GROWING NETWORKS

From the strucutre of the Internet to a strategy against AIDS*

by János Kertész (Budapest)

first published in: G. Sperl / M. Steiner (Eds.): Vermehrung: Die Geheimnisse der Multiplikation. Vienna: Edition Gutenberg (Was für Zeiten) 2004.

This paper was presented as Keynote Speech 3 on December 12 2003.

* Thanks are due to Margot Wieser for her insisting on my writing this essay.

Networks: The scaffold of complexity

Fashions govern our lives. Although we like to consider ourselves as most independent individuals especially in intellectual matters, trends and fashions often set our frames, even in science. »Network«: Hardly another word has become so popular in science recently. From sociology to economics, from linguistics to physics, from biochemistry to computer science hundreds of scientists are identifying and characterizing networks and processes related to them.

Why is this sudden burst of activity, why is this fashion of networks? In other words: Is there any deep reason behind this »movement«? The answer is yes. The ambition of scientists has driven them to try to understand most complex systems like the living cell, the human brain, the economy or other social systems. In diverse fields the detailed description of the interactions has reached a very refined level, the knowledge on details has grown enormously without enabling real understanding of the whole systems and indicating that it is time to look at them from a different point of view. There is hope that the mapping and understanding of the mere structure of the whole system as imposed by the interactions will bring us closer to the goal. Having this demand, starting with the late nineties there has been a real breakthrough in the science of networks and further fascinating results have been produced since then. This development is not independent from the former one: The high level of computerization makes it possible to analyze huge data systems related to the particular problems.

»Complex system« is a term introduced in physics for systems where, due to the large number of constituents and the interactions between them, the behavior becomes much more complicated than that of the single units. In simple terms: The whole is more than the sum of the parts. Cooperativity, competing interactions and non-linearity are key concepts in this respect and many examples could be named from magnets through glasses to granular systems. Clearly, the notion goes beyond physics. The examples mentioned in the previous paragraph are just illustrations for such systems.

We have learned from physics that in many cases the details of the interaction become unimportant for appropriately posed questions; seemingly very different systems have universal behavior. E.g., liquid-gas and some ferromagnetic transitions are in close analogy though the systems and the interactions are very different. There is good reason to believe that the approach of physics can be extended to a much broader set of systems even beyond physics.

What happens if we totally ignore the nature of the interactions and deal only with the topology generated by them? Let us denote the constituents of the complex system by dots or nodes. Some of them interact with each other, and if there is such an interaction we connect the corresponding nodes by a line or a link. What we obtain is a network. Graph theory, a well-established subfield in mathematics initiated by the Swiss mathematician Euler in the 18th century, deals with such objects (which are there called graphs) and a huge amount of knowledge has accumulated about them. Graph theory has been used widely in electrical engineering, in chemistry or in sociology. The revolutionary development referred to above has been achieved by using new methods to study the properties of networks mainly taken from the field of physics and by applying the approach to a very broad range of phenomena.

Networks everywhere

As already introduced above, a network consists of nodes and links (where the links can be directed or undirected, depending on whether the interactions between the units is unidirectional or bidirectional). First let us mention some examples for networks from very different fields.

Phenomenon	**Nodes**	**Links**
Cell Metabolism	Molecules	Chemical Reactions
Scientific Collaboration	Scientists	Joint Papers
WWW	Pages	URL Links
Air Traffic	Airports	Airline Connections
Economy	Firms	Trading
Language	Words	Synonymous meaning
Society	People	Acquaintances

