Hide-and-seek on complex networks

K. Sneppen\(^1\)\(^ *(\ast)\), 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

received 15 September 2004; accepted in final form 24 December 2004
published online 11 February 2005

PACS. 89.70.+c – Information theory and communication theory.
PACS. 89.75.Fb – Structures and organization in complex systems.
PACS. 87.80.Vt – Dynamical, regulatory, and integrative biology.

Abstract. – Signaling pathways and networks determine the ability to communicate in systems ranging from living cells to human society. We investigate how the network structure constrains communication in social, man-made and biological networks. We find that human networks of governance and collaboration have predictable communication on tête-à-tête level, reflecting well-defined pathways. In contrast, communication pathways in the Internet are more distributed. For molecular networks, the communication ability in the single-celled yeast resembles the one of human networks, whereas the more complicated \textit{Drosophila} is closer to the Internet. For all investigated networks, the global communication is worse than for their random counterparts, reflecting the fact that long-distance communication is disfavored.

Information exchange between distant parts of a complex system is essential for its global functionality. For example, without the adaptability to environmental changes, maintained by communication through signaling pathways, perturbations would be fatal for living cells. Similarly, human society needs communication networks to maintain global cooperativity. In such real complex systems one should primarily focus on specific, pairwise communication between particular nodes in the network. This “constructive” signaling should be treated differently than the non-specific broadcasting in, for example, virus spreading [1] or advertisements. The speed and reliability of the information transfer is closely linked to the network architecture [2–5]. We here introduce measures of communication pathways in networks by counting the number of bits of information required to transmit a specific message to a specific remote part of the network (fig. 1a), or reversely, to predict from where a message is received (fig. 1, b and c). The introduced measures are not to be confused with the Shannon entropies [6] that have earlier been assigned to the network degree distribution [7], respectively to the long-time amplification of the dominant eigenvector of the network adjacency matrix [8].

In practice, imagine that you at node \(i\) want to send a message to node \(b\) in a given network (fig. 1a). This could, for example, correspond to sending an e-mail over the Internet. Our requirement of specific signaling is implemented by assuming that the message should follow the shortest path, or if there are several degenerate shortest paths, it should follow one of

\(\ast\) E-mail: sneppen@nbi.dk

\(\odot\) EDP Sciences
Road entropy: Target entropy: Search information:

\[ R_i = -\sum_{j=1}^{b_i} c_{ij} \log_2 c_{ij} \]

\[ T_i = -\sum_{j} c_{ij} \log_2 c_{ij} \]

\[ S(i \rightarrow b) = -\log_2 \left( \sum_{p(i,b)} \right) \]

\[ \text{Search information } S(i \rightarrow b) \text{ measures your ability to locate node } b \text{ from node } i; \text{ b) Target entropy } T_i \text{ measures predictability of traffic to you located at node } i; \text{ c) Road entropy } R_i \text{ measures predictability of traffic around } i. \]

For each such path, \[ P\{p(i,b)\} = \frac{1}{k_i} \prod_{j\in p(i,b)} \frac{1}{k_j - 1}, \]

with \( j \) counting all nodes on the path from node \( i \) to the last node before the target \( b \). When the message has entered the right exit link from this last node before the target, the goal is secured. The factor \( k_j - 1 \) instead of \( k_j \) takes into account the information we gain by following them. For each shortest path we calculate the probability to follow this path, fig. 1a, if the signal had to be transmitted without any guidance:

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

Fig. 1 – Information measures on network topology: a) Search information \( S(i \rightarrow b) \) measures your ability to locate node \( b \) from node \( i \); and c) Road entropy \( R_i \) measures predictability of traffic around \( i \). \( S(i \rightarrow b) \) is the number of yes/no questions needed to locate any of the shortest paths between node \( i \) and node \( b \).

\[ \text{Search information } S(i \rightarrow b) \text{ is the number of yes/no questions needed to locate any of the shortest paths between node } i \text{ and node } b. \]

\[ c_{ij} \text{ is the fraction of the messages targeted to } i \text{ that passed through neighbor node } j. \]

\[ b_{ij} \text{ is the fraction of messages that go through node } i \text{ which also go through neighbor node } j. \]

