Communication Boundaries in Networks

A. Trusina,1,2,* M. Rosvall,1,2 and K. Sneppen2

1Department of Theoretical Physics, Umeå University, 901 87 Umeå, Sweden
2NORDITA, Blegdamsvej 17, DK 2100, Copenhagen, Denmark

(Received 23 September 2004; published 16 June 2005)

We investigate and quantify the interplay between topology and the ability to send specific signals in complex networks. We find that in a majority of investigated real-world networks the ability to communicate is favored by the network topology at small distances, but disfavored at larger distances. We further suggest how the ability to locate specific nodes can be improved if information associated with the overall traffic in the network is available.

Networks have proven to be intimately connected to systems composed of many different parts that each interact with only a few of the other specific parts. These specific interactions define a network where distant parts can communicate through sequences of local interactions. In this way, all parts of the complex network can be reached from other parts, but distant communication is obviously neither as easy nor as accurate as close direct communication [1]. That is, a network is de facto, a description of the limited ability to send signals in the corresponding complex system [2]. Thus the study of networks is, before anything else, a study of the relative importance or ease of communication ability at different distances. In this Letter, we therefore investigate how the topology of networks constrain the ability to send specific signals (communicate) at, respectively, short, medium, and long distances. For pedagogical reasons, we will do this by using analogies from the Internet and city structures, but emphasize that most of our methodology allows for a simple generalization to the interesting signaling networks such as the one formed by macromolecules inside a living cell.

We stress the fundamental difference between specific signaling and nonspecific broadcasting: Where the first focuses only on locating one specific node without disturbing the remaining network, the broadcasting amplifies by transferring the signals to all exit links of every node along all branching paths. Specific signaling is thus constructive communication, whereas broadcasting is of relevance for disease spreading, spam, or computer virus propagation [3,4]. The specific vs nonspecific difference is akin to the difference between a living cell with specific protein interactions and a hard boiled egg with its entangled proteins.

One can imagine various ways of searching a specific node in a network, dependent on the available information when the search is performed [5]. In the present Letter, we compare different ways to guide the search based on locating the shortest path between a source and a target in the network. Therefore we only characterize specific signaling, and quantify the search in terms of the number of questions needed to follow the shortest path to the target.

We believe that the assumption to consider only the shortest paths is valid, since in, for example, social networks the awareness about a person is related to the shortest path [1] and for the Internet because the deviation of traffic from the shortest path is only about 10% [6].

The search information of going from source $s$ to target $t$, $S(s \rightarrow t)$ [7], is the number of bits of information one needs to go from $s$ to $t$ using the shortest paths: When starting at node $s$, one has to find the right exit link, in the direction towards target $t$. The information value in knowing one particular exit channel is $\log_2(k_i)$, where $k_i$ is the degree of the source. At the subsequent node $j$ along the shortest path to the target, the number of questions is reduced to $\log_2(k_j - 1)$ since the incoming link is known. That means that the number of questions one has to ask when walking along the path from the source to the target is $S(s \rightarrow t) = -\log_2\left(\frac{1}{k_s} \prod_j \frac{1}{k_j - 1}\right)$. If there is more than one shortest path between $s$ and $t$, then

$$S(s \rightarrow t) = -\log_2\left(\sum_{\{p(s,t)\}} \frac{1}{k_s} \prod_j \frac{1}{k_j - 1}\right)$$

where the sum runs over the set $\{p(s,t)\}$ of degenerate shortest paths between $s$ and $t$; see Fig. 1. In a previous work [7], we investigated $S$ for a number of networks and found that one needs more information to orient in real than in random networks. A random network is obtained by randomizing links while conserving degree distribution. From $\Delta S(l) = \langle S(l) \rangle - \langle S_{\text{random}}(l) \rangle$ we see that essentially all the contribution to the global excess of $\Delta S = S - S_{\text{random}}$ comes from large distances $l > 3$. For some of the networks, as, for example, Internet, yeast, and fly, the $\Delta S(l)$ is even negative at short distances, which implies that these real networks are organized to optimize the search at these short distances. Thus local specificity is favored, whereas communication beyond the horizon $l = l_{\text{search}} \sim 3$ is disfavored (for a discussion re-
paths between node s and node t denotes the number of yes/no questions needed to locate any of the shortest paths between s and t. The quantity where we now weight each exit link from a node with its betweenness bj [12,13], defined as the fraction of messages that go through node l which also go through neighbor node k.

