# Exploiting the Scale-free Structure of the WWW

Niina Päivinen

Department of Computer Science, University of Kuopio
P.O. Box 1627, FIN-70211 Kuopio, Finland
email niina.paivinen@cs.uku.fi
tel. +358-17-16 2172, fax +358-17-16 2595

**Abstract.** This paper presents a new way of using scale-free network structure. Scale-free structure has been found in many real-world systems such as the World-Wide Web, the physical structure of the Internet, and social networks. In this study, some complex network models are introduced along with their properties, including the distribution of the number of links from each node and the error tolerance of the network. The construction of a special case of a scale-free network, a scale-free minimum spanning tree (SFMST), is presented. In this construction, the network is constantly growing and the connectivity and fitness of the nodes already in the network determine where the new node is linked to. Some freely available databases are processed using the SFMST method, and the usefulness of the obtained results in the field of pattern recognition are discussed.

## 1 About Networks

The World-Wide Web is an uncontrolled system into which practically anyone can add documents and link them to existing sites. The documents and the links between them build up a directed network with the documents as nodes and the links as directed edges. The enormous amount of the documents in the WWW makes it impossible to construct a network model containing all the documents and links between them. However, by using a robot that finds URLs from the web, Albert et al. [1] have found that the local distribution of links follows a power law. If this result is generalized to the whole WWW, all the documents in the web are only a few (on the average, about 19) clicks away from each other.

The most well-known models of network topology presented in the literature include the Erdös–Rényi (random graph) model, the small-world model, and the scale-free model [2]. In the *random* network model, each pair of nodes is connected to each other with some probability value. The probability that a node has exactly $k$ links follows the Poisson distribution. The construction of a *small-world* network begins with a lattice in which each node is connected to its nearest and next-nearest neighbours. Next, every link is rewired with some probability, meaning that one end of the link is shifted to another node selected at random. In both of these models the number of nodes is fixed, and each node has approximately the same amount of links. A *scale-free* network emerges when

new nodes are constantly added to the network. In the addition process, each new node is linked to a node already in the network in such a way that the linking probability is higher when the node is already well-connected. The name "scale-free" arises from the field of fractals and, in the context of networks, it means that there is no typical number of links from a node. Instead, a scale-free network contains a few nodes with many links and many nodes with only some links [3].

Real-world networks that are best modeled by scale-free structure include the WWW, the physical structure of the Internet, social networks, and electrical power grids. For example, new documents are constantly added to the WWW and some old ones are removed, so it seems logical that the best network model for the WWW has to be dynamic and not static, as is the much used random network model.

When dealing with real-world networks, the error tolerance is wanted to be as high as possible. A failure in a power grid may induce wide-spread blackouts which cause severe financial losses. It turns out that the functionality of a scale-free network is not disturbed even if a large number of nodes is not functioning correctly. However, there is a disadvantage: by removing a single node vital to the network, the whole network may be brought down [4]. These are the main results obtained by Albert et al. [5]. They have studied random and scale-free networks which have the same amount of nodes and links, and tested the effects of random failures of nodes as well as deliberate attacks to networks. Because in a random network each node has approximately the same amount of links, the failure of a single node always causes detectable damage to the whole network. On the other hand, if a random node fails in a scale-free network, it is most probable that the node was not highly connected and thus the failure might not cause any noticeable damage. The error tolerances change drastically if someone decides to disable the most well-connected nodes of a network. In practice, the effect to a random network is the same as in the case of a random failure, whereas the scale-free network can easily be fragmented into separate clusters with no links between them and the network does not function as intended.

Since power grids are best modeled using scale-free structure, wide-spread blackouts due to a single failure in the grid are very rare, but not impossible. If a saboteur has the information about the most connected nodes in the grid, he can do lots of damage with little effort.

## 2 Application

A network with a scale-free structure has some interesting properties. In order to take advantage of them, a new structure called a *scale-free minimum spanning tree* (SFMST) was defined. A minimum spanning tree of a weighted graph is a tree which connects all the nodes of the graph with a lowest possible cost [6]. There exists many algorithms for finding a minimum spanning tree; Prim's algorithm was modified to meet the requirement of preferential attachment to well-connected nodes to get a method for constructing an SFMST.

The construction of an SFMST begins with the calculation of the distances between the nodes. Then the distances are reversed, meaning that each value is subtracted from the maximum distance value, leading to a weight for each edge. At each step of the procedure a link with the greatest weight is added to the SFMST in such a way that the structure remains a tree (no cycles, no multiple links between the nodes). Then it is checked if the weights have to be updated: if the addition of a link increased the number of links from a node over a certain threshold value, then that node is made more attractive by increasing the weights of all the possible links to the node. It means the bolstering the fittest — a node with a high fitness is more likely to get additional links than a node with smaller fitness.