Fig. 2 – Hide-and-seek in complex networks. In (a-c) we show two networks obtained by (a) minimizing and (c) maximizing \( S \), while keeping the degree distribution identical to the Canadian hardwired Internet in (b). This network was selected as a typical communication network [9–11], with a broad degree distribution \( P(k) \sim k^{-1.3} \). The color of each node \( i \) shows the value \( A_i = \sum_{b} S(i \rightarrow b)/N \), that measures how easy it is to find other nodes when starting at node \( i \). In (d-f) we show same networks, but color coded according to how difficult it is to find the nodes, \( H_b = \sum_{i} S(i \rightarrow b)/N \).
the path, and therefore reduce the number of exit links by one. The above probability can be translated into an information measure, counting the number of bits of information that a message needs in order to follow the path \( p(i, b) \) exactly. Including possible degenerate paths we estimate the minimal information needed to go from \( i \) to \( b \) by the “search information”

\[
S(i \rightarrow b) = -\log_2 \left( \sum_{p(i,b)} P\{p(i,b)\} \right),
\]

(2)

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

The practical question is thus: Which position provides best access to the entire network? Surfing the Web, which web-page should be the start page when easy access to any other page is essential? The answer is the node with minimal access information, \( A_i = \sum_b S(i \rightarrow b)/N \). \( N \) is the number of nodes in the connected network. The networks in fig. 2, a to c, are color coded according to \( A_i \). Figure 2b illustrates that hubs, and often nodes directly connected to hubs, give best access to the system. Overall one can see that it is easy to access other nodes in the network in fig. 2a, whereas it is much more difficult in fig. 2c. In fact the network in fig. 2b is the Canadian Internet [12], whereas the networks in fig. 2, a and c, are obtained by rewiring the above network to, respectively, minimize and maximize \( S = \sum_i A_i/N \), while maintaining the network connected and conserving the degree of all nodes [13, 14]. We used simulated annealing to find the extremes.

Naturally, the next question is: Where is it best to hide? That is where \( H_b = \sum_i S(i \rightarrow b) \) is maximal. Note that maximizing everyone’s ability to hide \( \sum_b H_b = \sum_i A_i = S \cdot N \) is equivalent to maximizing the search information and therefore minimizing everybody’s ability to search. Thus, we illustrate the value of \( H_b \) in fig. 2, d to f, for the same networks as in fig. 2, a to c. In agreement with intuition, we indeed find that hubs are easily accessible by other nodes and thus are bad places for hiding. Rather one should hide on nodes on the periphery. Is it possible for a node to have a good access to other nodes but not be easy accessible at the same time? The compromise favors a position on a neighbor to a hub. For example, if we consider the network implementation of a city with roads as nodes and intersections as links, it is preferable with an address on a small road that connects directly to a major road/hub.

We will later see that many real-world networks are characterized by relatively high value of the overall search information \( S \) (see fig. 4). The main contribution to \( S \) comes from nodes that are separated by average or larger distances in the network, and thus a relatively large \( S \) implies that large-distance communication is disfavored. This could be due to functional, geographical or other constraints. The ability to search/hide is however not the only measure of the communication properties of a network. Another key aspect of communication handling is associated to prediction of local traffic to and across nodes in the network. This represents the “passive” aspect of information handling.

To define the predictability, let us consider messages arriving to a given node \( i \) in a network. Your task, being on node \( i \), is to guess the “active” neighbor/link from where the next message arrives. Without prior knowledge, all your local connections are equal and it would take you \( \log_2(k) \) yes/no questions to guess the active link, where \( k \) is the number of connections of your node. However, if the information about the traffic through links is available, the direction of the next message can be guessed with less questions if the search is biased towards the more used links. For simplicity, we assume that all communication takes place through the shortest paths and all nodes communicate in equal amounts with all other nodes.
The predictability, or alternatively the order/disorder of the traffic around a given node \( i \), is measured by an entropy of paths targeted to a given node \( i \), \( T_i \), and an entropy of all paths across the node, \( R_i \) (fig. 1, b and c). The predictability based on the orders targeted to \( i \) is

\[
T_i = - \sum_{j=1}^{k_i} c_{ij} \log_2(c_{ij}),
\]

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 \( j \). Similarly, we use \( b_{ij} \), defined as the fraction of messages that go through node \( i \) that also go through node \( j \), to quantify the entropy associated to traffic across node \( j \):

\[
R_i = - \sum_{j=1}^{k_i} b_{ij} \log_2(b_{ij}).
\]

Technically, \( b_{ij} \) is proportional to the betweenness [1] of the link between \( i \) and \( j \), whereas \( c_{ij} \) rather quantifies a sub-division of the network around node \( i \). We will refer to \( T_i \) as the target entropy, and to \( R_i \) as the road entropy, where a large \( T_i \) or \( R_i \) means a low predictability. In contrast, a low \( T \) or \( R \) implies that traffic around most nodes is polarized in the sense that only a few exit links are effectively used.