Table 1. Examples of networks

The first example in table 1 is taken from biochemistry. The complex chemical mechanisms of the cell are carried by the molecules, proteins, enzymes etc. These are the nodes of the chemical network. Those molecules that participate in a chemical reaction are linked. The second example is taken from sociology, in this case the sociology of science. The scientists can be considered as nodes of a huge network, namely that of scientific collaborations. The simplest way to measure the existence of an interaction between two scientists is to look for joint papers: If such a paper exists, the two corresponding nodes are linked. Of course, many other aspects of social life from friendship networks at schools to sexual relationships can be mapped by networks as has been done for long time in sociology. The third example is the most popular network of our times: The World Wide Web. The nodes here represent the homepages and the links are the URL pointers. Naturally, this is an example of a directed network since if there is a pointer from page A to B there is not necessarily also a link from B to A. Air traffic in example 4 can be modeled by a network in a natural way: The airports are the nodes and a link between two of them exists if there is a direct airline connection. Finally, the economy can be looked at from the network aspect in many ways and one is mentioned here: The firms are the nodes, which are connected if there is trade between them. The last example in table 1 is the widely investigated social network where the links can be, e.g., acquaintances.

It is easy to continue the above list. A qualitatively new feature from the empirical point of view is that, due to computerization, the quantity of the available data on networks has enormously grown. Here are just some numbers: The Hollywood movie actor data bank contains more than 450.000 persons with all their movies recorded. The Human Genome Project has produced a data bank of 30.000 genes and the sequence of 3 billion base pairs. The size of the WWW is estimated to exceed 5 billion pages. Every interaction on the stock markets is recorded, producing an ever-growing set of data from which economic networks can be mapped out. The empirical study of such huge data banks has lead to important discoveries.

The small world, friends of friends and scale freeness

»It's a small world«, we say if it unexpectedly turns out that there is an easy link via a few steps of acquaintances to a person. The famous letter experiment conducted by the American sociologist Stanley Milgram in the 1960's demonstrated that the »distance« between two arbitrarily selected persons is surprisingly small. This inspired the Broadway play and the related movie by John Guare, »Six degrees of separation« that made this term into a commonly used saying. There is a huge network of people on the earth with 6 billion nodes, where the average number of acquaintances-links per person-node is of the order of not more than thousand. Despite the huge number of nodes and the relatively small number of links it is said that on this network the average »distance« is only about six. How can it be so small?

There are similar phenomena in other networks too. The network of mathematicians is not so huge but still, it has a considerable size. They have a favorite game to calculate their own, so called Erdös (or *E*-) number, which is defined in the following way: Paul Erdös, the famous

Hungarian mathematician has the *E*-number 0. Those who have written a paper with him get the *E*-number 1. The *E*-number of a mathematician will be n+1 where n is the smallest *E*-number of his/her coauthors. The surprise is that an overwhelming majority of mathematicians has a very small (<5) *E*-number. A similar game can be played with the *N*- or Nobel numbers in other fields of science where Nobel laureates get *N*-number 0, etc.: Most scientists have a very small *N*-number. In the so-called Kevin Bacon game of American high school students the winner has to find the shortest path from an arbitrarily chosen actor to Kevin Bacon through pairs of actors linked by movies they played together in – again, the route is usually very short.

Measurements have proved that this *small world property* is present in many networks: Despite of the large number of the nodes and the relatively small number of links per nodes, the average distance between the nodes is rather small. This property is there in the WWW, in the Internet, in biochemical networks, in the air traffic net, etc.

Another generally observed, interesting property of many networks was also first identified by sociologists in social nets. They have observed that the number of second neighbor links is significantly higher than would be expected from a random assumption. In simple terms this means that, provided we are dealing with a friendship network, *friends of friends easily become friends*. On a graphical representation this means that the number of triangles will be very high. It is easy to quantify this observation by a number called clustering coefficient C, being the ratio between the number of existing triangles and the maximum number of triangles given the number of neighbors for each node. By this definition the clustering coefficient will be between 0 and 1, where 0 means negligible clustering and for full clustering we get 1. A wide variety of networks show high clustering: The movie actor web, scientific co-authorship, e-mail address books, or the WWW are just some examples. Generally, acquaintance-based linkage leads to high clustering. There are, however, cases that cannot be explained on this basis, like the high clustering in different food webs.