When constructing an SFMST, one has to define how much links a node has to have in order to get a bonus for large connectivity. In addition, the rewarding method has to be determined. Each new link has to affect to the weight but the effect must not be too big, and it might depend on the number of links of the node. Also the distance measure has to be determined. There are numerous possible dissimilarity measures which can be used in this purpose and the selection can be a difficult task. It requires some information about the data points, for example, are they binary, continuous-valued, discrete, or perhaps a collection of different kinds of quantities.

The SFMST can be constructed from both directed and undirected graphs; in this study, only undirected graphs were processed.

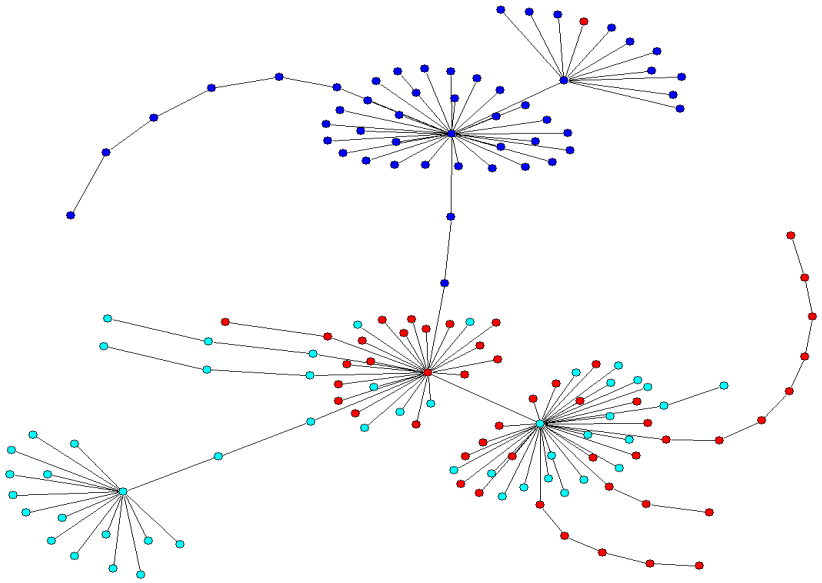

**Fig. 1.** An SFMST, iris dataset

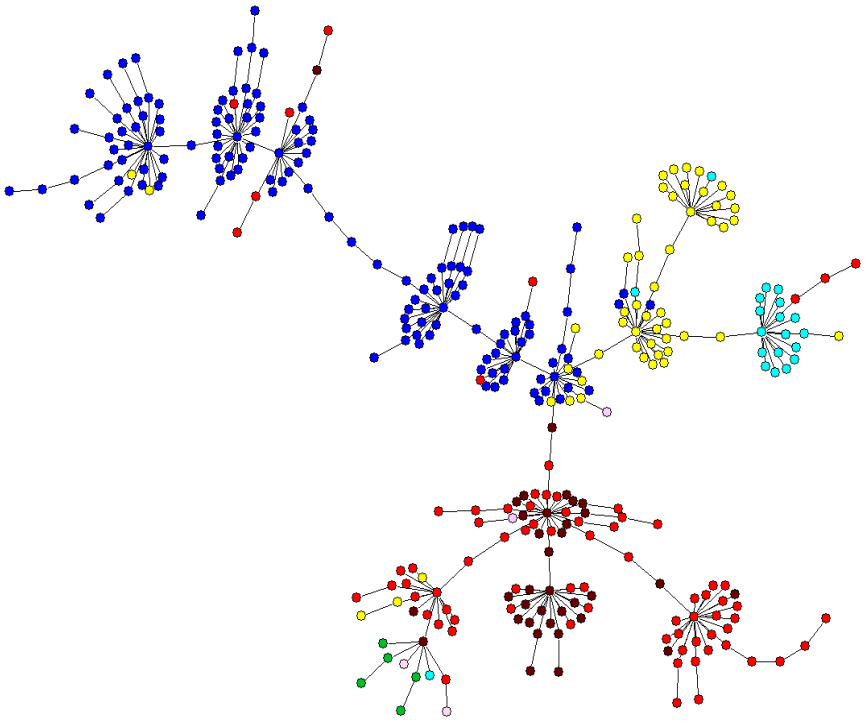

**Fig. 2.** An SFMST, ecoli dataset

## 3 Results

The SFMST construction method was tested using some freely available datasets from UCI Machine Learning Repository [7]. Each dataset consisted of multidimensional data points representing some measurements, one coordinate per one quantity. The data points were pre-assigned to classes, and the purpose was to determine if the points from different classes would settle themselves near each other in an SFMST.

The first dataset was the Fisher's iris plant dataset. It consists of 150 measurements from different iris plants, four measurements from each plant, and there are three different species. When defining three classes in such a way that each class contains one species, it is known that one class is linearly separable from the other two classes. As can be seen in Fig. 1, the linearly separable class (blue points in the figure) is located in one side of the SFMST, and the other classes (red and cyan points) are mixed with each other. Since the figure is a projection from a four-dimensional space to a two-dimensional plane, the edge lengths may not be proportional to the real distances between the data points.