Figure 3 shows the values of \( T_i \) and \( R_i \) for different complex networks. In fig. 3, a to c, we examine networks by color coding the nodes according to target entropy, \( T_i \). Figure 3, d to f, shows networks color coded according to the road entropy \( R_i \). The bluish hubs reflect that traffic to highly connected nodes is hard to predict. However, this is not always the case: the location of nodes with low predictability also depends on the overall topology of the network. The networks in fig. 3 are presented so that the entropy increases from, respectively, a and d to c and f. As the networks get more disorganized, the number of hubs with disordered

Fig. 3 – Prediction of local communication. The upper panel shows networks obtained by (a) minimizing and (c) maximizing the target entropy \( T = \sum_i T_i/N \) associated to traffic to nodes in the networks, while keeping the degree distribution identical [13,14] to the original network of Autonomous Systems in Canada shown in (b). (d to f) show the networks that (d) minimize and (f) maximize the road entropy \( R = \sum_i R_i/N \) associated to traffic across nodes in the networks. In a to c the nodes are color coded according to the value of \( T_i \), while we in d to f color code according to \( R_i \).
traffic increases. Also, nodes of low degree become more confused as they tend to position themselves between the hubs. It is interesting that this positioning of low-degree nodes increases the number of alternative pathways in the system, and thus tends to minimize the search information $S$. Therefore the minimal $S$ network in fig. 2a is similar to the maximal $R = \sum_i R_i$ or $T = \sum_i T_i$ networks in fig. 3, c and f.

Whereas the maximal $T$ and $R$ networks are topologically similar, this is not the case for the minimal $T$ and $R$ networks in fig. 3, a and d. The network of minimal $T$ in fig. 3a concentrates all signaling into a simple star-like structure with hierarchical features [15]. As a consequence, nearly everybody can easily predict from where the next message will come. In contrast, minimizing $R$ results in a topology characterized by hubs on a string forming an “information super highway” (fig. 3d). A low $R$ (or $T$) means that traffic is concentrated, whereas a large $R$ (or $T$) implies a distributed and thus more robust [1, 11] traffic pattern.

We have further studied Erdős-Renyi (ER) and scale-free (SF) model networks to understand the effect of the size and degree distribution on $S$, $T$ and $R$. For all network topologies we found $S \propto \log_2(N)$, with $S/\log_2(N) \approx 1.1$ for ER and $S/\log_2(N) \approx 1.4$ for SF networks, independent of $\gamma \in [2.2, 3]$ when $P(k) \sim 1/(k_0 + k)^\gamma$ and $\langle k \rangle$ is fixed, here $= 4.6$. The “polarizability measures” $T$ and $R$ are smaller for SF compared to ER, and naturally all measures $S$, $T$ and $R$ increase with $\langle k \rangle$, when $\langle k \rangle \ll N$. Beyond degree distribution, fig. 2, c and f, illustrate that high $S$ is associated to extended or modular features of the network, and a high $T$ or $R$ (fig. 3c, f) is associated to a homogeneous topology around most of the nodes.

With this background we now, in fig. 4, compare real networks with their randomized counterparts [13], thus renormalizing for effects associated to their degree distribution. The details of the comparison are shown in table I. For each network we show the $Z$-score for $S$, $T$ and $R$. A large positive $Z$-score means that the corresponding network has relatively large entropy. For example, we see that the hardwired Internet is quite “messy” in all senses: The traffic is unpredictable, implying that the network is robust, and at the same time one needs relatively large information handling to transmit packages across the system. In contrast, the social networks, exemplified here by the network of company executives in USA, CEO [16], and the scientific collaboration network, hep-th [17], shows a pronounced pattern of high traffic predictability and large cost of locating any particular node. These features are characteristic

| Network  | $N$  | $S$  | $S_r$ | $T$  | $T_r$ | $R$   | $R_r$ |
|----------|------|------|-------|------|-------|-------|-------|
| Internet | 6474 | 16.34| 15.03(2) | 0.583 | 0.499(3) | 0.809 | 0.793(3) |
| CEO      | 6193 | 20.693 | 12.597(3) | 1.58 | 3.316(3) | 1.831 | 3.513(1) |
| hep-th   | 5835 | 19.72 | 13.48(1) | 0.847 | 1.385(5) | 1.211 | 1.668(1) |
| Yeast [18]| 921  | 13.3  | 12.5(1) | 0.38 | 0.38(1)  | 0.722 | 0.742(3) |
| Yeast [19]| 417  | 12.2  | 10.7(2) | 0.30 | 0.33(2)  | 0.662 | 0.708(6) |
| Fly      | 2915 | 14.03 | 13.96(6) | 0.56 | 0.53(1)  | 0.931 | 0.925(2) |

Table I – Networks together with their size $N$ and their $I = S$, $T$ and $R$ values. In each case we compare the measured $I$-value by comparing with $I_r$ for randomized networks with preserved degree distribution in the considered largest component. The Internet is the network of autonomous systems [12]. In the CEO network, chief executive officers are connected by links if they sit in the same board [16]. hep-th is a network of scientists connected by links if they coauthor a publication [17], yeast is the protein interaction network in Saccharomyces cerevisiae detected by the two-hybrid experiment [18, 19], and fly refers to the similar network in Drosophila melanogaster [20]. Both of these networks are pruned to include only interactions of high confidence, and in both networks we compare with their random counterparts where both bait and prey connectivity of all proteins are preserved. All results are robust to a 10% random removal of links except for the fly network which with such a pruning tends to be closer to the yeast network.
to the ordered network topologies in fig. 3a and d.