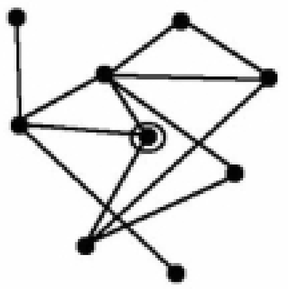


Figure1: Example of a network. The encircled node has degree k=3 and clustering coefficient C= 1/3 (out of 3 possible triangles 1 is present). The global clustering is the average of the clustering coefficients at the nodes.

A third, very robust property of a great variety of networks is what is called »*scale freeness*«. Every node in a network can be characterized by a number k, called degree, which is the number of links at that node. (For directed networks in-degrees and out-degrees have to be introduced.) We are now interested in the statistics of the degrees. It turns out that in most networks usually very many nodes have only one or a few neighbors and less have higher degrees, although nodes with very high degrees can also be found.

This can be described by what is called degree distribution. Most distributions have a characteristic value and it is very unlikely to find one that has much larger ones, but in the mentioned networks we do find such nodes with very high degrees. There is a special family of distributions, the power law: The probability to find a node with degree k is proportional to $k^{-\gamma}$. The specialty is that a power law has no characteristic value which would set a scale, i.e. it is scale free and the consequence is that the chance to find a node with very high degree is relatively high. Indeed, empirical studies showed that in a great variety of networks the distribution follows a power law, with γ being between 2 and 3.

It was the WWW and the Internet where this property was first found around the turn of the century and since then many networks from cell metabolism to movie actors, from scientific co-authorship to power grids, from sexual partnerships to language have turned out to be scale free.

Simple models of complex networks

The route to understanding in science goes via modeling. How is one to construct a model of real networks, which is mathematically or at least by computer tractable and reflects their pro-

1 Cf. D.J. Watts, D..J. / Strogatz, S..H.: Collective dynamics of small-world networks. In: Nature 393 (1998), p. 440.

perties? The first attempt was proposed half a century ago by the above-mentioned Erdös and another Hungarian mathematician, Rényi. They introduced the so-called random graph model, a beautiful mathematical object, which has been explored since then by generations of mathematicians. In this model a given large set of *N* nodes is taken and they are connected at random. The parameter of the problem is the average degree of the nodes which can change between *0* (no links in the network) and *N-1* (all links present). We are interested in large networks, i.e., *N* is a large number. There is a remarkable transition in this network if the links are put one after the other into the network: When there is one link per node on average suddenly, besides the small fragments, a giant interconnected component occurs. Or, the other way around: If we start from a fully connected graph and remove the links one by one randomly, the graph falls apart into small segments at the transition point.

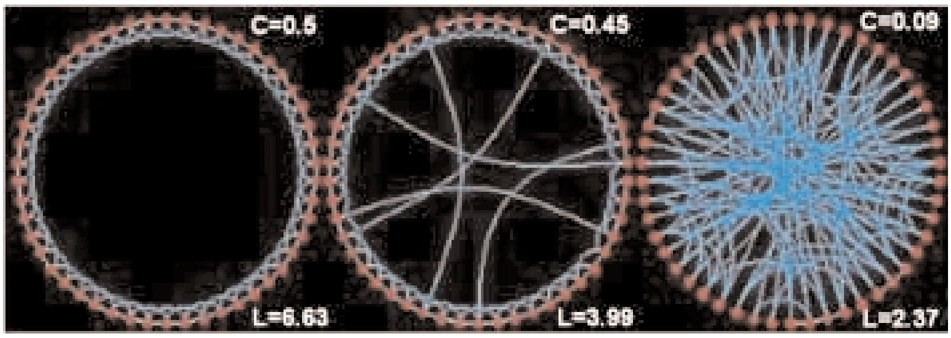


Figure 2: Watts-Strogatz network. A few long-range links introduced to the regular network assure the small world property (strongly reduced average distance *L*), while the clustering coefficient *C* remains relatively high. If most links are rewired the Erdös-Rényi model is obtained. [1]

