Models of Complex Networks

Kalin Vetsigian

December 9, 2001

Many biological, social, communication systems can be described as complex networks. The most studied ones are the Internet, WWW, citation networks, coauthorship networks [2], metabolic networks [6]. In the last few years, as more and more data became available, it was discovered that all of these completely different networks exhibit certain common topological properties. Probably the most striking one is the power law distribution, $P(k)$, of the number of nodes, k, to which a randomly chosen node is connected. The number of edges coming from a given node is said to be the *degree* of the node. The observed degree distributions are well fit by $P(k) \sim k^{-\gamma} \exp -k/\kappa$, where the exponent γ is usually between 1 and 3, and $\kappa \gg 10$ specifies an exponential cutoff which sometimes is present. Another surprising result is that despite the enormous size of networks such as WWW, the average distance, ℓ, between two nodes is very small, even though $\langle k \rangle \equiv \sum kP(k) \sim O(1)$. Such networks are known as *small worlds*. A third property which is in the focus of attention is *clustering*. In social networks context clustering expresses the fact that the friends of a person tend to know each other. Clustering is characterized by a *clustering coefficient*. The clustering coefficient of node i with degree k_i is given by $C_i = \frac{2E_i}{k_i(ki-1)}$, where E_i is the number of edges that exist between neighbors of i and $k_i(k_i-1)/2$ is the maximum number of possible edges among them. The clustering coefficient of a graph is the average for all nodes. A nice summary of many empirical studies is given in [1].

It is a challenge to create mathematical models that can explain the emergence of these universal properties and deduce any universal conclusions that follow from them. Currently, there are three types of approaches: random graph models, evolving networks models and small-world models. In this paper I will consider the first two.

The study of random graphs was initiated by Erdös and Rényi in the 50's and 60's. They considered a graph of N nodes in which each of the $N(N+1)/2$ possible edges is present with probability p. The greatest discovery of Erdös and Rényi was that many important properties of the graph appear quite suddenly as one varies p, i.e. for a property Q there is a critical probability $p_c(N)$ in the large N limit such that

$$\lim_{N\to\infty} P_{N,p}(Q) = \begin{cases} 0, & \text{if} \frac{p(N)}{p_c(N)} \to 0 \\ 1, & \text{if} \frac{p(N)}{p_c(N)} \to \infty. \end{cases} \tag{1}$$

For example there is a critical probability at which a giant cluster forms, i.e. a connected cluster containing a finite fraction of the nodes.

It can be trivially shown that for this random model $P(k) = \exp(-\langle k \rangle)\langle k \rangle^k / k!$, i.e. Poisson distribution. This is in sharp contrast with the observed power law behavior. This model also predict clustering coefficient $C = p$ that is orders of magnitude smaller than those observed for real networks with the same $\langle k \rangle$. ℓ is systematically underestimated but of the right order of magnitude.

After the discovery of the scale free nature of $P(k)$ people attempted to generalize the concept of random graphs by putting by hand the desired degree distributions [4]. This led to random graph theory with arbitrary degree distributions. In this theory one computes averages over an ensemble of all possible graphs with N nodes, and each graph is weighted by $\prod_{i=1}^{n} P(k_i)$, where $\{k_1, ..., k_N\}$ are the degrees of its nodes. In computations with this theory the degree distributions for the N nodes are usually considered to be independent. Strictly speaking, not all degree distributions $\{k_i\}$ are consistent with a valid graph, but as long as $\langle k \rangle \ll N$ (which is always the limit of interest), the probability of such bad sequences is negligible, and the correlations created by throwing them away are negligible.

Though this subfield is very young there are already a number of beautiful results. Molloy and Reed [8] showed that a giant cluster exists iff

$$\sum_k k(k-2)P(k) > 0, \tag{2}$$

i.e. there is a phase transition in the space of distributions $\{P(k)\}$. Another important result is that the average distance between nodes is approximately equal to

$$\ell = \frac{\ln(N/z_1)}{\ln(z_2/z_1)} + 1, \tag{3}$$

where $z_1 = \langle k \rangle$ and z_2 are the average numbers of first and second neighbors of a node. ℓ scales logarithmically with N irrespective of the degree distribution, and thus the random graph models satisfy the small world property. For $\gamma < 3$ a finite exponential cutoff should be included in order for this result to be meaningful. Comparison with real networks indicates that though the trend of $\ell \sim \ln(N)$ is right, the slope is underestimated and as a result ℓ is systematically lower than the observed values [1]. This is a manifestation of the non-random aspects of the topologies of real networks.

