plain

Network Entropy

An investigation of an entropy mesuare for networks based on signaling.

ANNA ANDERSSON

Master's Thesis at KTH
Supervisor: Kim Sneppen
Examiner: Erik Aurell

TRITA-FYS 2005:72
ISSN 0280-316X
ISRN KTH/FYS/05:72–SE

Abstract

A network representation has shown to be usefull when studying complex systems with a large number of connected componets, these systems can be found in a variety if different places for example in biology, in social contexts and in technical applications. As of now there are not that many measures that can quantify network properties so that different networks can be compared. One such property is the entropy of the network. In this master thesis the properties of an signaling based entropy are investigated. The signaling is assumed to be along the shortest paths in the network and this partitions the network around the nodes and the entropy reflects how homogeniesly the network is around the nodes.

By looking at the entropy average over all the nodes in the network we can see that the meassure can differentiate between different structural, topological, properties in random networks. We also analyse a number of different real networks. To better understand the meassure we investigate how the entropy is distributed and find that it reflects the positioning of the nodes.

Sammanfattning

Det dyker upp complexasystem med ett stort antal komponenter samman länkade genom ett nätverk på många ställen tillexempel i biologi, sociala sammanhang och innom teknologin. Det finns ännu inte så många övergripande mått som karakteriserar och möjliggör jämförelser av egenskaper hos olika nätverk. En sådan egenskap är nätverkets entropi. I det här examensarbetet utforskas egenskaperna hos ett entropimått som baserats på komunikation, T för target entropy". Kommunikationen antas ske via de kortase vägarna i nätverket, detta partitionerar nätverket runnt noderna och entropin baseras på hur homogent nätverket är kring de enskillda noderna.

Genom att titta på ett medelvärdet av entropin hos alla noder i nätverket ser vi att måttet kan särskilja olika strukturella, topologiska, egenskaper hos slumpgenererade nätverk. Vi analyserar ett antal nätverk hämtade fran olika verkliga samanhang. För att forstå måttet bättre tittar vi på hur entropin är fördelad och finner att den, som väntat, avspeglar vart i nätverket en nod befinner sig.

Contents

Chapter 1

Networks

The network representation of a compexsystem is a graph showing how different components in the system interact with each other. Almost any system can be represented in a network format and because network structures are fashionable at the moment many systems are represented as networks. The systems that will be considered in this work are ones where there is a flow of information between nodes. When the flow enables the whole network to perform a function that is more complex than the function of each node we say that we have a complex network. ie. a system where the components communicate to perform a function.

To get a feeling for this one can look at a company where each emplyee carries out theire individual tasks but the company could not function without a network for communication. The same is true for living organisms. For example, if a cell is exposed to some extreme condition such as a heat shock or starvation; when this happens sensors detect the change and via its regulatory network the cell reacts to the change in many different ways to maximize the chances of survival.

The aspect of networks that will be studied in this work is communication within a network. It is known that networks found in nature are far from random, and it is interesting to study if there is a connection between the structure of a network and the tasks they perform.

1.1 How to describe a network

The components of a network are the nodes and the links. Some examples of networks and what the components are, ia given in table 1.1

For many networks information can flow in both direction if there exists a link, e.g. if two computers are connected by a cable, messages can be sent in both

Network	Node	Link
City	Street	Intersection
Social	People	Relation
Biological Regulatory NE	Proteins	Binding, Chemical reaction
Internet	Computer	Cable
World Wide Web	Web page	Hyperlink

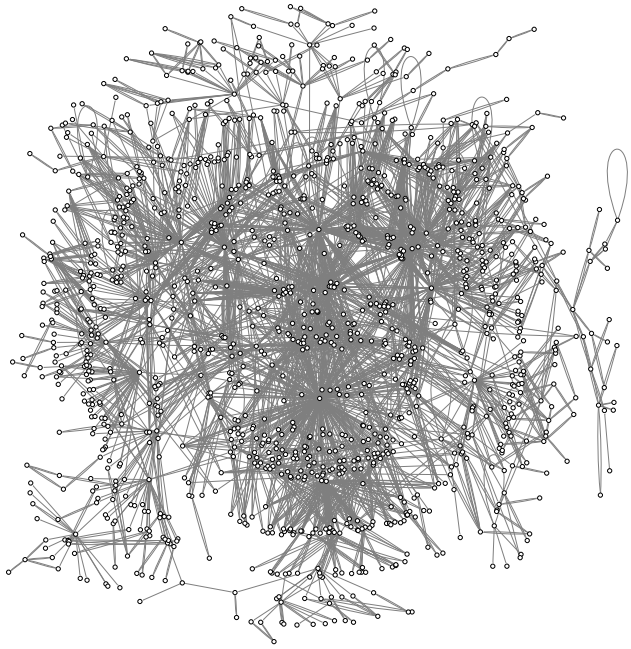


Figure 1.1: Network representation of the E-coli regulatory network. The Nodes are proteines and the links, which here are undirected, represent binding.

directions. In other networks the links are directed and information can only flow in one direction, e.g. if web page A has a link to web page B, that link can be used to go from A to B but not to go from B to A.

Connectivity

The most elementary feature of a network is how many nodes and links it has. The number of nodes in the network is the size of the network and the links shows how connected the network is. The proportion of links to nodes determines the connectivity; the average connectivity is how many link ends a node has on average.

Thus the total connectivity is twice the number of links. A node that has a high connectivity compared to the average connectivity is called a hub. For a directed graph one separates the connectivity into incoming connectivity, k_{in}, and outgoing connectivity, k_{out}.

Degree distributions

The connectivity of each node is also called the degree of that node. Looking at the degree of all the nodes in the network and how many of each degree there is we get the degree distribution. This is a is a feature of the network that can bee used to charaterize networks in to different groups.

Erdös Renyi Networks

If we are to make a network and all we know is how many nodes and links we have then it is convenient just to connect the nodes in a random manner. The degree distribution of the graph will then be a Poison distribution. There will be a few nodes with high connectivity, but most nodes will have a connectivity which is close to the average connectivity. Not even the largest hub will have a connectivity that is very much larger than the average connectivity. The properties of these graphs were studied in detail in the 1950's by Erdoš and Renyi (ER) [1] and others. However it turs out that many networks found in reality are not of this kind.

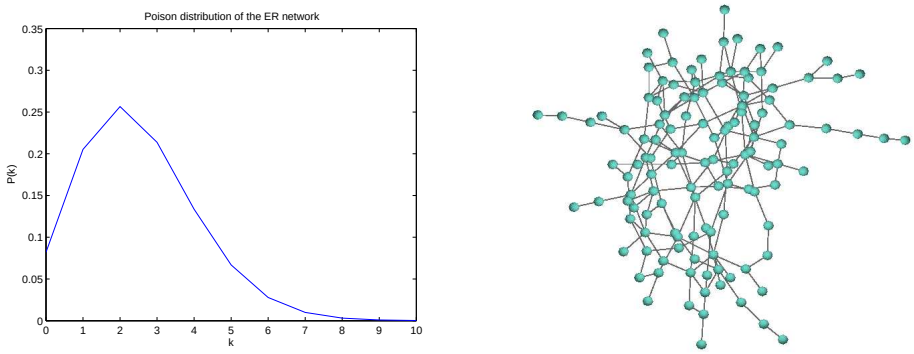


Figure 1.2: Example of an ER graph and its degree distribution. Note that all nodes have similar and low connectivity.

Scale free Networks

In contrast to ER networks, many real networks have a degree distribution that is scale free. The probability for a node to have degree k is:

$$P(k) = \frac{1}{k^{\gamma}}$$

Here γ is the exponent which determines how steep the degree distribution is.In scal free networks many have low degree and a few extremely high degree compared to the average connectivity. Figure 1.1 below shows the degree distribution for three networks. The network with $\gamma = 2.1$ has the flatest slope and the largest hub, and the network with $\gamma = 3.0$ has the steepest slope and the smallest largest hub.

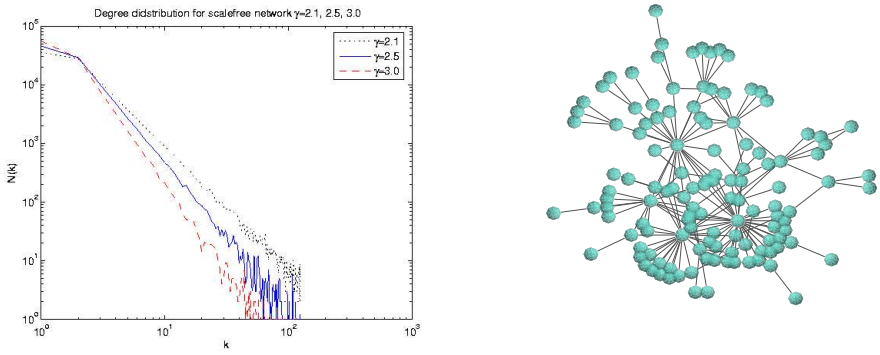


Figure 1.3: The degree distribution of three random scalefree networks with different values of γ and an example of what a scale free network can look like. The nodes have a large variation in connectivity. Note especially the large hubs, characteristic for scale free networks.

