PRAMANA
— journal of
physics

Vol. 64, No. 6
June 2005
pp. 1121–1125

Measuring information networks

K SNEPPEN[1], A TRUSINA[1,2] and M ROSVALL[1,2]
[1]NORDITA, Blegdamsvej 17, Dk 2100, Copenhagen, Denmark
[2]Department of Theoretical Physics, Umeå University, 901 87 Umeå, Sweden
E-mail: sneppen@nbi.dk

Abstract. Traffic and communication between different parts of a complex system are fundamental elements in maintaining its overall cooperativity. Because a complex system consists of many different parts, it matters where signals are transmitted. Thus signaling and traffic are in principle specific, with each message going from a unique sender to a specific recipient. In the current paper we review some measures of network topology that are related to its ability to direct specific communication.

Keywords. Networks; communication; entropy; search information.

PACS Nos 89.70.+c; 89.75.Fb; 87.80.Vt

A key feature of molecular as well as most of the other networks is that they define the channels along which information flows in a system. Thus, in a typical complex system one may say that the underlying network constrains the information horizon that each node in the network experiences [1]. This view of networks can be formalized in terms of information measures that quantify how easy it would be for a node to send a signal to other specific nodes in the rest of the network [2,3]. To do this one counts the number of bits of information required to transmit a message to a specific remote part of the network, or conversely, to predict from where a message is received (see figure 1).

In practice, imagine that you at node i want to send a message to node b in a given network (left panel in figure 1). Assume that the message follows the shortest path. That is, as we are only interested in specific signals we limit ourselves to consider only this direct communication. If the signal deviates from the shortest path, it is assumed to be lost. If there are several degenerate shortest paths, the message can be sent along any of them. For each shortest path we calculate the probability to follow this path (see figure 1). Assume that without possessing information one would chose any new link at each node along the path with equal probability. Then

$$P\{p(i,b)\} = \frac{1}{k_i} \prod_{j \in p(i,b)} \frac{1}{k_j - 1}, \tag{1}$$

where j counts all nodes on the path from a node i to the last node before the target node b is reached. The factor $k_j - 1$, instead of k_j, takes into account the

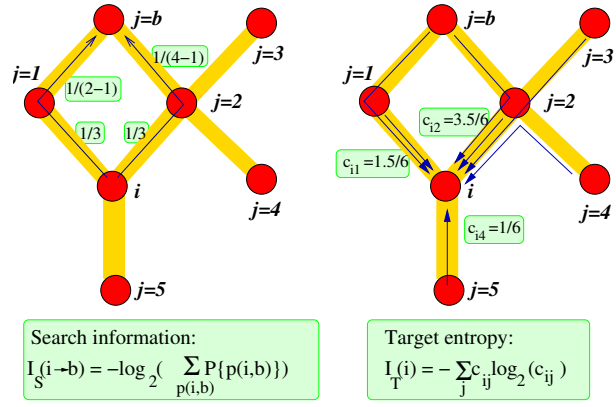

Figure 1. Information measures on network topology: Left panel: Search information $S(i \to b)$ measures your ability to locate node b from node i. $S(i \to b)$ is the number of yes/no questions needed to locate any of the shortest paths between node i and node b. For each such path $P\{p(i,b)\} = \frac{1}{k_i} \prod_j \frac{1}{k_j - 1}$, with j counting nodes on the path $p(i,b)$ until the last node before b is reached. Right panel: Target entropy T_i measures predictability of traffic to you located at node i. c_{ij} is the fraction of the messages targeted to i that passed through neighbor node j. Notice that a signal from b in the figure can go two ways, each counted with weight 0.5.

information we gain by following the path, and therefore reduces the number of exit links by one. In figure 1 we show the subsequent factors in going along any of the two shortest paths from node i to node b. The total information needed to identify one of all the degenerate paths between i and b defines the 'search information'

$$I_S(i \to b) = -\log_2 \left(\sum_{p(i,b)} P\{p(i,b)\} \right), \tag{2}$$

where the sum runs over all degenerate paths that connect i with b. A large $I_S(i \to b)$ means that one needs many yes/no questions to locate b. The existence of many degenerate paths will be reflected in a small I_S and consequently in easy goal finding.