FIG. 1 (color). Information measures on network topology: (a) Search information S(s → t) measures your ability to locate node t from node s. (b) Weighted search information Sw measures your ability to locate target t from source s, when you tend to follow the traffic given by the betweenness bj. S(s → t) is the number of yes/no questions needed to locate any of the shortest paths between node s and node t. For each such path, P(p(s, t)) = \( \prod_{j=1}^{n} \frac{1}{1-p_{j;j}} \), with j counting nodes on the path p(s, t) until the last node before t. Ssw(i → j) is the similar quantity where we now weight each exit link from a node with its betweenness bj [12,13], defined as the fraction of messages that go through node l which also go through neighbor node k.

To uncover how the topology and the search information, S, are coupled, we (in Fig. 3) investigate a number of model networks. In Fig. 3(a) we see that the search is easy at distance l ∼ 3 in the structured modular network, whereas in an unstructured random network provides better search options for l > 3. The hierarchical club network, on the other hand, clearly does worse than a random network on all scales. This is a surprising and counterintuitive result, showing that hierarchies are not always optimal for search. That (club) hierarchies are used in many human organizations may thus be seen as a way to regulate and thus limit the information exchange, rather than optimize it [11].

The search information S defined above is based on a minimal approach where one at each node knows nothing about the relative importance of the neighbors. However, in, for example, social networks one often knows who is best connected to the rest of the system. This knowledge can be obtained self-consistently at any node in a network by monitoring the traffic of orders past this node. To explore how the search can be simplified by additional knowledge, we count the information needed to search if one already knows the overall traffic flow. In this case, the relative weight of exit links from a node i is given by the betweenness, \( \beta_{ij} \) [12,13], of the links from i. Using this weighting, the minimal number of binary questions needed to choose one of the shortest paths between source s and target t defines the weighted search information,

\[
S_{sw}(s \rightarrow t) = -\log_2 \left( \sum_{p(s,t)} b_{x,j=1} \prod_{j \in p(s,t)} b'_{j,j+1} \right),
\]

where j labels the node on the path p(s, t), starting at j = 1 for neighbor node to s. \( b_{j,j+1} = \sum_{k} \beta_{j,k} \sum_{j+1} \beta_{j+1,k} \) is the betweenness of the link from node with label j to node with label j + 1, divided by the sum of the betweennesses of all k links from j. \( b'_{j,j+1} \) is similarly defined, except that the normalization excludes the link to the preceding node of j on the shortest path between s and t.

To understand the difference between S and Sw we consider a city (defined through the city network where each node is a road and each link an intersection [14]). By

FIG. 2 (color). Analysis of real-world networks: city is the information city network of Malmö [14], internet is the network of autonomous systems [20], the CEO is the network of cooperative executives in the USA [21], and Fly is the protein-protein network of Drosophia Melanogaster [22]. Panel (a) compares S(l) with the search S_random(l) measured in a randomized version of the network. One observes that a search on short distances l ∼ 2−3 is relatively optimized in the real networks. In (b) we compare S with the search obtained when one uses the information associated with overall traffic in the network. The global traffic information helps search at all long distances.