Is this model applicable to those problems listed in the previous section? The average distance in the giant component becomes indeed small as more and more links are introduced, however, the clustering coefficient becomes large only when the number of links per nodes strongly increases. In 1998, Watts and Strogatz introduced a very interesting and simple model to overcome this problem: They started from a regular network where by construction the clustering coefficient was large, and then a fraction *p* of the links is rewired at random. They showed that there is a broad range in *p* where the clustering coefficient is still high but, due to the random links the small world property is already present. The idea was taken from sociology: People live in small communities where they mutually know each other (high clustering) while some of them posses links to members of other communities. Figure 2 shows a realization of the Watts-Strogatz network. Although the WS model has nice and interesting properties, an important feature of many real networks is missing: It is not scale free. The degree distribution is peaked at a characteristic value similarly to the Erdös-Rényi graphs.

Barabási and his collaborators realized that one crucial aspect of most real networks has not been taken into account in the above model: They are produced by growth.

Growth: A key aspect

Indeed, in most of the examples listed in Table 1 the networks emerge from a growth process. It is not that a given number of nodes can be considered as given and the links are then put into the system according to some rules like in the Erdös-Rényi or the Watts-Strogatz model. The number of nodes is not constant, it increases with time and the created structure influences how the new nodes are connected to the older ones. What is this rule of new connections? Barabási and Albert proposed a very simple one: Preferential attachment. A new node brings a given number of »hungry« links and such a link is connected to a node of the already existing network chosen with a probability proportional to its degree. The idea is based on such commonplaces like »the rich get richer« or popular people make new friends more easily.

The simple rule of preferential attachment leads to scale free growing networks. With the help of this model many interesting features of the scale free networks can be studied. First of all it clearly shows the importance of the rare but highly connected nodes, the so-called hubs. They are responsible for the small world property, for the fact that in many networks the average distance between the nodes is so small. In the Erdös, Nobel or Kevin Bacon games the successful strategy is that one looks for such hubs and finds a path that leads through them. If you want to test if the »six degrees of separation« is true, choose an arbitrary person on earth and try to make a chain of handshakes to her/him. First it becomes clear that it is much easier to find a short path to well known personalities. If you want to make the task harder you choose a less known person, like an Indian peasant. However, he will also be connected with a few links to some hubs and the hubs themselves are interconnected. (It is worth mentioning that this line of thought was written down first in a short story by the Hungarian author Frigyes Karinthy as early as 1929 with reference to a globalizing world.[2])

Preferential attachment is a simple principle but in reality we do not choose our friends by giving weight to the candidates proportional to the number of their friends and do not make a decision accordingly. If we compose a website and put some pointers on it, again the principle of choice is not preferential attachment. However, it was shown that in many cases effectively preferential attachment works. There have been measurements on scientific collaboration networks supporting this and simple models indicate that acquaintance–based networks tend to follow the preferential attachment principle. It should be stressed that there must be other mechanisms leading to scale free networks too. An immediate indication of this is that the clustering coefficient of the original model as described here leads to a far too low clustering coefficient. However, flexibility and simplicity have made the Barabási-Albert model very popular – e.g., it is easy to modify it in order to cure the problem with the clustering coefficient. Before closing this section I will show a small graph grown by this modified preferential attachment algorithm (Figure 3). The size of the nodes indicates their degree thus the hubs can easily be identified. The high number of triangles in the graph shows large clustering.