The value of $I_S(i \to b)$ teaches us how easy it is to transmit a specific message from node i to node b. To characterize a node, or a protein in a network, one may ask how easy it is on average to send a specific message from one node to another in the net:

$$\mathcal{A}_i = \sum_b I_S(i \to b). \tag{3}$$

$\mathcal{A}$ is called the access information. In figure 2 we show $\mathcal{A}_i$ for proteins belonging to the largest connected component of the yeast protein–protein interaction network obtained by two hybrid methods [4,5]. The network shown nicely demonstrate that

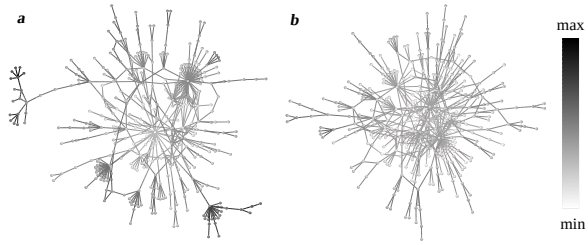

Figure 2. (**a**) Analysis of the protein–protein interaction network in yeast defined by the connected component of the most reliable data from the two-hybrid data of Ito *et al* [5]. The value of the shown access information $\mathcal{A}_i$ increases from light colored in the center to darker in the periphery. The dark colors mark nodes that have least access to the rest of the network. (**b**) shows a randomized version of the same network [6]. One sees that hubs are more interconnected and that typical $\mathcal{A}$ values are smaller (less dark).

often highly connected nodes are on the periphery of the network, and thus do not provide particularly good access to the rest of the system. This is not what one see in a randomized version of the network, where all in and out degrees are maintained, the network kept globally connected, but partners reshuffled (for a procedure, see [6,7]). In fact we quantify the overall ability for specific communication

$$\mathcal{I}_S = \sum_i \mathcal{A}_i = \sum_{i,b} I_S(i \to b) \tag{4}$$

and compare with the value $I_S(\text{random})$ obtained for a randomized network. In figure 3 we plot the Z score defined as

$$Z = \frac{\mathcal{I}_S - \langle \mathcal{I}_S(\text{random}) \rangle}{\sqrt{\langle \mathcal{I}_S(\text{random})^2 \rangle - \langle \mathcal{I}_S(\text{random}) \rangle^2}} \tag{5}$$

for the protein–protein network for both yeast (*Sacromyces Cerevisia*) [4,5] and fly (*Drosophilia*) [8] as well as for the hardwired Internet and a human network of governance (CEO) defined by company executives in the USA where two CEOs are connected by a link if they are members of the same board [9]. One sees that $\mathcal{I}_S > \mathcal{I}_S(\text{random})$ for most networks, except for the fly network. Thus most networks have a topology that tends to hide nodes. In fact this can be quantified further by considering the average information $\langle S(l) \rangle$ needed to locate a node that is a certain distance l away from a given node (the average is over all nodes and all neighbors at distance l from these nodes in a given network). For most of the investigated networks, including the yeast network, we find that $\langle S(l) \rangle - \langle S_{\text{random}}(l) \rangle$ has a minimum below zero for some rather short distance $l = 2$ to $l = 3$, whereas it becomes positive for $l > 3$. Thus most information networks have good local communication, but prefer to hide for more distant communication. We interpret this as a topology that reflects a tendency to favor specific signals, and disfavor the distant and therefore typically nonspecific signals. To understand what feature

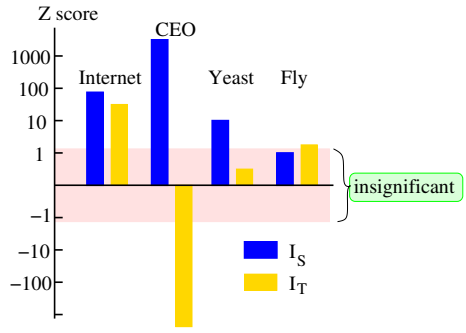

Figure 3. Measure of communication ability of various networks. A high Z-score implies relatively high entropy. In all cases we show $Z = (I - I_r)/\sigma_r$ for $I = I_S$ and I_T, by comparing with I_r for randomized networks with preserved degree distribution. σ_r is the standard deviation of the corresponding I_r, sampled over 100 realizations [2]. Results within the shaded area of two standard deviations are insignificant. All networks have a relatively high search information I_S. The network of governance CEOs show a distinct communication structure characterized by local predictability, low I_T, and global inefficiency, i.e., high I_S.

that give rise to this 'information horizon' one may recast it in terms of average connectivities as function of distances from a typical node. The observed information horizon means that the average distance from a random node to the highly connected hubs is larger in real world networks, than randomly expected.