This is the part of network theory that is most studied and recently there are many models for how these networks form in reality. Some of the most famous methods are briefly described below:

Preferential Attachment: Is a model for a growing network, where the node added to the network connects to existing node with a probability that is proportional to the connectivity of that node.[2][3][4]

Merging: The basic idea is that two nodes are merged into one larger node and a new node with random links is added. There are many variants of this method. [5]

Copying models: This model was designed for protein networks and mimics the duplication of genes/proteins.[6][7]

Maximizing entropy with boundary conditions: These methods use a Shannon like entropy, $\sum_k P(k)lnP(k)$ where $P(k)$ is the probability from the degree distribution. Maximizing this entropy with specific boundary conditions gives scale free networks, while varying the boundary conditions changes the topology of the network.[8][9][10]

Scale free networks have one property that is desirable in nature, and that could explain why they are so common. The large hubs make them very robust against random attack. If something happened to a random e.g. node a mutation, or a brakedown in a computer, the probability that it will effect the rest of the network is small since most nodes are peripheral and not essential for the flow of information.

Structure

Structure is what makes a real network different from a random network. In a random network the nodes have been connected by pure chance, whereas in a real network the structure is not random, and this is important for the large scale purpose of the network. It is in the structure of real networks the most interesting questions in this field lie. Does the structure itself provide any clues as to explain how the organism, or other large system work? Are networks optimized for the flow of information? How does the structure effect the robustness? It is, ingeneral hard to pinpoint what structure is, especially at a larger scale, but I will, here introduce some structural properties that can be enforced in random networks and studied in real ones.

Degree Correlations

The degree of a node is the same as the connectivity, the number of links a node has. Degree correlations tell one thing more than the connectivity: what degrees mare at both sides of a link. The collection of all the degree correlations in a network is called the correlation profile and it shows wheter there are any patterns in how nodes of different connectivity are connected in the network.[11]

When a network is randomized all degree correlations are lost. If one wants to study something that is very much dependent on the degree correlations there are ways of randomizing the network while keeping the degree correlations, but depending on the network, the subspace of networks left to work with might be too small.

Hierarchy and Anti-hierarchy

Of course, one can manipulate the degree correlations. If high connectivity nodes are preferentially connected to high connectivity nodes we call the network hierarchical and if the opposite is true and the high connectivity nodes are connected

to low connectivity nodes we call the network anti-hirarcical [12]. Another name for this structure is associative and dissociative mixing. Many biological networks are anti-heirarcichal as where social and collaboration networks are hierarchical. Enforcing structure in this way on random networks is done to see how structure the influece the properties of networks.

Diameter

There is no consistent definitiuon of the diametre of a network. In this work the diametre of a network is the length of the longest shortest path. An other defininition that is frequently used is the average shortest path.[13]

Chapter 2

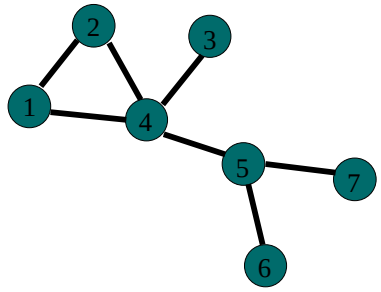
Working with networks

In the analysis of networks there are are many standard methods used to generate different kinds of networks and randomize real networks. In this section a short description of these methods will be given.

2.1 Matrix representation

It is convenient to represent a network as a matrix A where the element a_{ij} is 1 if there exists a link between the two nodes i and j and 0 if there is not a link. This matrix is called the adjacency matrix. In figure 2.1 an example of a network and the corresponding adjacency matrix [14] can be seen:

As can be seen from this example the adjacency matrix is often sparse and to make the algorithms effective the actual matrix is seldom used.



(a) The network

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

(b) The adjacency matrix corresponding to the network

2.2 Generating networks

It is often usefull to compare networks that are fully random but have different degree distributions and it is then essential to be able to generate such networks. The probability function for the degree distribution for a scale free network is given by:

$$P(k) = A\frac{1}{k^{\gamma}}$$

Where A is a normalizing constant. Calculating A and rewriting the equation in a form that gives the connectivity list, K , from a series of uniformly distributed random numbers, X , we get:

$$K = \left(\frac{1}{X}\right)^{\frac{1}{\gamma-1}}$$

Once the connectivity of each node is known and the nodes have been sorted in order of connectivity the network has to be connected. This procedure is illustrated in figure 2.2. First, the node with the highest connectivity is connected to as many other nodes as it needs, and then the second node is connected, and so forth.

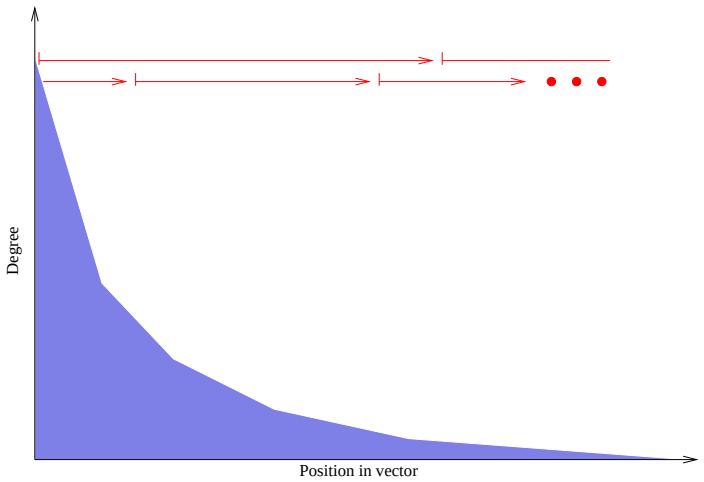


Figure 2.1: To connect the nodes with a given connectivity start with the highest connected node and connect it to the nodes with lower connectivity, when the largest node is fully connected, take the next largest node takes on where the previous stopped and connect it in the same manner, continue doing so until all nodes are fully connected.

After the network is connected it is randomized according to the algorithm given below.

For networks with an exponent γ close to three it can be hard, or even impossible, to connect the network. A trick one can use then is to introduce k_0, $P(k+k_0) = \frac{1}{(k+k_0)^\gamma}$. This shifts the degree distribution slightly to the right and induces an upper plateau, ie. , the difference in the number of nodes with degree 1 and 2 is not as large as it would have been in the simple case.

2.3 Randomizing networks, link swapping

Randomization of a network is useful when we want to compare a real network with some thing that has the same basic properties.

Randomizing keeping degree distribution

The basic idea behind the randomization is that you, at random, take two connected pairs of nodes in the network and swap their links.[15]

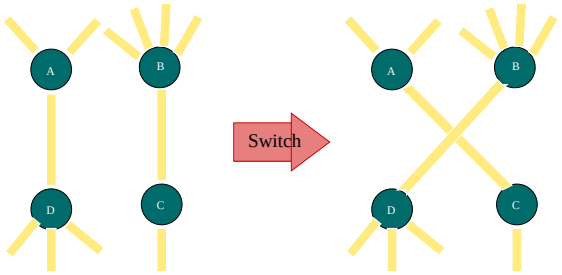


Figure 2.2: The figure illustrate how links are switched in the randomization process.

If we want to preserve the degree distribution, we choose connections to be changed by picking two links. The first link is always accepted and thus the probability to choose any link is:

$$P(l_1) = \frac{1}{L}$$

where L is the total number of links in the graph. The second link is chosen in the same way as the first but only accepted if none of the nodes are the same. This probability is:

$$P(l_2) = \frac{1}{L}\ \big(1 - P(N_A)\big)^2 \big(1 - P(N_B)\big)^2$$

Where $P(N_A)$ and $P(N_b)$ are the probability to choose either of the first pair of nodes. The probability of choosing any node is proportional to the number of links in the network and the connectivity of that node.

$$P(N_x) = \frac{k(N_x)}{2L}$$

N_x is any node and the 2 comes from the fact that the connectivity is twice the number of links.

This kind of randomization destroys any large scale structure that exists in the real network, such as degree correlations. However it conserves the degree of each node. So large scale structure is broken but local properties are conserved.

Randomizing and making an ER network

Randomizing in this way is more violent than the method described above. It destroys the original degree distribution and makes an ER-network. Thus it changes the local degree of each node. The only features it conserves is the number of nodes and links.

To speed up the randomization this algorithm is based on choosing nodes. The idea is to take a random node and choose one of its links randomly. Break that link and then choose two new random nodes and connect them. In this way links to hubs are broken more often and since the new links are formed in a fully random way the network gets an ER-distribution.

Adding structure to the network; Hierarchy and antihierarchy

The method to make a network hierarchical or antihierarchical is very similar to the randomization process where the degree distribution is kept. If the network is to become hierarchical two links are chosen according to the same principle as above then the nodes are sorted according to connectivity and the nodes with highest connectivity are connected to each other. If the network is to become antihierarchical the node with the highest connectivity is instead connected to nodes with the lowest connectivity.[12]

Maximizing/minimizing the entropy

To find a structure of the network that extremizes the entropy, I use simulated annealing and the Metropolis algorithm. Each step is done according to the link-swaping process described above. The step is always accepted if it takes the quantity in the right direction and if not then it is accepted according to a probability that is given by the function below:

$$P(accept) = \begin{cases} 1 & if\ move\ optimizes \\ e^{-\Delta E\beta} & if\ move\ not\ optimizing \end{cases}$$

Where ΔE is the quantity that we want to maximize/minimize, and β is an arbitrary number that determines how often an unfavorable move is accepted. β can be varied in different ways to improve the chanses to converge to a global maximum/minimum.

