Statistics in World Wide Web

Guojun Zhu

The World Wide web is a ever-growing random network with some special characters. Recently research on it reveals that there are self-organization phenomenon and scale-free power law in it. Furthermore, it has been pointed out clusters are formed in it. All of these cannot be well explained by the original theory of random graph. A new model are proposed and proved to be correct in some cases. And a search strategy in WWW based on it has been tested.

The *World Wide Web (WWW)* was born from a small research project a few years ago and has grown into a vast repository of information. Furthermore, this trend seems not to cease. It is totally open: any individual or institution can create a website with any number of documents and links. Although whether two documents or sites are connected by a edge is comletely random, the macroscopic characters of WWW can be followed someway. Because of the decentralized nature of its growth, the WWW has been widely believed to lack structure and organization as a whole. Recent research, however, shows a great deal of self-organization[6]. Statistics on the WWW, which is quite different from old models, seems very interesting and challenging.

WWW can be viewed as a huge *directed graph* whose vetices are documents or sites and whose edges are links (URLs) that pointed between them.[3] Many problems, although not all, can be followed in this model. Although whether there are two points are connected by a edge is comletely random, some macroscopic topological characters, such as its connectivity $< K >$, its clustering coefficient $< C >$, its average minimum path $< d >$[5, 6], are especially interesting.

Statistics Results

Pow-law

Recently, several groups reported the distributions of several quantities, such as incoming links (URLs pointing to a certain HTML document), outgoing links (URLs found on an HTML document) of a document, pages on a site, follow the scale-free power law.[3, 4].

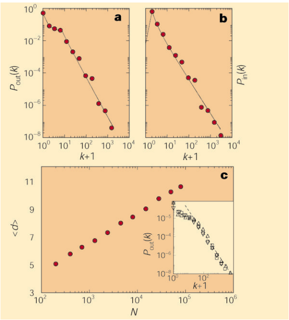

FIG. 1: Distribution of links on the WWW. **a**, Outgoing links; **b**, incoming links; **c**, the average minimum path distance between two documents[3]

In the figure 1, we can see that the tail of the distributions follows $P(k) \approx k^{-\gamma}$, with $\gamma_{out} = 2.45$ and $\gamma_{in} = 2.1$. Where the distribution of pages on the site gives a γ in $(1.6, 2)$[4].

Other

There are some other characters have been found. One is the average minimum path distance d between arbitrary two documents, which can be viewed as the magnitude of the WWW. It is reported $< d >= 0.35 + 2.06 \log(N)$, where N is the total number of the documents. In fig1 (c), we can find it. $< d >$ is tested as 11.2 for the documents inside the domain nd.edu, which has about 3×10^5 documents. The expected $< d >$ is 11.8, which is very near. And it has the definite relation with the total number of documents N[3].

Another important character is pages in the WWW appears to form clusters[5, 6]. Use the conditional probability $P_C(k'|k)$ that represents a link belonging to a node with connectivity k points to a node with connectivity k'. If it is independent of k, we are in the presence of a topology without any correlation among the nodes' connectivity, $P_C(k'|k) = P_C(k') \sim k'P(k')$. On the contrary, the explicit dependence on k is a signature of nontrivial correlations among the nodes' connectivity, and the possible presence of a hierarchical structure in its topology. $P_C(k'|k)$ is not easy to measure directly, while $< k_n n >= \sum_{k'} k'P_C(k'|k)$, the nearest neighbors average connectivity of nodes with connectivity k is tested. And the result clearly implies the existence of nontrivial correlation properties for the WWW.

BA Model

The structure of WWW is quite different from previous models of the random graph/network. Albert-László Barabśi and Réka Albert has developed a new model, which can be used to study it. Since this graph is a ever growing one. Each time step can be any of the three possible operations below. Assume we start from m_0 isolated nodes. [8]

- With probability p, we add m new links. The starting points are selected randomly and the end points of the links are selected with probability

$$\Pi(k_i) = \frac{k_i + 1}{\sum_j (k_j + 1)} \tag{1}$$

incorporating the fact that new links preferentially point to popular nodes.

- With probability q we rewire m links: For this we randomly select a node i and a link l_{ij} connected to it. Then we remove this link and replace it with a new link $l_{ij'}$ in which j' is chosen with probability $\Pi(k'_j)$ given by 1.
- With probability $1 - q - p$ we add a new node and m new links from this new node to older nodes i with probability $\Pi(k'_j)$.

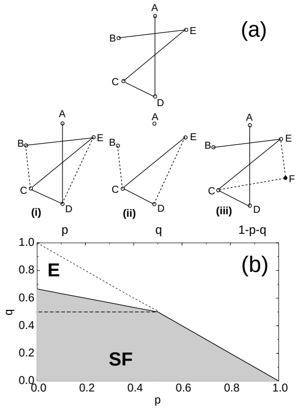

FIG. 2: **a**, Illustration of the possible elementary processes in the model, $m_0 = 3$, $m = 2$. **b**, Phase diagram. The scale-free (SF) regimes for $m = 1$ is shaded; the remaining $p + q < 1$ corresponds to the exponential (E) regime. The boundary between E and SF is shown as a dotted line when $m \to 0$, or as a dashed line when $m \to \infty$. [8]

In the SF regime,

$$P(k) \propto [k + \kappa(p, q, m)]^{-\gamma(p,q,m)} \tag{2}$$

when $k \gg \kappa$, it is approximately a power law distribution, where γ can be 2 to $infty$. When we set $p = q = 0$ and $m \to \infty$, we can get the exponent $\gamma = 3$, which was derived by a simpler model not including the rewiring.

Conclusion

The study on the statistical properties of WWW has just begun. It will benefit both the WWW and the statistics itself. We can expect more and more work will be done in the future.

[1] A. Barabási, R. Albert, *Science*, **286**, 509, (1999)
[2] S. Bornholdt, H. Ebel, *Phys. Rev. E*, **64**, 035104, (2001)
[3] R. Albert, H. Jeong, A. Barabási, *Nature*, **401**, 130, (1999)
[4] B. A. Huberman, L. A. Adamic, *Nature*, **401**, 131, (1999)
[5] R. Pastor-Satorras, A. Vázquez, A. Vespignani, *Phys. Rev. Lett.*, **87**, 258701, (2001)
[6] J. Kleinberg, S. Lawrence, *Science*, **294**, 1849, (2001)
[7] L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman, *Phys. Rev. E*, **64**, 046135, (2001)
[8] R. Albert, A. Barabási, *Phys. Rev. Lett.*, **85**, 5234, (2000)
[9] A. Barabási, R. Albert, H. Jeong, *Physica A*, **272**, 173, (1999)
[10] A. Barabási, R. Albert, H. Jeong, *Science*, **287**, 2115b, (2000)