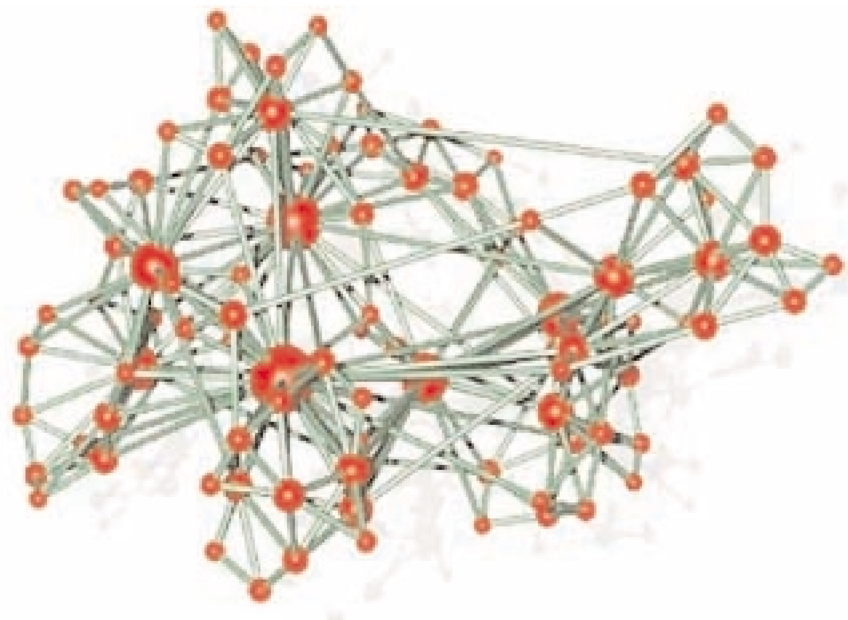


Figure 3: A network grown by the modified preferential attachment algorithm with enhanced clustering coefficient (originally due to Holme and Kim). The size of the nodes indicates their degree.[3]

2 Karinthy Frigyes: Láncszemek. In: Minden másképpen van. Budapest 1929 [in Hungarian].

3 Cf. Szabó, Gábor: Structure and Dynamics of Complex Networks [PhD thesis] Budapest University of Technology and Economics 2004.

Robustness and vulnerability of scale free networks

In order to be one of the generic structures observed from human interactions down to the biochemical reactions scale free networks must have some very specific feature: They must be robust against random failure. This is a most important requirement if complex structures are constructed irrespective of their human or natural origin. Scale free networks do show this robustness. This was demonstrated by the following experiment: Consider a scale free network – it does not matter whether it is a map of a real net or it is just grown by one of the above-mentioned algorithms. Eliminate randomly the nodes one after the other at random and monitor when the network will fall apart. This imitates, e.g., a computer network where the machines get it wrong with some probability and the question is what portion of the functioning units are needed in order to still allow networking. Here it is a crucial point that the failure is at random: We eliminate the nodes irrespectively of their degrees.

The procedure is very similar to that mentioned in the context of the Erdös-Rényi model where we saw an abrupt transition between the fragmented case (with links below the threshold) and the case where there was a giant component (above the threshold). The result with scale free networks is astonishing: The larger the network is, the smaller is the portion of nodes needed to hold together the network! This is a clear effect due to the hubs. Since nodes are selected at random, the chance of eliminating a rare but very important node is small. Therefore the network will be very robust against random failures.

The situation changes dramatically if an intentional attack is performed. How to paralyze a computer network? Obviously, the hubs have to be the targets. Indeed, it was shown that if the removal of the nodes happens with a probability proportional to their degree, the network falls apart at a nonzero ratio of the remaining nodes even in the very large network limit. The conclusion is that besides their considerable robustness against random failure scale free networks are very vulnerable against intentional attacks. That is why Barabási called the hubs the Achilles heels of the networks.

These results obviously have far-reaching consequences. One is on the military side, which receives particular importance in the age of terrorism. The terrorists of 9/11 chose just the main military, governmental and economic hubs of the US as targets with the hope that the attack would paralyze the country. Of course, defense experts have always been aware of the importance of central units and appropriate measures are taken, including built-in redundancy. However, the modern science of networks can provide new aspects to this issue. If a network is scale free – and there is good reason to believe that much of society is organized in this way – then the good news is that there is not a single, most important node that the whole structure depends on: There is a whole hierarchy of nodes. This is at the same time the bad news too: Special protection has to be extended to a much broader set of units.