Part I

Methods

Chapter 3

Signaling

The network representation is a picture of a system seen from an information flow viewpoint. It is then desirable to study the network structure from a signaling perspective. However in most systems the actual signal paths are not known, at least not on a larger scale. If they are known then the interactions become a complex dynamical process. To be able to model the signaling on a large scale certain assumptions have to be made. The most fundamental assumption made here is that the structure of real networks is connected to the signaling in the network.

Broadcasting

The simplest way of modeling signaling is by assuming that the signal is broadcasted over the network. Then all possible paths are taken. This can easily be implemented by multiplying the adjacency matrix A by itself. Each multiplication represents one step. after two steps we have A^2 each element $(A^2)_{ij}$ represents the number of paths between i and j with pathlength 2. The sum over the elements in the squared matrix $|A^2|$ is the total number of paths of length two.

The drawback with this method is that all possible paths are taken into account, and there is no self avoidance. A signal that started in node x can after 11steps be at distance 1 from x because it has been jumping back and forth between two nodes five times.

A related broadcasting method that have introduced, [**?**] is to base the signaling on dynamic approach and use the leading eigenvalue and eigenvector of the adjacency matrix and multiply that to take a step. This approach leads to a high entropy state that, from a network perspective, looks very ordered, and therefore we are trying another method for the signaling.

3.1 The shortest path assumption

The basic assumption in this work is that the signals have a cost. This can be, for example, the fuel for a car or the cost of producing a molecule in a reaction network. This means that taking a step from one node to another along a link costs one unit. In this work the cost is only based on distance so all links cost equally much to go along.

If a second assumption is made; that there are no barriers in the network and all nodes communicate with each other, one gets specific signaling between all nodes along the shortest paths.

Comparing shortest path specific signaling (SPSS) to broadcasting, the later takes all paths into account whereas SPSS only takes the subset of paths that minimize the distance between two nodes. SPSS can be seen as the first wave in the broadcasting, the first time the signal reaches one specific node is always through the shortest path.

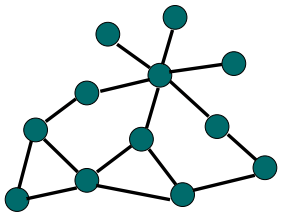


Figure 3.1: Network for which the cost of SPSS signaling is compared to the cost of broadcasting.

To illustrate how the total cost of sending signals from all nodes to all other nodes compares between the two models, this cost is calculated for the small network above.
Broadcasting: The longest path in the network is 4 and the number of nodes is 12. Assuming a unit cost and broadcasting one mesage we get:

$$Signaling\ \ cost = \sum_{n=1}^{4} |A^n| = 968$$

If every node is to send different messages to all nodes this number needs to be multiplied by the number of pairs in the the network: $(N(N-1))*$signaling cost $= (12(12-1)*968) = 127776$

The cost of sending one SPSS message between two nodes is equal to the average signaling distance in the network which is 2.8. To calculate the cost of sending different signals to all nodes we take the average signaling distance in the network

and multiply it by the number of signals sent:

$$Signaling\ \ cost = \langle shortest\ path \rangle \times N(N-1) = 2.8 \times 12(12-1) = 274$$

It is clear that broadcasting becomes very expensive compared to the SPSS. In this example the ratio of signaling cost in SPSS to broadcasting is 2.1×10^{-3}, but as the longest shortest path gets longer for larger networks the ratio becomes much smaller.

The way that we have chosen to model the signaling is based on distances in the network, and since our goal is to connect structure and function, this is a good signaling model.

3.2 Betweenness

To see how the signals are distributed on the network the Newman betweenness is used. Thies method to compute this is described in detail later, but first an introduction to the idea.

This measure uses the shortest path assumption for the signaling. Each node sends one signal to every other node in the network. That signal takes one path that has the length of the shortest distance between the nodes.

When calculating the betweeness, one takes one node at the time and lets all other nodes send one signal to that node. The procedure is repeated for all nodes. The number of signals sent is $N(N-1)$, since no node signals to itself. When taking the whole network into account one talks about the total betweenness b_i and when the betweeness is calculated for one node i , this is called the local betweenness c_i.

Newman's betweenness algorithm

Since the betweenness is a central part of this work the algorithm as it was introduced by Newman will be given in detail as described in referece[17].

1. The shortest paths to a node, j, from every other node are calculated using a breadth first search.
2. A variable c_k, taking an initial value 1, is assigned to each node, k.
3. Going through the nodes, k, in order of their distance from node j, starting with the farthest, the value of c_k is added to the corresponding variable on the the predeces or node of k. If k has more than one predecess or, then c_k is divided equally between them. This means that if there are two shortest paths between a pair of nodes, the nodes along those paths are given a betweenness of $\frac{1}{2}$ each.

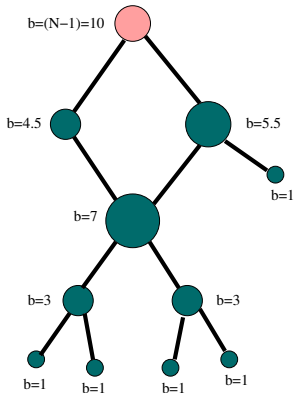


Figure 3.2: Illustration of the local betweeness. The size of the nodes is proportional to the relative betweeness of signals going to the top node.

4. When we have gone through all the nodes in this fashion the resulting values of the variables c_k represent the number of geodesic paths to node j that run through each node on the lattice, with end points of each path being counted as part of the path. To calculate betweenness for all paths, the c_k are added to a running score b_k maintained for each node an the entire calculation is repeated for each of the N possible values of j. The final running scores are precisely the betweenness of each of the N nodes.

Chapter 4

Introducing the Target Entropy

For a signaling network it is useful to have an entropy that measures how ordered the signaling is around each node. Does most of the communication go through one of the neighbors or do all neighbors contribute equally? This is an important question since this is what determines how robust the network is to directed attacks. If there are many equivalent paths it will be harder to deliberately stop the communication in the network. Here we introduce an entropy measure that is based on the signaling and measure the vulnerability of the network.

4.1 Definition

The target entropy T for a node is defined as [18]:

$$T_j = \sum_i c_i log_2 c_i$$

Where c_i is the probability that a signal arriving at node j will come through node i and shows how large a part of the network is signaling to node j through node i. This is illustrated in figure 4.1.

As an entropy, T measures how the signaling is distributed around a node. If most signals come through one or a few of the neighbors, T will have a low value, since it will be easy to predict where the signal will come from. However, if equally many signals are sent through each neighbor node, T will have a high value.In figure 4.1 the left node has a low T value and the right a high value.

The entropy T is not additive like thermodynamical entropies, since c_i is based on the structure of the network and the network is static. The structural

dependence of c_i is illustrated in figure 4.1. Changing the structure of the network slightltly, by swapping one link, T changes for both the marked nodes. In general any change in the topology will induce a change in the signaling patterns and thus in T.

This implies that an average over T does not give an entropy, instead it is an average which measures how the signaling is distributed around a node on average. The average T_i over all nodes,$i = 1...N$ will be denoted $\langle T \rangle$.

In the coming sections the properties of this measure will be investigated. Questions such as how this measure depends on size, degree distrbution, signaling distances and structural properties need to be answered.

4.2 Aims and questions

In the following sections we want to investigate what this measure shows. The primary goal is investigate what the average $\langle T \rangle$ shows in terms of detecting structure in a network. But we will also investigate where the contribution to the average comes from and thus see if it reflects the homogeneity around nodes of different connectivity. And finaly we will investigate what distances are typical in the network.

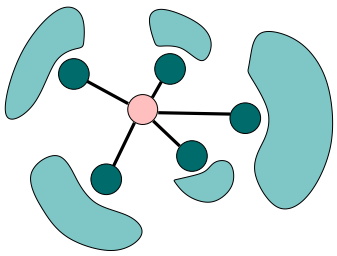


Figure 4.1: The local betweenness of the nodes around node j represents how large part of the network signals to node j through that node.

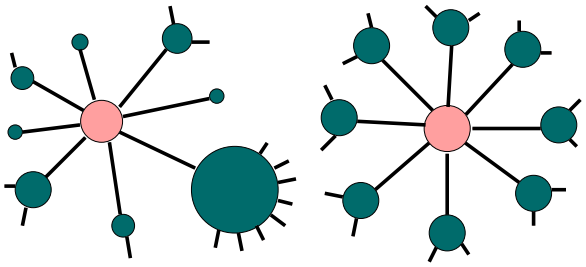


Figure 4.2: Two examples of of different situations giving different T values. The size of the nodes is proportional to the betweeness of the node. In the left network most signals come from one node and thus the entropy T has a low value. In the right example all neighbors are equivalent and the system is unpredictable and the T value is high.