Random graph theory gives a way to construct networks with power law degree distributions but does not answer the important question of how these distributions emerge. Most networks of interest evolve in time and it is natural to study their dynamics. Understanding the dynamical mechanisms which determine the network growth can explain the power-law behavior and probably much more. A simple growth model [5] is the one in which we add new nodes one by one and connect them to the existing nodes by a fixed number, m, of new edges. A key idea is that in order to get a power law we mist have a *preferential attachment*, i.e. the likelihood for the new node to connect to an old

node should depend on the degree of the old node. Preferential attachment is characterized by a probability $\Pi(k)$ to attach to a node with degree k. It can be shown that if $\Pi(k) \sim k^{\alpha}$ then we get a power law degree distribution only if $\alpha = 1$ and this always gives exponent $\gamma = 3$ independent of m. $\Pi(k)$ can be measured for real networks evolving on not too slow time scale. For Internet, and the citation networks for Medline and Los Alamos archive [3] we have $\alpha \simeq 1$. However, the exponent γ is different from 3 for these networks. There are also some coauthorship and collaboration networks for which $\alpha = 0.8 \pm 0.1$ [7].

There are ways to generalize $\Pi(k)$ in order to get power law distributions with arbitrary γ [1]. For example $\Pi(k) = A + k$ leads to $\gamma = 2 + A/m$. Another possibility is to include an accelerated growth. This is inspired by the observation that the average degree $\langle k \rangle$ of WWW and Internet increases with time. $m(t) \sim t^{\theta}$ leads to scale free networks with exponent controlled by θ.

In real networks the connectivity of a node does not depend on its age alone. Correspondingly, Bianconi and Barabási [9] proposed a fitness model in which each new node has a different fitness η_j which is drawn from a probability distribution $\beta(\eta)$ and

$$\Pi(k_i) = \frac{\eta_i k_i}{\sum_j \eta_j k_j}, \tag{4}$$

where the denominator normalizes the distribution and the sum is over all nodes present. This model has the curious property that it can be mapped to a non-interacting Bose gas as described in [1]. The analog of Bose-Einstein condensation is that the fittest node acquires a finite fraction of all the edges.

There are no theoretical predictions for the diameter ℓ but simulations show that $\ell = A \ln(N - B) + C$ fits the data very well. ℓ for scale free dynamic models is greater than that predicted by Eq. 3. In general, the topology of networks created by preferential attachment models is different from those created by random models even when $P(k)$ is the same. This is because the dynamical process generates non-trivial correlations that affect all topological properties. Therefore the random and evolution models cannot be substitutes for each other.

There also no analytical results about the clustering coefficient of dynamic growing models. But simulations indicate that $C \sim N^{-3/4}$ which depends more slowly on N than $C \sim N^{-1}$ for random graphs and in turn is very different from the small world models which predict C independent of N.

An interesting finding which cannot be explained by the models discussed above is that the metabolic networks of 43 cellular organisms representing all three domains of life exhibit the same average distance between nodes [6] despite significant differences in constituents and pathways. The sizes of the networks studied varied between 200 and 800 nodes. These results are in contrast with the prediction that $\ell \sim \log N$.

Intense theoretical and empirical work on network topology just started in the last few years. As more data becomes available and analyzed probably other interesting properties of natural networks will emerge and correspondingly the attention might shift toward different characteristics and different ways to clas-

sify networks. The edges of a network are the simplest representation of interactions between components - nodes. From statistical physics it is known that there are certain features that depend mainly on the lattice type and dimensionality independent of the nature or details of the interactions. This gives a hope that the abstraction of a complex system as a network is a useful one.

As an immediate application, understanding network topology of Internet can lead to the design of new more efficient communication protocols. The currently existing protocols were optimized keeping in mind the classical random network theory of Erdös. It can also lead to better understanding of the spread of computer viruses and efficient ways to fight with them.

References

[1] Réka, A.& Barabási, A., *arXiv : cond-mat*/0106096 (2001).

[2] Newman, M., *Phys. Rev. E* **64**, 025102 (2001)

[3] Newman, M., *arXiv : cond-mat*/0104209 (2001)

[4] Newman, M., Strogatz, S. H. & Watts, D., *Phys. Rev. E* **64**, 026118 (2001)

[5] Barabási, A. & Réka, A., *Science* **286**, 509-512 (1999)

[6] Jeong, H., Tombor B., Albert, R., Oltvai, Z.& Barabási, *Nature* **407**, 651-654 (2000)

[7] Jeong, H., Néda, Z. & Barabási, A., *arXiv : cond-mat*/0104131 (2001)

[8] Molloy, M. & Reed, B., *Random Structures and Algorithms* **6**, 161 (1995)

[9] Bianconi, G. & Barabási, A., *arXiv : cond-mat*/0011029 (2000)