Spreading in a network

Networks are the scaffolding of complex systems and it is the processes on them that make them particularly interesting. How is the structure related to the dynamic phenomena taking place on the networks? This is presently the most intensively investigated question in the science of networks.

In communication networks the information is expected to spread as fast as possible and in order to design such networks one has to understand aspects of path optimization and jamming. However, we are faced with such spreading phenomena in everyday life: The development of trends and fashions is just another example. Of course, in this case, there is a »supernode« which has links to practically all members of the society – and these are the communication media. The efficiency of chain letters and related games demonstrate how highly »wired« our society is. It is amazing that within days a petition (e.g., against the Iraq war last year) can run around the globe and, if the aim finds resonance, millions of signatures are collected. But false alarms can spread equally fast as was the case last year in Hungary when the rumor of an explosion in the nuclear power station swept through the country resulting, among others, in the evacuation of schools – the rumor was based on a chat by kindergarten kids. Again, based on network concepts, police could trace back the initiator of the rumor within hours (it was the journalist friend of the kindergarten teacher). The positive and negative consequences of the possibility of this rapid mobilization have to be understood in the future.

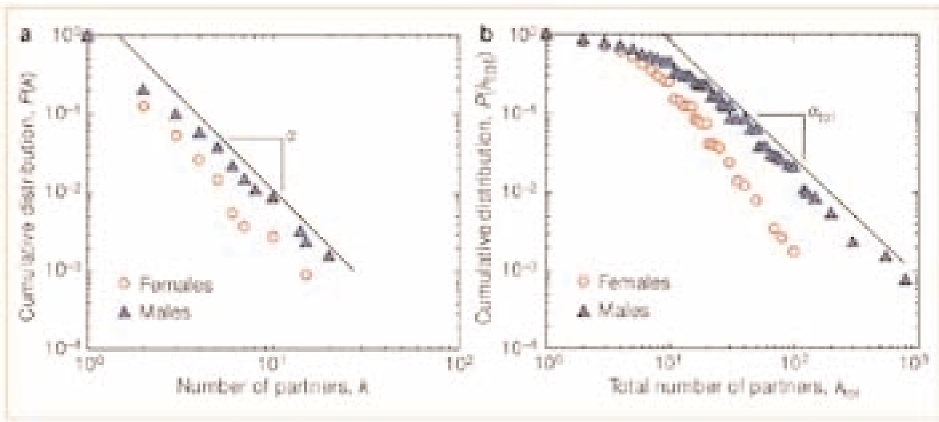
Unfortunately, not only information can spread on communication networks, but viruses too. Everyone who uses the Internet has had annoying experiences with the epidemics of the computer age: Computer viruses and worms. Software experts are usually able to rapidly develop a medicine, a virus-killer: A special program that eliminates the virus or at least isolates the infected elements into quarantine. Having discovered a virus, a serum is soon produced: Again it is a program that recognizes the beast and does not let it into the computer achieving immunity for that specific disease.

In spite of the fast reaction of the computer doctors, computer viruses have remarkable success. Epidemics caused by »Love Bug« or »W32« viruses have a persistence that seems to be in contradiction with the efforts against them. Why is it not enough to immunize a considerable part of the computers to stop the spreading of the disease? The answer is in the scale free topology of the Internet. The immunization is the analogy of the removal of nodes in the example of the previous section: In a scale free network practically all computers should be immunized to eliminate the giant component of vulnerable units, which is the carrier of the epidemics.