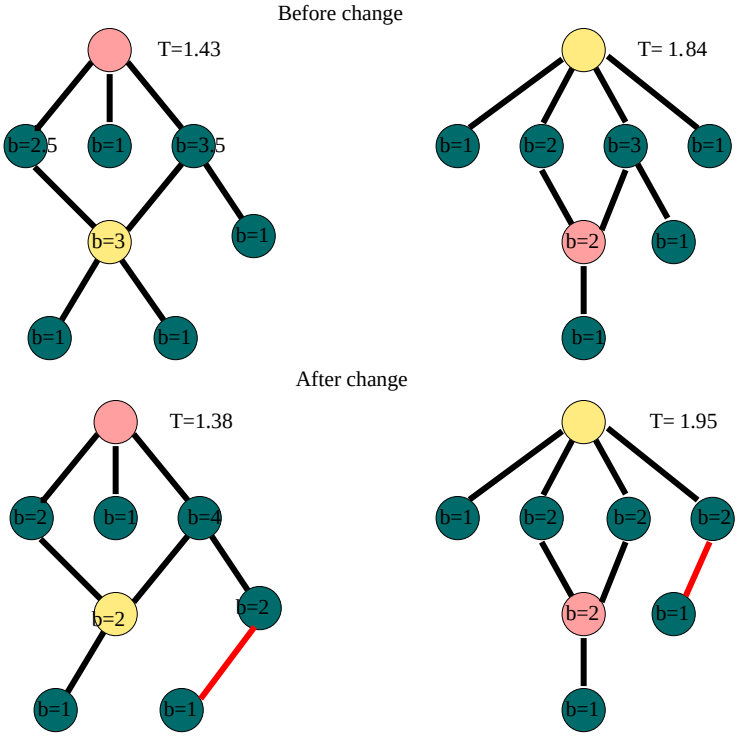


Figure 4.3: Since the all betweenness are based on the structure of the network they are dependent. This is illustrated in this figure; swapping a link, changes the signaling on the network and thus the T values.

Part II

Results

Chapter 5

$\langle T \rangle$ on Full Networks

In this Chapter we will focus on the average $\langle T \rangle$. What does the average of the entropies tell us about the structure of a network, what network properties give high/low average entropies? As T is defined, the average $\langle T \rangle$ on the full network will tell us how homogenous the network is from a signaling perspective. We will also look at some real networks and see how they can be compared to each other.

The size of the network

The first question we need to answer is how the number of nodes and links affect $\langle T \rangle$. These dependencies are shown in figure 5.1. $\langle T \rangle$ stays constant when the number of nodes is changed. The number of links, however, affects $\langle T \rangle$: an increase in the number of links leads to a more connected, and thus more homogeneous, network so $\langle T \rangle$ increases as the average connectivity increases.

The dependence on the number of links appears to be linear for the 1500 node $\gamma = 2.5$, it has a similar behaviour for networks with exponent $\gamma = 2.1$ and 3.0 From these results we can see that if we want to do a comparative analysis between random networks it is important to keep the average connectivity constant. Real networks seldom have the same average connectivity and this differences will contribute a lot to the absolute differences in $\langle T \rangle$.

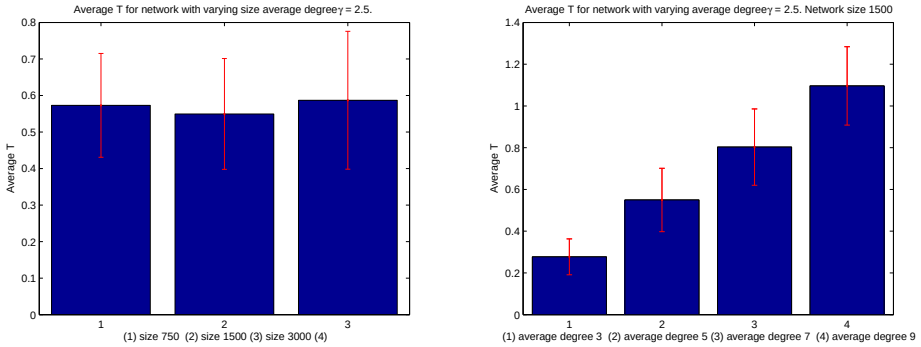


Figure 5.1: Average T is independent of network size but increases as the average degree gets larger because more links lead to a more connected and homogeneous network.

5.1 Degree Distribution Dependence

The classical way to classify networks is in terms of degree distributions. This is natural since many network properties depend on the degree distribution. The question we address in this section is: how does the degree distribution Affect $\langle T \rangle$. We have looked at scale free networks with an exponent γ between two and three since this range capture many real networks, we have also studied ER-networks since thire lack of structure should be reflected in $\langle T \rangle$. In our simulations we constructed networks with different degree distributions, but constant size and average degree. The results can be seen in figure 5.2

In figure 5.2 the most striking thing is that the ER-networks have considerably higher average $\langle T \rangle$ compared to the scale free networks, one can also note that there is hardly any variation in $\langle T \rangle$ for ER. The high average means that all the neighbors to all nodes are very similar, there are no prefered directions and thus no structure in the network. The comparitively small standard deviation indicates that different ER networks are structurally very similar.

This is an important result since an entropy should show the difference between order and disorder. ER networks are random without any restrictions and are the result of connecting nodes and links by chance and there is no order in the structure. It is therefore very nice that average $\langle T \rangle$ has its highest value for these networks.

To understand the difference between the ER and scale free networks we go back to the degree distribution 1.1. The power law distributions have a larger diversity in the connectivity and thus on average a node has neighbors

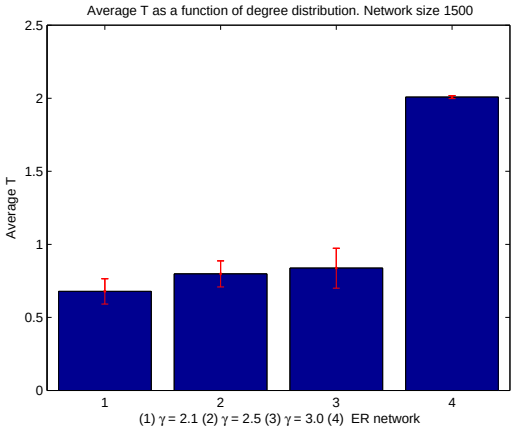


Figure 5.2: From this fugure it can be seen that $\langle T \rangle$ for the ER network is significantly higher than for the scale free networks. This reflects the homogeneity of the ER-networks compared to the scale free. Amongst the scale free networks $\langle T \rangle$ decreases with the exponent, showing that as the difference in connectivity increases the signaling becomes more predictable.

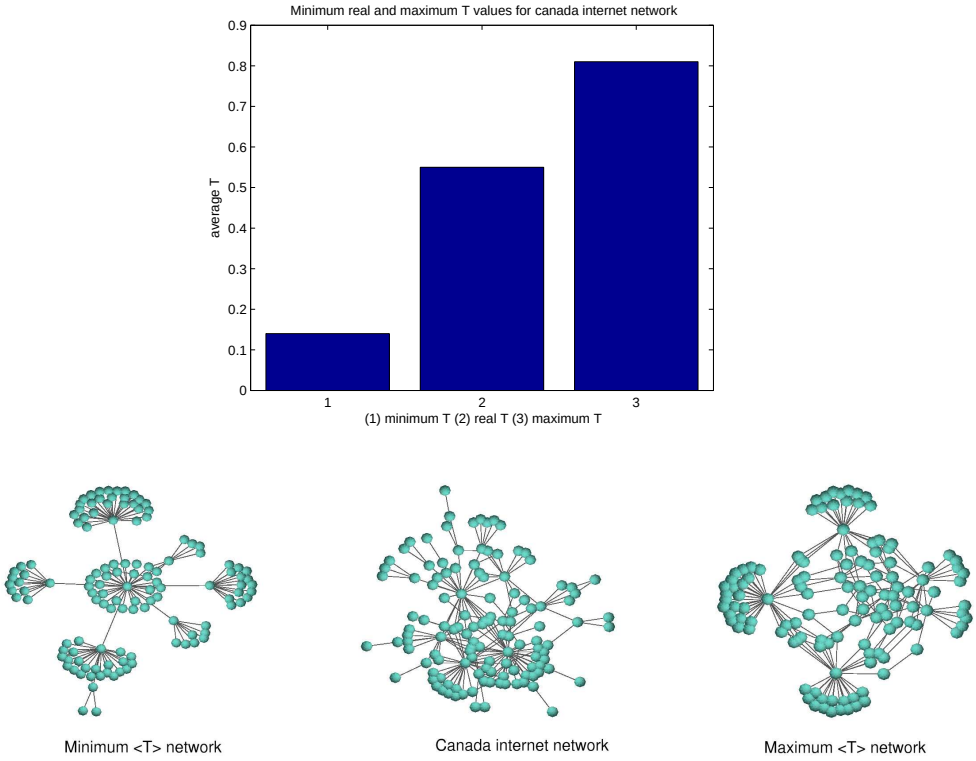
that are more different compared to the ER nodes where the diversity in the connectivity is comparatively small and thus the neighbors to a node are on average similar.

Comparing the different scale free networks, average $\langle T \rangle$ gets higher the as γ gets higher. From the degree distributions we know that the higher value γ has the steeper is the slope of the curve and thus the largest hub becomes smaller ie. the nodes are on average more similar which explains why $\langle T \rangle$ increases with γ.

