i) = \frac{k_i}{\sum_j k_j}.
\]

(3.1)
As a model of the AS-Internet, the BA model has several limitations. The model generates a power-law degree distribution with exponent $\gamma = 3$ compared with $\gamma = 2.1$ obtained from the measurements. The BA model also creates networks where the value of the maximum degree $k_{\text{max}}$ is too small compared with the value measured in the Internet. Nevertheless, the BA model can be extended to obtain degree distributions with other power-law exponents (Krapivsky et al. 2001; Albert & Barabási 2002; Dorogovtsev & Mendes 2003). However, a network model based solely on the reproduction of the power-law exponent of the degree distribution has its limitations as it will not describe the Internet hierarchical structure (Newman 2002; Pastor-Satorras & Vespignani 2004) neither will it reproduce the rich-club connectivity of the AS-Internet (Zhou & Mondragón 2004a).

Equations (2.5) and (2.6) show that a model that recreates the rich-club connectivity $\phi(k)$ of the network under study will also generate a good approximation of the network’s degree–degree distribution $P(k, k')$ (Krioukov & Krapivsky 2006; Zhou & Mondragón 2007). It is possible to modify the BA model to generate networks with a rich club by adding internal links as the network evolves (Dorogovtsev & Mendes 2000; Bu & Towsley 2002; Bar et al. 2004; Caldarelli et al. 2004; Zhou & Mondragón 2004a). This addition of internal links in the model is supported by observations of the Internet evolution (Pastor-Satorras et al. 2001; Vázquez et al. 2002). Simon (1955) introduced a model based on the addition of new nodes and the addition of new links between nodes that belonged to the same class, where a class is the set of nodes with the same degree. Simon’s model generates networks with a power-law scaling in the node degree. This growth mechanism allows different growth rates for different classes of nodes and hence it can create a well-connected rich club (Bornholdt & Ebel 2001).

The new growth mechanism has two components, a new node attaches with $m$ old nodes or $m$ new links appear between old nodes. In both cases, the attachment is done using the BA growth model or a modification of the BA model. For example, Bu & Towsley (2002) used a generalized linear preference model with $P(i) = (k_i - \beta)/(\sum_j k_j - \beta)$, where $m$ and $\beta$ are parameters that are adjusted to produce the correct power-law decay. The Bu & Towsley model can create a well-connected rich club but generates networks with a maximum degree $k_{\text{max}}$ smaller than the one measured in the Internet. It is possible to increase the maximum degree $k_{\text{max}}$ produced by a model using the nonlinear preferential attachment (Dorogovtsev & Mendes 2000; Krapivsky et al. 2001)

$$P(i) = \frac{k_i^\alpha}{\sum_j k_j^\alpha}, \quad \alpha > 1.$$ (3.2)

In this case, the rich nodes get all the connections and the degree distribution $P(k)$ does not decay as a power law.

From the Internet history data, it is known that the probability that a new node links with a low-degree node follows the linear preferential attachment given by equation (3.1) (Pastor-Satorras et al. 2001; Vázquez et al. 2002), whereas high-degree nodes have a stronger ability of acquiring new links than predicted by equation (3.1) (Chen et al. 2002). Taking into account these
observations the nonlinear preferential attachment

$$P(i) = \frac{k_i^{1+\delta \ln k_i}}{\sum_j k_j^{1+\delta \ln k_j}}$$

(3.3)

has been used to model the AS-Internet (Zhou & Mondragón 2004a; Zhou 2006; Zhou et al. 2007). The function $f(x) = 1 + \delta \ln x$ was used because it describes the strength of the preferential attachment with only one parameter. The parameter $\delta$ is set such that maximum degree $k_{\text{max}}$ generated by the model matches the one observed in the Internet. This model, referred to as the positive-feedback preference (PFP) model, reproduces many of the properties measured in the Internet (figure 1). In the model, the rich club is generated by the addition of new links between old nodes. The degree distribution $P(k)$ obtained from computer simulations has a pre-asymptotic power-law decay. The exponent of the power law depends on the addition of new links between old nodes and the preferential attachment (Zhou 2006). The PFP model is one of the most successful models describing the Internet (Mahadevan et al. 2005). However, we do not know if the model can predict the future shape of the Internet. Krioukov & Krapivsky (2006) noted that the PFP model does not produce degree distributions with asymptotic power-law decay. The power law observed in the model is pre-asymptotic and it is the modelling of the rich club that is responsible for this pre-asymptotic power-law decay.