In figure 3 we also show another quantity, namely, the ability to predict from which of your neighbors the next message to you will arrive. This quantity measures predictability, or alternatively the order/disorder of the traffic around a given node i. The predictability based on the orders that are targeted to a given node i is

$$I_T(i) = -\sum_{j=1}^{k_i} c_{ij} \log_2(c_{ij}), \tag{6}$$

where $j = 1, 2, \ldots, k_i$ denotes the links from node i to its immediate neighbors j and c_{ij} is the fraction of the messages targeted to i that passed through node j. As before our measure is implicitly assuming that all pairs of nodes communicate equally with one another.

Notice that I_T is an entropy measure, and as such is a measure of order in the network. In analogy with the global search information I_S one may also define overall predictability of a network

$$\mathcal{I}_T = \sum_i I_T(i) \tag{7}$$

and compare it with its random counterparts. In general, as the organization of a network gets more disorganized, $\mathcal{I}_T$ increases and the number of hubs with

disordered traffic increases. Also, as one considers networks with increasing values of $\mathcal{I}_T$, nodes of low degree tend to be positioned between the hubs [2].

In summary, networks are coupled to specific communication and their topology should reflect this. The optimal topology for information transfer relies on a system-specific balance between effective communication (search) and not having the individual parts being unnecessarily disturbed (hide). In figure 3 we saw that the human network of governance (CEO, [9]) were highly 'predictable', and at the same time very inefficient in transmitting information. In contrast, the hardwired Internet was found to be locally unpredictable, and therefore robust against local failures. Further, the fruit fly network, *Drosophilia melanogaster* [8], had better connections between distant parts of the network than the yeast, *Saccromyces cerevisiae* [4,5]. Such global communication patterns may reflect that the multicellular organism must sustain life in cells with many more different local environments than the single-celled yeast.

In a wider perspective the measures of ability to direct specific communication is a complement to more traditional measures of network complexity, such as node degree [10–14], particular motifs [15–17], topological hierarchy [18,19], modularity or measures associated to the nonspecific diffusion of signals across the network [20,21]. To understand real world networks, measures that relate function and topology are required. The measures I_S and I_T presented here focus on communication and as such are ideally suited for characterizing networks where information transfer is the main purpose.

References

[1] M Rosvall and K Sneppen, *Phys. Rev. Lett.* **91**, 178701 (2003)
[2] K Sneppen, A Trusina and M Rosvall, cond-mat/0407055 (2004)
[3] M Rosvall, A Trusina, P Minnhagen and K Sneppen, cond-mat/0407054 (2004)
[4] P Uetz *et al*, *Nature (London)* **403**, 623 (2000)
[5] T Ito *et al*, *Proc. Natl. Acad. Sci. USA* **98**, 4569 (2001)
[6] S Maslov and K Sneppen, *Science* **296**, 910 (2002)
[7] S Maslov, K Sneppen and A Zaliznyak, *Physica* **A333**, 529 (2004)
[8] L Giot *et al*, *Science* **302**, 1727 (2003)
[9] G F Davis and H R Greve, *Am. J. Sociology* **103**, 1 (1997)
[10] A-L Barabasi and R Albert, *Science* **286**, 509 (1999)
[11] M E J Newman, cond-mat/0303516 (2003)
[12] R V Sole and S Valverde, Complex networks, in *Lecture Notes in Physics* edited by E Ben-Naim, H Frauenfelder and Z Toroczkai (Springer, Berlin, 2004) pp. 169–190
[13] B J Kim, A Trusina, P Minnhagen and K Sneppen, nlin.AO/0403006 (2004)
[14] K Sneppen, M Rosvall, A Trusina and P Minhagen, *Europhys. Lett.* **67**, 349 (2004)
[15] D Watts and S Strogatz, *Nature (London)* **393**, 400 (1998)
[16] S S Shen-Orr, R Milo, S Mangan and U Alon, *Nature Genetics* **31**, 64 (2002)
[17] J-P Eckmann and E Moses, *Proc. Natl. Acad. Sci. USA* **99**, 5825 (2002)
[18] H Tangmunarunkit, R Govinadan, S Jamin, S Shenker and W Willinger, *Tech. Rep. 01-746* (Computer Science Department, University of Southern California, 2001)
[19] A Trusina, S Maslov, K Sneppen and P Minnhagen, *Phys. Rev. Lett.* **92**, 178702 (2004)
[20] J M Kleinberg, *Nature (London)* **406**, 845 (2000)
[21] K Eriksen, I Simonsen, S Maslov and K Sneppen, *Phys. Rev. Lett.* **90**, 148701 (2003)