FIG. 3 (color). Analysis of model networks in terms of the quantities in Fig. 2. The modular network is constructed of C modules with C nodes in each, with 0.2C connections between nodes in modules and 0.2C connections between the modules (C = \( \sqrt{N} \)). The club tree is a hierarchical network with club structure at each level. We used a version with (k) = 6 neighbors per node and with \( k \approx \log(N)/\log((k)/2) \) hierarchical levels. The scale-free network is an example of networks with broad degree distributions, here scaling as 1/k^{1.4}. In all cases, we simulate N = 5000 node networks.
orienting yourself with the strategy behind $S$, small and large roads are weighted equally. However, $S_w$ captures that large roads more often take you closer to the target than small roads. For all investigated networks one on average gains by using the weighted search strategy. However, the contribution is not homogeneously distributed over distance. As one can see in Fig. 2(b) the weighted strategy is more efficient at longer distances, $l > 3$. However, $S_w > S$ for $l = 3$, and thus it turns out to be inefficient to follow the flow when the target is nearby: by following the flow, you will nearly always overlook small roads in your neighborhood. In terms of navigating in a city, the $S_w(l) - S(l)$ shows that it pays off to follow the large roads until you are within a few turns from your end target. Then it naturally pays off to change strategy and disregard the main stream. The distance where $S_w(l) - S(l)$ becomes negative therefore defines a characteristic search horizon, $l_{\text{local}}$, at which one should switch from local to global search strategy.

We next study the relative advantage of local vs global search strategies for some model networks in Fig. 3(b). Like the real-world networks, also the model networks have $S_w > S$ at small distances and $S_w < S$ at large distances. In particular, the club tree (hierarchy) does extremely bad at short distances because there is a strong bias to go along the main flow, and one thus needs a lot of effort to locate peripheral neighbors. For a random scale-free network, on the other hand, the overall traffic very fast guides you to the center, and therefore $S_w$ is a good search strategy at nearly all distances. The scale-free network represents topologies with very broad degree distributions, and in these, one nearly always benefit by following the flow [5]. In between these two networks is the modular network, where the global flow confuses local search [$S(l < 3) > S_w(l < 3)$] but helps traffic to other modules and thus to the more distant targets. Returning to the real-world networks in Fig. 2(b), their $l_{\text{local}} \sim 2$ horizon for traffic guided search may be seen as a combination of a short $l_{\text{local-global}} \sim 1$ horizon due to their broad degree distribution [scale-free in Fig. 3(b)], and a larger $l_{\text{local-global}}$ horizon related to modular or hierarchical features.

One may ask whether the two search strategies can be combined, such that one uses local information for local search, and global traffic information for long-distance search. In terms of traffic in a city, the picture is that there are multiple types of traffic, from pedestrian to short distance targets, bicycles to intermediate distance targets, to cars for the distant targets. In accordance to this picture, we introduce the limited betweenness measures $b_{ij}(r)$ for the links $j$ around a node $i$, defined by traffic between all pairs of nodes that moves only at maximum a distance $r$ between the source and the target. Given this set of $r$ dependent traffic weights, we, in analogy with Eq. (2), define a set of search measures $S_w(r)(l)$. For $r = 1$, $S_w(1)(s \rightarrow t) = S(s \rightarrow t)$, whereas $S_w(r \rightarrow \infty)(s \rightarrow t) = S_w(s \rightarrow t)$ and thus $S_w(r)$ naturally interpolates between the nonweighted and the traffic-weighted search approaches. In Fig. 4(a) we examine $S_w(r)(l)$ for the club-hierarchy network from Fig. 3. In accordance with Fig. 3(a) we again see that the longer distances indeed are best searched by using long-distance traffic. Also the intermediate distances $l = 3–7$ are best searched by using a search weighted by traffic traveling intermediate distances $r \sim 5$, as quantified by $S_w(5)(l)$.