One can also note that the range in $\langle T \rangle$ in which the different exponents average is not big and the error bars have large overlaps. this implies that the degree distribution has not got a large impact on the over all $\langle T \rangle$.

5.2 Manipulating $\langle T \rangle$ keeping the degree distribution fixed

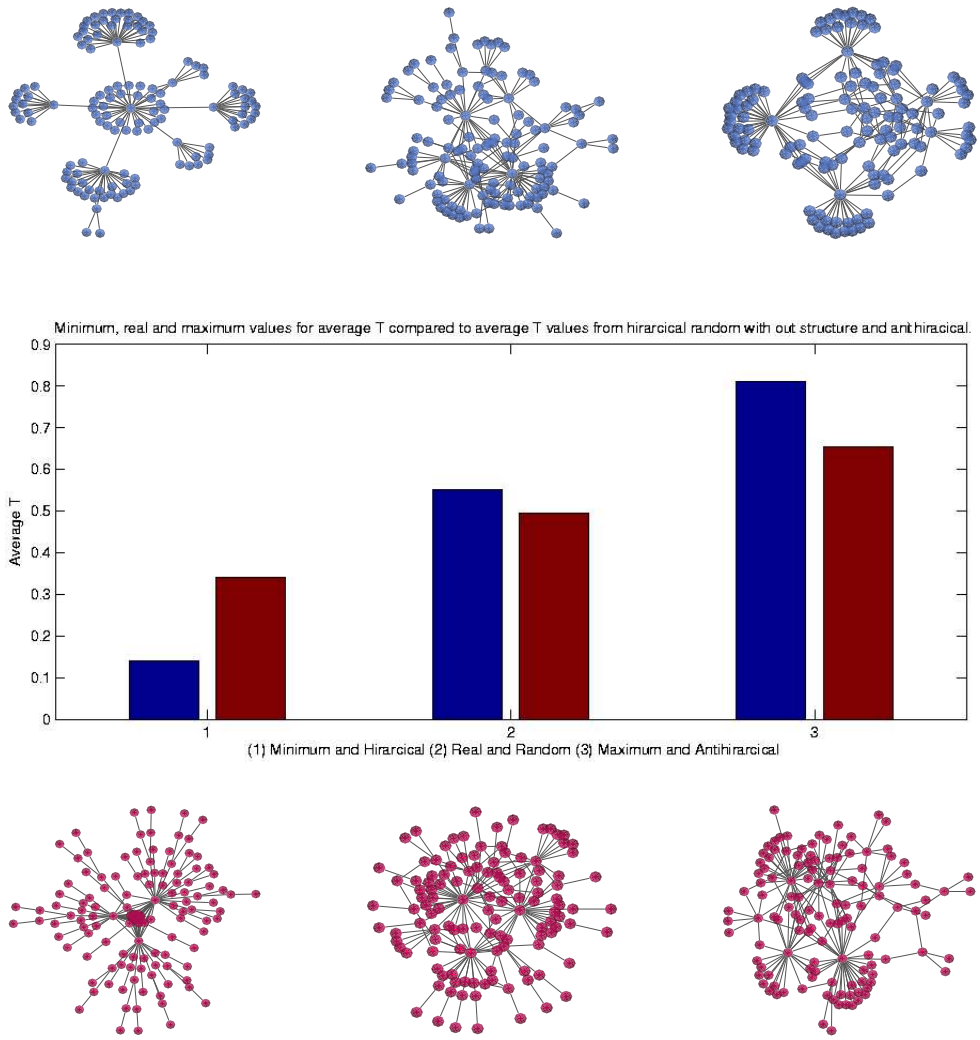
A question that was adressed in previous work [18] was: If everything is kept constant (the number of nodes, links and the degree distribution) how much can $\langle T \rangle$ vary? If you have one network how much can the $\langle T \rangle$ be changed by



rewiering links. In the article a part of the Canadian Internet was studied, the results can be seen in figure**??**.

The span of $\langle T \rangle$ values is big, thus structure in the network is captured well by $\langle T \rangle$. The network structure can be studied in figure 5.2 and from that we see that the structure that minimizes $\langle T \rangle$ is one where the hubs are connected to the largest hub. The maximizing structure repels the hubs.

Taking this one step further we investigate what kind of structural properties could increase or decrease $\langle T \rangle$. The known way to enforce structure in networks is to manipulate the degree correlations. One way to structuralize the network is to connect nodes that are either as alike as possible or connect the nodes that are as different as possible. If we base this on the connectivity of the nod it corresponds to making the network hierarchical or antihierarchical. Since we are here investigating structure it is necessary to have a structureless null model that has the same degree distribution as the original network, that is the degree distribution conserving randomized version of the original



network.

From the values of $\langle T \rangle$ we can see that hierarchy and antihierarchy do not fully explain the structure of the maximized and minimized networks. Adding hierarchy/antihierarchy does raise/lower the $\langle T \rangle$ but the minimal and maximal structures are more complex.

Comparing the networks visually it can be seen that the hierarchical network and the minimum $\langle T \rangle$ network are very different. In the hierarchical network all the hubs are connected and the diameter is larger; the network appears to have a more stringy structure. I the maximum $\langle T \rangle$ and the antihierarchical structure the hubs are at approximately the same average shortest distance from each other however in the maximum network all the intermediate nodes are placed between the hubs whereas in the antiheirarchical network they are not. The real Canada network has a higher $\langle T \rangle$ an average than the randomized versions which means that the real network is more homogeneous than the random, but it is hard to see any clear features that would explain this by looking at the network.

We can see that structure does influence the average entropies in networks, which was the aim of this investigation. However this simple manipulation of degree correlations is not enough to structurize networks to give extreme values of $\langle T \rangle$.

Combining degree distribution and degree correlations

We have seen that we have two controllable structural properties that affect $\langle T \rangle$: the degree distribution and the degree correlations. From this two questions arise:

- How is the γ dependence influenced by degree correlations?

- How are the degree correlations affected by the degree distribution?

To investigate this we work with random networks with specified degree distributions and a fixed number of nodes and links. In figure 5.2 networks with different degree distributions have been manipulated to give hierarchical structure or antihierarchical structure. From the results we can see that a hierarchical structure always decrease the average $\langle T \rangle$, whereas antihierarchy always increases $\langle T \rangle$.

In the top figure we can see that making the network anti-hierarchical changes $\langle T \rangle$ much less than making it more hierarchical. This is because $\gamma = 2.1$ networks are naturally hierarchical [ref hirarcypaper] with a few large hubs that are connected to each other and most of the network. It will be hard to find structures where the hubs are not connected and the network will never truly become antihirarcical. Thus the structure does not change much from the random and this can be seen in $\langle T \rangle$.

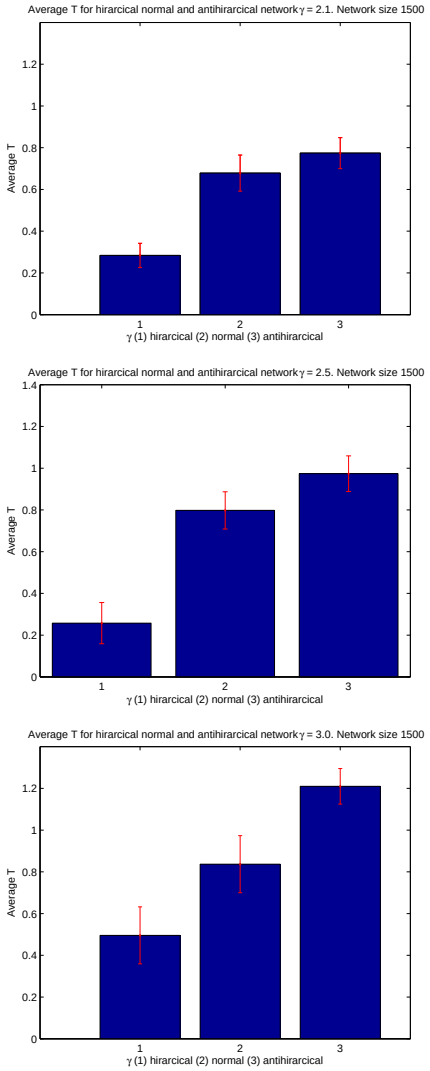


Figure 5.3: Adding structure to the network by adding a hierarchy, i,e., highly connected nodes being linked to highly connected nodes, lowers the average value of T.

The trends in the behavior of $\langle T \rangle$ described above is true also for the $\gamma = 2.5$ network. In the $\gamma = 3.0$ network these trends have disappeared and degree correlation manipulation influences $\langle T \rangle$ equally in positive and negative

directions. The $\gamma = 3.0$ network is naturally antihierarchical but the signaling structure can still be made more homogeneous by rearranging the degree correlations.