In fig. 4 we also investigate networks of physical interactions among proteins in two organisms, yeast [18, 19] and the fly [20]. Whereas the fly network is quite close to its randomized counterpart, yeast is reminiscent of the social networks. The large $S$ for yeast reflects that many of the largest hubs are positioned on the periphery of the network [13], and therefore have relatively large entropy $A_i$, see fig. 5. This tendency of hub separation reflects optimization of local communication, at the cost of global specific signaling. On the other hand the protein network of the multicellular and more advanced fly, *Drosophila melanogaster*, displays a more complicated and in fact more robust topology as witnessed by the significantly positive $Z$-scores for $T$ and $R$ entropies.

We have presented measures that quantify the ease of global search, $S$, and the predictability...
ity of local activity, $T$ and $R$, and illustrated how they characterize both the single node signaling and the overall communication set by the global topology. 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). We found that $S$ is largest for the considered real networks, smaller for the randomized versions with preserved $P(k)$, and smallest for random (ER) networks with only $\langle k \rangle$ preserved. Thus the topologies of the considered networks disfavor distant communication, both through their large-scale organization and through their broad degree distribution.

In particular, the network of corporate CEOs and scientific co-authorship were found to have “predictable” communication pathways, and at the same time to be very inefficient in transmitting information. In contrast, the communication around most nodes in the Internet is homogeneously distributed. Finally, the protein network of the fruit fly, Drosophila melanogaster, is more integrated than that of the yeast, Saccharomyces cerevisiae, and has better connections between its distant parts. This better global communication ability may reflect the fact that the multicellular organism must sustain life in cells with many more different local environments than the single-celled yeast.

∗∗∗

We thank R. Donangelo, P. Minnhagen and J. Wakeling for useful comments and revisions of the manuscript. We acknowledge the support of Swedish Research Council through Grants No. 621 2003 6290 and 629 2002 6258.

REFERENCES

[1] Newman M. E. J., SIAM Rev., 45 (2003) 167.
[2] Kleinberg J. M., Nature, 406 (2000) 845.
[3] Adamic L. A., Lukose R. M., Puniyani A. R. and Huberman B. A., Phys. Rev. E, 64 (2001) 046135.
[4] Eckmann J.-P. and Moses E., Proc. Natl. Acad. Sci. U.S.A., 99 (2002) 5825.
[5] Rosvall M. and Sneppen K., Phys. Rev. Lett., 91 (2003) 178701.
[6] Shannon C. E. and Weaver W., The Mathematical Theory of Communication (University of Illinois, Champaign, IL) 1963.
[7] Sole R. V. and Valverde S., Information Theory of Complex Networks, edited by Ben-Naim E., Frauenfelder H. and Toroczkai Z. (Springer, Berlin) 2004, p. 169.
[8] DEMETRIUS L. and MANKE T., Physica A, 346 (2005) 682.
[9] Faloutsos M., Faloutsos P. and Faloutsos C., Comput. Commun. Rev., 29 (1999) 251.
[10] Broder A., Kumar B., Maghoul F. et al., Comput. Networks, 33 (2000) 309.
[11] Albert R. and Barabasi A.-L., Rev. Mod. Phys., 74 (2002) 47.
[12] Website maintained by the NLANR Group at http://moat.nlanr.net/.
[13] Maslov S. and Sneppen K., Science, 296 (2002) 910.
[14] Maslov S., Sneppen K. and Zaliznyak A., Physica A, 333 (2004) 529.
[15] Trusina A., Maslov S., Sneppen K. and Minnhagen P., Phys. Rev. Lett., 92 (2004) 178702.
[16] Davis G. F. and Greve H. R., Am. J. Sociol., 103 (1997) 1.
[17] Newman M. E. J., Phys. Rev. E, 64 (2001) 016131.
[18] Uetz P. et al., Nature, 403 (2000) 623.
[19] Ito T., Chiba T., Ozawa R. et al., Proc. Natl. Acad. Sci. U.S.A., 98 (2001) 4569.
[20] Giot L. et al., Science, 302 (2003) 1727.