The spreading of diseases is in many respects analogous to that of human epidemics. Computer viruses can spread on the Internet only where the connections are clearly given. This is not the case for most infectious diseases, which can spread through accidental contacts or even through the air. The situation is different for sexually transmitted diseases, which spread on a very specific social network, namely that of sexual contacts. This network has been studied in recent decades in different contexts, but its mapping is extremely difficult due to the very personal nature of the problem. However, instead of mapping out a whole network of sexual contacts where everyone should declare the names of her/his partners and they of their partners etc., for many purposes it is enough to have a statistic about the number of different partners, a question to which people give out information more easily. This statistic does not allow us to learn the topology of the network, but it gives the degree distribution. Such a study was recently carried out in Sweden where 5000 heterosexual, sexually active males and females were asked about the number of their different sexual partners during the last year and in their whole lifetime. The results are illustrated in Figure 4.

There are remarkable features in these graphs. One is that, although the sample was heterosexual, the male declarations are always higher than the female ones. The possible explanation for this is on the one hand the higher social expectation for male sexual activity, which may lead to distortions with opposite signs in the two groups. On the other hand, the female statistic does not include prostitutes, while the male declarations may refer to them.

Most remarkable, however, is that the distribution shows a clear power law character. This indicates that the human sexual network is scale free, with many people having only few partners and a few having enormously high number of them. The latter are the hubs of the sexual network and they play a specific role in the transmission of diseases.



F*igure 4*: The (cumulative) distribution of sexual partners during the last year (left graph) and in the whole lifetime (right figure). Note the logarithmic scales. Straight lines in these plots mean power law dependencies and indicate the scale free structure of the underlying network.[4]

What is the consequence of this finding to the strategy against sexually transmitted diseases like AIDS? In an optimal situation all endangered individuals should be handled equally. What happens if the resources are limited for some reason? A natural approach would perhaps suggest a random choice; however, this is not an efficient way to beat the disease: We have seen that in a scale free network with random elimination the chance for an epidemic to survive is high. Instead, it is suggested that the strategy of the »aimed attack« should be followed: The hubs should be identified and treated with special care. There are very efficient ways to identify the hubs: Since they have so many connections, it is enough to choose an arbitrary node and follow the connections, which will likely lead to a hub very soon. Of course, such a strategy raises several moral questions, which I do not want to address here.

4 Cf. Liljeros, F./Edling, C.R./Amaral, L.A.N./Stanley, H.E/Åberg, Y.: The Web of human sexual contacts. Nature 411, 211 (2001), pp. 907-8.

Summary

In this brief survey I have tried to present some of the recent results of the new science of networks. On the one hand we saw that universal features emerged from the detailed study of diverse networks: A great variety of networks from diverse fields share three properties: They are »small worlds«, they have high clustering coefficients and they are scale free. Based on this recognition, efficient models have been introduced and it has become clear that for many networks the aspect of growth is of central importance. The knowledge gained in these studies are closely related to our everyday life and it is expected to be useful in a wide range of disciplines from biochemistry through computer science to epidemiology.

The new science of networks is based on the recognition that after having learned much about the nature of the interactions we should concentrate on the scaffold of complexity. There is hope that our understanding of networks will proceed faster and that it will promote a synthesis where the two parts, both the features of the networks and the detailed properties of the interactions can be taken into account.

Selected Bibliography

Barabási, Albert-László: Linked. The New Science of Networks. Cambridge/Mass.: Perseus 2002.
Buchanan, Mark: Nexus. Small Worlds and the Groundbreaking Science of Networks. New York: W.W. Norton 2002.
Watts, Duncan J.: Six Degrees. The Science of a Connected Age. New York: W.W. Norton 2003.

János Kertész was born in 1950 in Budapest. He is physicist and has been professor at the Budapest University of Technology and Economics since 1992. He graduated from the Eötvös University and worked at the Hungarian Academy of Sciences. He has been a visiting scientist at the University of Cologne and at the TU Munich, for shorter terms at several research institutes in Germany, US, UK, Finland, Italy and France. Research interests: Statistical physics, including applications to social sciences.
Contact: kertesz@neumann.phy.bme.hu