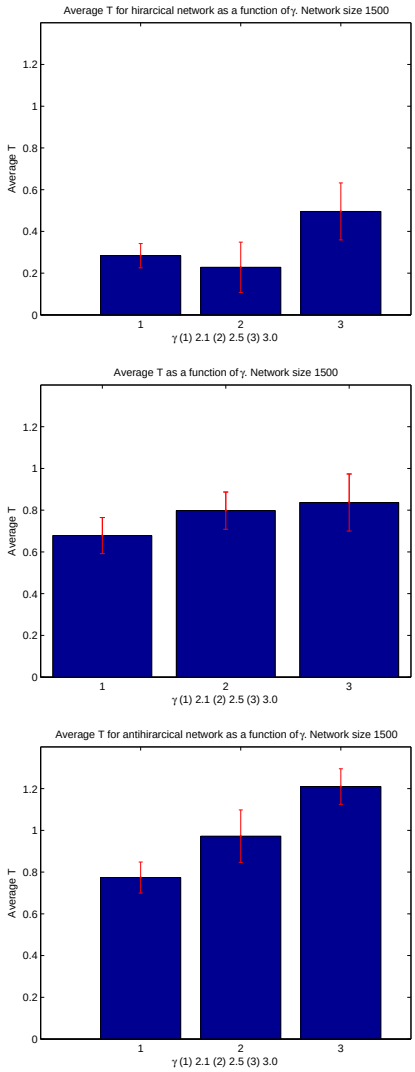


Figure 5.4:

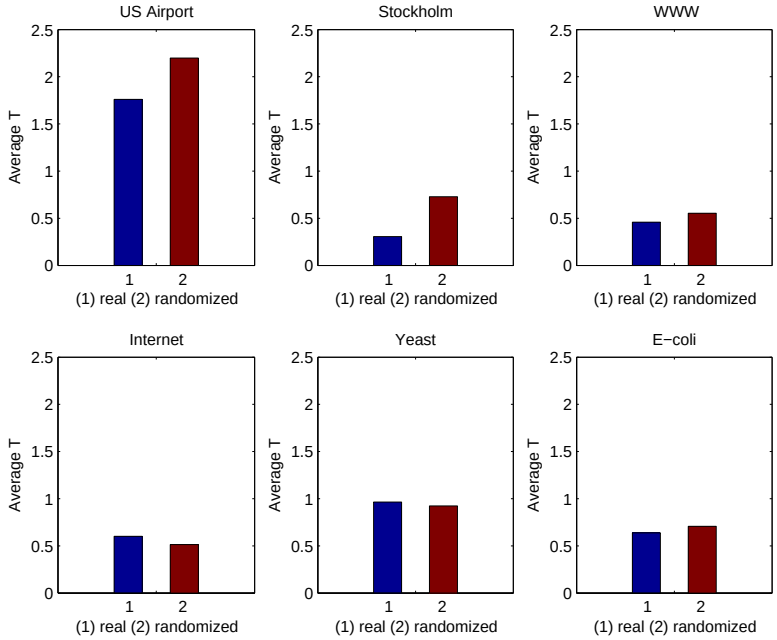


Figure 5.5: Degree distribution and average T for real and random networks.

Figure 5.2 shows how $\langle T \rangle$ depends on the exponent γ if structure is added. The first thing we once again can conclude is that structure affects $\langle T \rangle$. In the top figure we can see that the degree distribution influences $\langle T \rangle$ in a noticeable way. The $\gamma = 2.1$ networks have higher $\langle T \rangle$ than the $\gamma = 2.5$ networks. This is an effect of the rigidness of the $\gamma = 2.1$ networks; the big hubs limit what structures are possible.

5.3 $\langle T \rangle$ for real networks

To get some perspective of what the $\langle T \rangle$ measure captures we have also studied some real networks. These were chosen to give a fairly wide spectrum of networks. From the averages in figure 5.5 we see that the US airport network has a $\langle T \rangle$ value that is significantly higher than other values. This is the network that the by far the highest connectivity. From 5.1 we know that the average degree strongly influence the value of $\langle T \rangle$ and that is what we are seeing.

Network	Nodes	$\langle k \rangle$	γ	$\langle T \rangle$	$\langle T_{rand} \rangle$
US airport	332	12.8	-	1.76	2.2
Stockholm	3325	3.1	-	0.31	0.73
WWW	10000	4.2	2.4	0.46	0.55
Internet	6474	3.8	2.2	0.60	0.51
Yeast	848	4.2	-	0.97	0.92
E-coli	1522	4.6	2.4	0.64	0.71

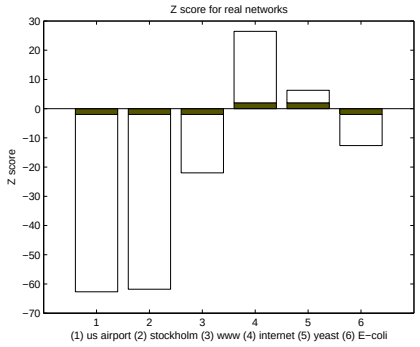


Figure 5.6: Z score for the real networks. The darker area mark two standard deviations and shows that the networks studied are significantly different from random networks.

The value of $\langle T \rangle$ tells how homogeneous the network is, but to determine how well $\langle T \rangle$ captures the structure in the network and to compare how structured these networks are we need to use the Z-score.

$$Z = \frac{\langle T \rangle - \langle T_{rand} \rangle}{\sigma_{rand}}$$

The results are displayed in figure 5.6. A positive Z-score means that the randomization makes the network more homogeneous. Likewise a negative Z-score means that the randomization make the networks less homogeneous. The magnitude of the Z-score reflects how different the real network is from the random, how much structure there is in the real network, in units of standard deviations in the random networks. The level of significance corresponding to two standard deviations is shaded in the figure. Using this as the significance level all we find that the differences between real and random are statistically significant. The real networks have structual properties that make them either more homogeneous or less homogeneous compared to the random.

Only three of these networks can be said to be scalefree and their exponents are given in table 5.3. The US-airport network has a degree distribution that is exponential and the Stockholm and Yeast networks have not got hubs that are large enough to categorize them as scale free. Thus from these data no correlation between the degree distribution and $\langle T \rangle$ or the Z-score can be seen, any such effect is taken out by the structure and average connectivity.

From these results we can see that the $\langle T \rangle$ measure captures structure in real networks and enables us to compare their homogeneity and structuredness.

Chapter 6

Distribution of T_i

In the previous section we looked at how the average value $\langle T \rangle$ for a network captured the structure of the full network. In this part we will see how nodes of different connectivity contribute to that average; the measure mainly capturing the homogeneity around the high connectivity nodes or low connectivity nodes. In this section we will only look at random generated networks since we are not here trying to capture structure in the network, but rather trying to understand how the degree distribution affects the contribution to $\langle T \rangle$.

In figure 6.1 we have plotted the T values of each node in a 900 node $\gamma = 2.5$ network. Also plotted is the theoretical upper bound of T $log_2(k)$. This figure illustrates the T values of all different nodes in some different random networks of size 900. For reference the theoretical upper limit is plotted with circles.

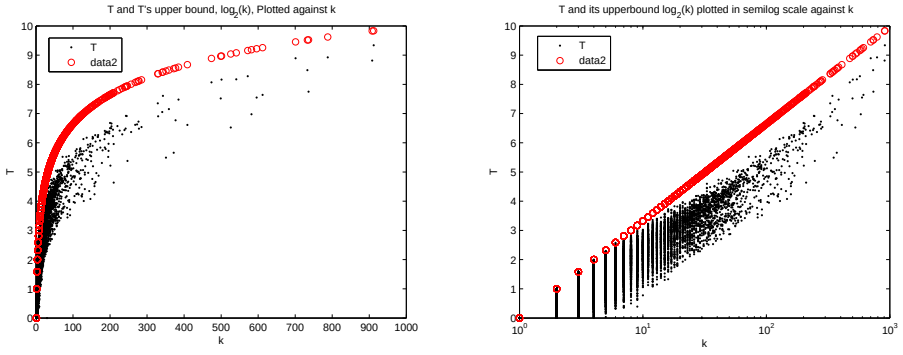


Figure 6.1: Contribution to $\langle T \rangle$ as a function of degree on normal and semi log scale.

These plots show the strong dependce of T_i on connectivity; the upper bound for low connectivity nodes is lower than that of high connectivity nodes. The nodes with intermediate connectivity in this random network(nodes of connectivity $\approx 11 - 200$) are further from the maximum values than the low connectivity nodes and the largest hubs.

In the log-log plot in figure 6.2a we can see that the main contribution to $\langle T \rangle$ comes from low connectivity nodes. This becomes even clearer in figure 6.2b which shows the cumulative contribution to $\langle T \rangle$. It can be seen that half of the contribution to $\langle T \rangle$ comes from nodes of connectivity 9 or lower.
In this investigation we have looked at networks with different $\gamma(2.1, 2.5, 3.0)$

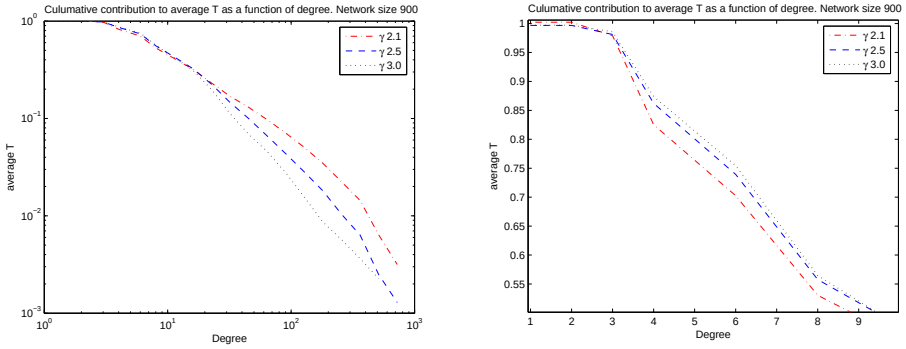


Figure 6.2: Half of the contribution to $\langle T \rangle$ comes from nodes that have connectivity 9 or lower

where the average degree has been held constant. As seen in figure 6.2 show have similar behavior.