Figure 4(b) shows the optimal search strategy in a real network, here the information city network of Malmö [14]. Again the search efficiency is improved by adjusting the traffic horizon to the search distance. In fact, the search can be further optimized by, at each step, adjusting the traffic horizon $r$ to the remaining distance to the target. In the language of city networks, when searching a distant road, one first uses information from car traffic, but as the distance to the target becomes smaller than, say, five intersections, one instead uses bicycle and then subsequently pedestrian traffic. This overall feature of optimizing search works best when one weights the exit link from each node $j$ by the fraction of overall traffic target explicitly to $j$. Thus, the optimal search indicated by the shaded area in Fig. 4 is a search strategy, where one at each step $j$ from $s$ to $t$ bias the search according to the subpart of the traffic that is targeted at $j$, and which has a source at distance not further away than the target $t$ (see Fig. 5). Compared to the normal betweenness, the target betweenness effectively partitions the network around each node $j$, such that each exit link is weighted by the fraction of the network that it leads to. This therefore provides a more efficient guess on the direction to the rest of the network from $j$ than the normal betweenness.

![Fig. 4](color). Investigations of search strategies in a model network (a) and the information city network of Malmö (b). The figures illustrate three simple search strategies, and one optimized (shaded). $S_w(r)(l)$ is the search information based only on traffic between nodes separated by not more than $r$ steps in the network. Nodes on short distances are best found by using local search strategy ($S_w(r-1) = S$), whereas distant nodes are best searched by using information from global traffic. However, nodes at intermediate distances are best found by using traffic between nodes at intermediate distances. To optimize search, we also investigate a search strategy $S_{\text{scale-adjusted}}$, where one at each step $j$ along the path adjust the traffic horizon to the remaining distance to the target (see Fig. 5).
FIG. 5 (color). Illustration of optimal search strategy on an information city network of Gamla stan, Stockholm [14]: At each step the contribution to the traffic bias is limited by sequentially decreasing horizons (circles). The radius of each horizon reflects the node distance to the target.

Obviously the optimal search strategy can be used only if one has access to this distant-dependent traffic information. However, as in a city, such information can, for example, be quite well estimated in social networks. Consider Milgram’s famous result of a mail locating a target person in a chain of typically six acquaintances between two persons in the USA [15]. The nontrivial result of Milgram’s experiment is not that the distance between two persons is just six, since the dimension in social networks is high [16], but the fact that short paths were found in the experiment (see, for example, Ref. [17]). In terms of our optimal search strategy, Milgram’s experiment is interpreted in the following way: Every participant that receives a mail aimed to a distant target person gives this in his turn to a friend, with a chance weighted to how often this friend travels on distances up to the scale of the target distance. With such a search strategy, adjusted to the horizon to the target, the mail will find a short path to the target person with high probability (low information cost). Effectively this partitions the network into hierarchical levels [18], as in Fig. 5, a method utilized in hierarchical computer networks [19]. We speculate that humans inherently tend to use such a scale-free search strategy, and by this, facilitate robust communication on all scales ranging from within a single remote village to the whole planet. The information gain by doing so in the city of Malmö is illustrated by the difference between the black curve and the shaded area in Fig. 4(b).

In the present work, we quantified the information cost associated with the transmission of specific signals across a network. By comparing real and random networks, we found that many real-world networks have relatively good communication at distances $l \sim 2 \rightarrow 3$, whereas long-distance communication is disfavored. The presence of such a horizon suggests that networks are indeed constructed to facilitate short-scale signaling. Beyond this horizon one should use more intelligent methods to enable signal transmission, and, for example, use overall traffic flow. In particular, we find that optimal search is obtained by using the “scale invariant” strategy, where directions are selected according to the average traffic to nodes at distances similar to that of the searched target node.

We acknowledge the support of the “Models of Life” Center at NBI and from the Swedish Research Council through Grants No. 621 2003 6290 and No. 629 2002 6258.

*Electronic address: trusina@tp.umu.se
†Electronic address: http://www.nordita.dk/research/complex
[11] The hierarchical search algorithms are effective for tree-like hierarchy searched from the top where $S$ measured from the top is $\log_2(N/2)$, which is a smaller search information than for any other network of the same size.