We finish this section with a cautionary note. In the Internet, the network’s connectivity can be obtained by direct probing of the network or by the routing tables. Direct probing is done by recording the nodes that are visited by a packet.
travelling the network. The routing tables describe which nodes a packet should visit when crossing the network. Both measurement methods have several limitations as they can create a skewed description of the network (Petermann & de los Rios 2004). It is also the case that so far, all the measurements of the Internet are incomplete (Krioukov et al. 2007) so any model based on these measurements will not represent the actual structure of the whole Internet.

4. Visualization

A method to visualize if a network generated by a model looks similar to the original network would be a useful tool to validate a model. As networks are very large, drawing them as a set of discs joined by lines is not helpful; the density of connections obscures the network features. Recently, a visualization of the adjacency matrix has been used to distinguish different networks (Chakrabarti et al. 2007; Guo et al. 2007). As the labelling of the nodes is arbitrary, the idea is that the label of a node should be related to its degree. Then a linear order relationship, based on the connectivity, is used to arrange the node labels (Nagle 1966). Using this labelling, the adjacency matrix is plotted as a binary diagram: white if there is no link between node $i$ and $j$; black otherwise. The binary

Figure 2. (a) The AS-Internet bitmap. (b) A BA network bitmap; the network has the same number of nodes as the AS-Internet. (c) The router Internet network; the network has 192 244 nodes. (d) The characteristic component of the router graph.
The diagram gives an overall impression of the network connectivity. The degree tuple of a node is defined as \( \{k_{i,0}, k_{i,1}, \ldots, k_{i,j}\} \), where \( k_{i,0} \) is the degree of node \( i \); \( k_{i,j}, j=1, \ldots, n \) is the degree of its neighbours; and \( j < m \) if \( k_{i,j} > k_{i,m} \). The nodes are labelled in increasing order of their degree, the node with lowest degree is labelled 1. If two nodes have the same degree the node with the highest degree neighbour will rank higher. If the highest degree of the neighbours is the same, the second highest degree neighbour is used for the ranking and so on until all the neighbours are considered.

Figure 2a–d shows the plot for the AS-Internet, a model of the AS-Internet using the BA model, the router Internet and a sketch of a characteristic component of the router Internet. The bitmaps of the networks look like a fractal structure constructed by the repetition and scaling of a characteristic component. The dimensions of the component in figure 2d are related to properties of the network in the following way. The length of its base is the number of nodes with degree \( k \) given by \( N \sum_{i=k}^{k+1} P(i) \). The height of the steps is equal to the number of nodes with degree \( k' \), again given by \( N \sum_{i=k}^{k+1} P(i) \). The tread of the \( k' \)th step is given by the number of links that have degree \( k \) at one end and degree less than \( k' \) at the other end. The density of points inside the main component is related to the conditional probability \( P(k'|k) \). The density of points in the tip of the component is related to the rich-club coefficient via \( \Delta L(k) \) (see equation (2.5)).

For networks like the Internet, it is known that the degree–degree distribution \( P(k_0, k) \) gives a very good description of the network structure (Mahadevan et al. 2006). What the bitmap diagrams show are features of the whole network that are related to the degree–degree distribution; our cognitive response is to integrate this information as patterns. Different patterns correspond to different networks with different degree–degree distribution. In figure 2, the BA network was produced to model the AS-Internet. From the scaling of the characteristic component, it is clear that the BA model does not reproduce the scaling behaviour of the AS-Internet. The density of points in the BA network is more evenly distributed than in the AS-Internet, reflecting that the BA model generates neutral networks compared with the AS-Internet which is disassortative. The router Internet shows a feature not present in the AS-Internet and BA network, a dense number of points aligned in diagonal lines. This feature is present because the network is assortative (Echenique et al. 2005), high-degree nodes tend to connect with high-degree nodes and low-degree nodes with low-degree nodes.

5. Conclusions

The starting point for a model is to describe the network’s degree distribution. This has been the approach used by many researchers to create a model of the AS-Internet. However, the extension of these models to reproduce other characteristics of the network has been erratic. In part, this is because we do not know which basic topological characteristics a model should reproduce. From our experience trying to model the AS-Internet, a starting point to obtain a good model of a scale-free network is to recreate the rich-club connectivity of the network under study. The rich-club connectivity is simple to measure and a model that recreates this connectivity will also give a good approximation to the degree–degree distribution.

Phil. Trans. R. Soc. A (2008)
The validation of a model usually means the comparison of several topological characteristics which can be time consuming. The visualization of the adjacency matrix based on a linear order relationship between the node connectivity is a good initial step in assessing the validity of a model. It is computationally fast and cheap.

The author would like to thank the EPSRC EP/C520246/1 for support.

References