This investigation concludes that the connectivity of a node is a primary component in T for a node. In scale free networks the large number of low connectivity nodes make their contribution to the average $\langle T \rangle$ significant. This makes $\langle T \rangle$ a rather democratic measure that captures both the homogeneity around high and low connectivity nodes.

Chapter 7

What is a typical distance in a network?

How much of the network is it necessary to take into account to get a good picture of the network? Or, at what distance from a node can the rest of the network be assumed to be mean-field or silent? And how large a fraction of the network does that correspond to? Do different networks have different typical distances and what sets theis distance.

7.1 Method for distance calculations

To investigate this we have taken all the nodes at distance one and let them signal, this obviously gives a maximum $log_2 k$ for each node thus a on the full networks scale maximum $\langle T \rangle$. In the next step the nodes at distance one and two signal, the procedure is continued until all nodes have been reached. At each step the $\langle T \rangle$ for that part of the network is calculated. This is calculated for all nodes in the network. To see how nodes of different degree behave the connectivities were divided into bins having roughly the same number of nodes in each bin.

7.2 Typical distance

The average over the generated random networks show that these networks have the expected maximum $\langle T \rangle$ at signaling distance one. There is a minimum at adistance two and at distance three the networks stabilize at their final level. This behavior indicates that diversity in the networks occurs at

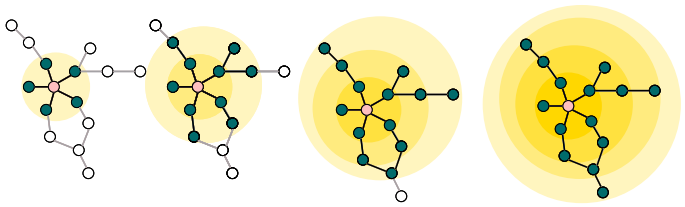


Figure 7.1: The figure illustrates the method for distance calculations. In the first step only the nearest neighbors signal, giving $T = log_2 k$, and taking the full network into account a maximum $\langle T \rangle$ for the network. In the following steps the signaling distance is increased by one until the full network is signaling.

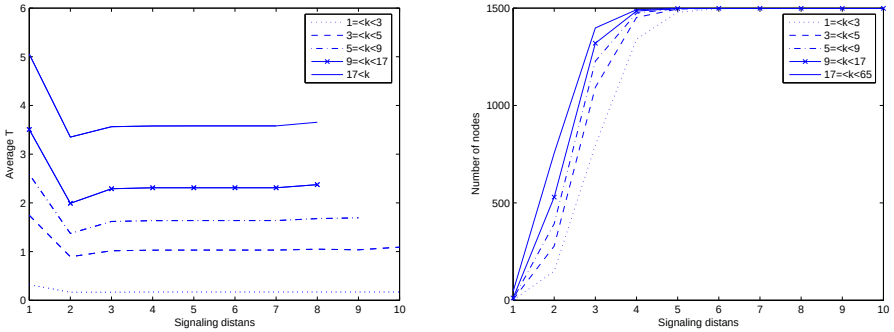


Figure 7.2: The random network has 1500 nodes an average connectivity of 5 and $\gamma = 2.5$

distance smaller than three and that distances larger than three do not influence the value of $\langle T \rangle$. Three is not a particularly large distance in these networks where the (diameter is around 10) and it could be argued that T_i values at larger distances than three could be considered to be independent, and thus additive. However, looking at 7.2 we can see that at distance 3 most of the network has been reached, and thus the fraction of the nodes that have independent T values is very small.

Comparing the behavior of nodes of different degree we can see that the lower the degree of the nodes the closer is the final value to the maximum. This is something that could alrady be seen in figure 6.1.

The E-coli network is in many respects similar to the generated random networks: the number of nodes is 1522, average connectivity is 4.6 and $\gamma = 2.4$.

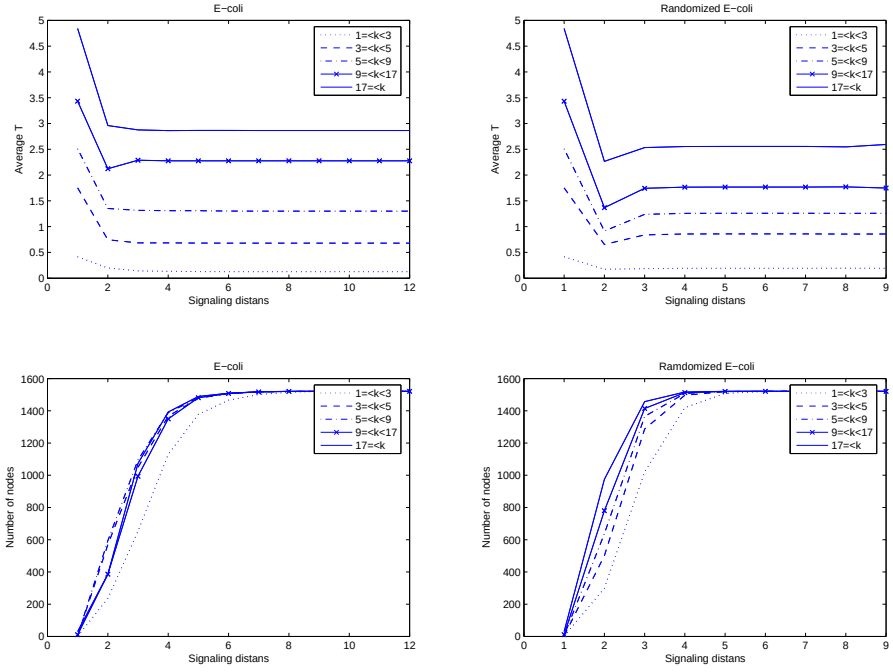


Figure 7.3: Real and random average T and number of nodes reached for the E-coli network

Therefor there is a big similarity between the generated networks and the randomized is therefor big. Other randomized networks such as the internet and the us-airport network also have this behavior in their randomized versions so this seems to be a general result. It probably reflects the fact that most nodes in a random network are connected to a hub and at signaling distance two the effect of the hub sets in since all other nodes that are connected to that hub start to signal.

The real networks are somewhat different. There is in general not a clear minimum at distance two. It is logical that there is a minimum at distance since at that distance the connectivity of the nearest neighbors of each node influence T. In a random network the nearest neighbors are of different connectivity and thus generating a minimum in the entropy. It is interesting to see is that in the real networks this is not the case, there is no minima thus implying that there is a homogeneity in the connectivity of the nearest

neighbors.

In the figure displaying the number of nodes reached at a certain distance we can see that the distance at which most of the network has been reached is, larger than the random case, where it is five or six. Further, the intermediate connectivity nodes are behaving more like high connectivity nodes giving a clearly distinguished form the nodes with the lowest connectivity.

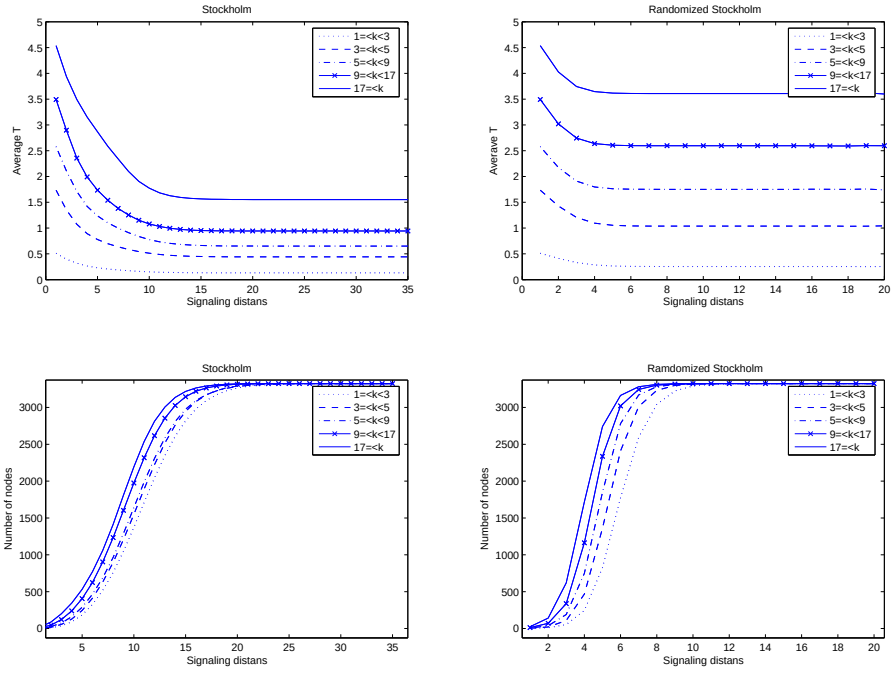


Figure 7.4: T has been calculated for different signaling distances, the signaling distance has been taken as how far out you go from the node. At distance one T will be $log_2(k)$. From this analysis one gets a profile of order in the network around the network. comparing the real network and the random we See that the in the real network the nodes of different degree are more alike compared to the nodes in the randomized. The longest distance in the network decreases when the network is randomized. It can also be seen that the distance out from the node before the final value is reached is larger in the random compared to the real.

The city networks, and to some extent the Yeast and internet networks, have clearly different behavior, the distances are larger and the T_i values of the hubs decrease much more resulting in that the values at which $\langle T \rangle$ saturates are

more similar, there is a smaller difference between hubs and low connectivity nodes compared to the other networks. The explanation for the characteristic behavior of the city's can be under stood if we look at the network of Stockholm. Cities are special in the sense that they have have a two dimentional lay out and are divided in to clear districts which make them modular. In Stockholm this is made even clearer since the city is built on islands.

It is the modularity and that the modules are connected by low connevtivity nodes that decreases the effect of the connectivity. Out side of the module of the node its connectivity does not really matter. being highly connected only matters until the signal reaches the next module, then it is the connectivity of the node connecting the modules that matters, and in that respect all nodes within the module are similar. Looking at the distance at which $\langle T \rangle$ saturates and at the average size of the modules we can see that this correlates.

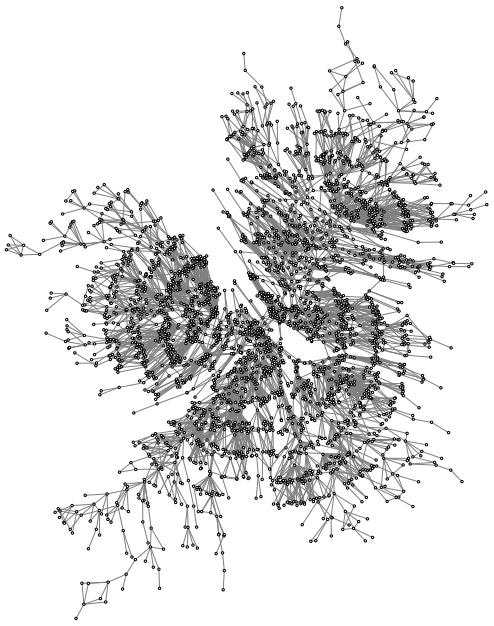


Figure 7.5: Stockholm in a network representation where the streets are nodes and the intersections links. From this figure one can clearly see the modular nature of city networks. In Stockholm this is effect is magnified by the fact that the city consists of islands.

From this investigation we have seen that the typical distance in a random

network is small, around two or three. If, however, the network is modular this distance increases to the mean distance within the modules. The distance at which T saturates correlates with when most of the network has been reached. We have also seen that in real networks the nodes seem to have a neighbors that have similar connectivity.

Chapter 8

Conclutions

This entropy mesuare is indirecty based on the topology of the network. We have shown that it is sensitive to structural properties of networks such as hirarcy and antihierachy. We have seen that the connectivity of the node is very important for the T value at one node and that when the fullnetwork is wieved nodes of both high and low degree contribute and we can say that the mesuare is democratic. From the distance study we have seen that $\langle T \rangle$ saturates when most of the network has been reached.

The next step in the investigation whould be to see how it correlates with robustness and if it can be used to make predictions about the rubustness.

Bibliography

[1] P. Erdős and A. Rï¿½yi, Publ. Math. Debrecen, 6, 290 (1959).

[2] G. Youle. Philosophical Transaction of the Royal Society of London, (Series B), 213:21-87 (1925)

[3] H.A. Simon Biometrika, 42(3/4):425-440 (1955)

[4] A.-L. Barabasi and R. Albert, Science, 286, 509 (1999)

[5] P. Minnhagen M. Rosvall, K. Sneppen and A. Trusina cond-mat/0406752 Physica A 340 (4)pp. 725-732 (2004)

[6] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani cond-mat/0108043

[7] R.V. Sol'e, R. Pastor-satorras, E.Smith and T. Kepler Adv. Complex. Syst 5, 43-54 (2002)

[8] G.Wilk, Z.Wlodarczyk cond-mat/0212056

[9] Michel Bauer and Denis Bernard cond-mat/0206150 '

[10] Ramon Ferrer and Ricard V. Solï¿½Statistical Physics of Complex Networks, Lecture Notes in Physics, Springer (Berlin), 114-125

[11] S. Maslov and K. Sneppen, Science 296, 910 (2002).

[12] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen. Phys. Rev. Lett. 92, 178702, (2004).

[13] Mark E.J. Newman "Random graphs as models of networks", Handbook of Graphs and Networks: From the Genome to the Internet, Editors Stefan Bornholdt, Heinz Georg Schuster, wiley-vch, pp.35-68

[14] Sanjay Jain and Sandeep Krishna "Graph theory and the evolution of autocatalytic networks", Handbook of Graphs and Networks: From the Genome to the Internet, Editors Stefan Bornholdt, Heinz Georg Schuster, wiley-vch, pp.355-395

[15] Sergei Maslov, Kim Sneppen "Correlation profiles and mortifs in complex networks", Handbook of Graphs and Networks: From the Genome to the Internet, Editors Stefan Bornholdt, Heinz Georg Schuster, wiley-vch, pp.169-198

[16] Lloyd Demetrius and Thomas Manke Physica A 346 (2005) 682-696

[17] M.E.J Newman Phys. Rev. E64,016132.

[18] K. Sneppen, A. Trusina and M. Rosvall, Europhys. Lett. 69 (5), 853 (2005)

[19] N. E. Friedkin, The UNC Press (1983).

[20] S. Valverde and R. V. Solï¿½ Eur. Phys. J. B 38, 245 (2004).

[21] A. Trusina, K. Sneppen, and M. Rosvall, Phys. Rev. Lett. 94, 028701 (2005).

[22] D. Watts and S. Strogatz, Nature 393 (1998).

[23] R. Albert, H. Jeong, and A. Barabasi, Nature, 406, 378 2000.

[24] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, Phys. Rev. E 65, 056109 (2002).

[25] S. Milgram, Psychol. Today 1, 61 (1967).

[26] M. Kochen, Ed., "The Small World" (Ablex, Norwood, 1989).

[27] J. M. Kleinberg, Nature 406, 945 (2000).

[28] D. J. Watts, P. S. Dodds and M. E. J. Newman, Science 296, 1302 (2002).

[29] J. M. Kleinberg, in T. G. Dietterich, S. Becker and Z. Ghahrmani (eds.), Proceedings of the 2001 Neural Information Processing Systems Conference, MIT Press, Cambridge, MA (2002).

[30] M.E.J Newman cond-mat/030945

[31] M.E.J Newman and M. Girvan Phys. Rev. E69, 026113.

[32] Joyong Park and M.E.J Newman cond-mat/045566

[33] Reka Albert and Albert-Laszlo Barabasi cond-mat/0106096

[34] M. Rosvall, A. Gronlund, P. Minnhagen, K. Sneppen cond-mat/0505400

[35] A. Trusina, M. Rosvall, K. Sneppen cond-mat/0412064

[36] M. Rosvall, P. Minnhagen, K. Sneppen cond-mat/0412051

[37] M. Rosvall, A. Trusina, P. Minnhagen K. Sneppen cond-mat/0407054

[38] Kasper Astrup Eriksen, Ingve Simonsen, Sergiei Maslov and Kim Sneppen cond-mat/0212001

[39] Enzo Marinari, Remi Monasson, Guilhem Semerjian cond-mat/0507525

[40] Ricard V. Solï¿½and Sergi Valverde Lect. Notes Phys. 650,189-207

[41] Stockholm: M. Rosvall, A. Trusina, P. Minnhagen and K. Sneppen. "Networks and Cities: An Information Perspective" Phys. Rev. Lett. 94:2, 028701 (2005).

[42] Airports: From Pajek dataset at: http://vlado.fmf.uni-lj.si/pub/networks/data/

[43] Internet: Website maintained by the NLANR Measurement and Network Analysis Group at: http://moat.nlanr.net/

[44] WWW: Homepage of A.-L. Barabasi: http://www.nd.edu/ networks/resources.htm

[45] YPD: 1: P.E. Hodges, A.H. McKee, B.P. Davis, W.E. Payne and J.I. Garrels, "The Yeast Proteome Database (YPD):a model for the organizition and presentation of genome-wide functional data" Nucleic Acids Res., Jan 1;27(1):69-73 (1999)

2:P.E. Hodges, W.E. Payne and J.I. Garrels, "The Yeast Proteome Database (YPD):a curated proteome database for Saccharomyces cerevisiae" Nature 407, 651-654 (2000)

[46] E.coli prot.: 1:P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, J. Collado-Vides, S.M. Paley, A. Pellegrini-Tool, C. Bonavides and S. Gamma-Castro, "The EcoCyc Database" Nucleic Acids Res., Jan 1;30(1):56-8 (2002)

2:I.M. Keseler, J. Collado-Vides, S. Gamma-Castro, J. Ingraham J, S. Paley, I.T. Paulsen, M. Peralta-Gil and P.D. Karp, "EcoCyc:a comprehensive database resource for Escherichia coli" Nucleic Acids Res., Jan 1;33(Database issue):D334-7 (2005)