In Fig. 2, the 336 seven-dimensional data points of the ecoli database are presented in a scale-free minimum spanning tree. There are eight different classes

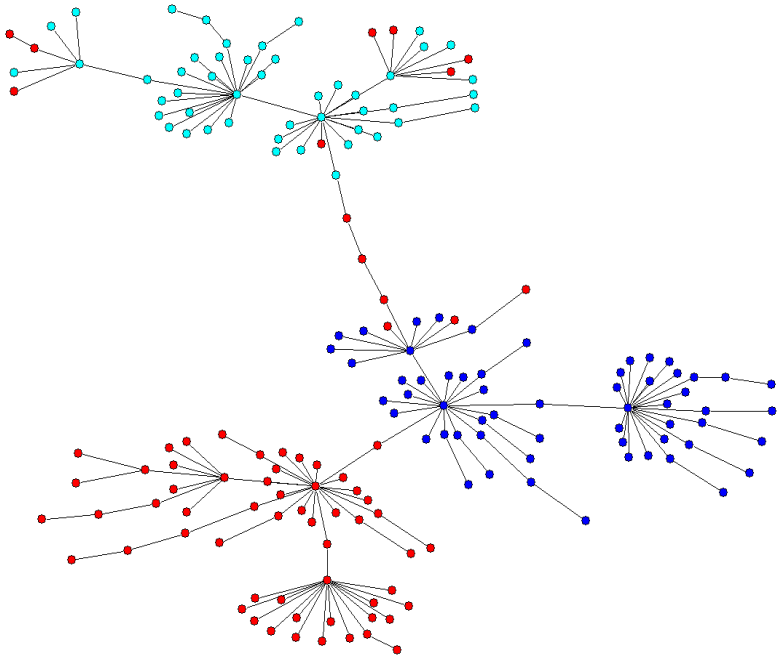

**Fig. 3.** An SFMST, wine dataset

which are color-coded. The database is meant to be used in the prediction of protein localization sites, and each class represents one localization site.

Figure 3 contains the 178 13-dimensional data points of the wine database. The measurements are the results of a chemical analysis of different wines grown in the same region, but in three different cultivation sites, in Italy. Each cultivation site corresponds with a class.

As can be seen in the figures, the points representing different classes are in general situated quite near each other in the SFMST. This suggests that it could be used as a pattern recognition method, by defining highly-connected nodes as cluster centers and the nodes linked to the center as the other cluster members.

The SFMST has been used as a clustering method and the performance was compared with the standard $k$-means method [8]. The results were encouraging.

## 4 Discussion

A scale-free structure found in directed networks was used in undirected networks. The usefulness of the resulted structure in pattern recognition problems was considered.

Pattern recognition is about finding somehow similar data points from a lot of data. If these data points are arranged in a scale-free structure, using a dissimilarity measure suitable for the application field, hopefully the similar

points are located within only a few links from each other. In this procedure, the scale-free property is induced to the data set under study in order to gain new information about the data.

What does it mean for a structure to be inherently scale-free? For example, fractals are scale-free in the sense that they look quite similar at all levels of magnification (self-similarity). The WWW can be assumed to be scale-free based on the study of the link distributions, and this structure has arisen without global orders of how and where new documents can be linked. A system where global order arises from local interactions, without supervision, is known as a self-organizing system [9]. Maybe there is a connection between scale-free and self-organizing systems?

If it is known that a certain network has a scale-free topology, this fact can be taken into account when some information is wanted to be extracted from the network. It has been found out that even the best Web search engines can find only a fraction of documents of the WWW [10]. If the searching method took advantage of the scale-free structure of the WWW, the results might be significantly better.

**Acknowledgements.** The author wishes to thank professors Tapio Grönfors and Seppo Lammi from University of Kuopio for their advisement in this research. The network pictures in this document were created with Pajek [11].

## References

1. Albert, R., Jeong, H., Barabási, A.L.: Diameter of the world-wide web. Nature **401** (1999) 130–131
2. Barabási, A.L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks. Physica A **272** (1999) 173–187
3. Strogatz, S.H.: Exploring complex networks. Nature **410** (2001) 268–276
4. Tu, Y.: How robust is the Internet? Nature **406** (2000) 353–354
5. Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. Nature **406** (2000) 378–382
6. Aho, A.V., Hopcroft, J.E., Ullman, J.D.: Data structures and algorithms. Addison-Wesley, Reading, Massachusetts (1983)
7. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html (1998) Accessed 21st June 2004.
8. Päivinen, N.: Clustering with a minimum spanning tree of scale-free structure. Manuscript submitted to Pattern Recognition Letters (2004)
9. Haykin, S.: Neural networks. A comprehensive foundation. Prentice Hall International, Inc., London (1994)
10. Lawrence, S., Giles, C.L.: Accessibility of information on the web. Nature **400** (1999) 107–109
11. Batagelj, V., Mrvar, A.: Pajek — program for large network analysis. http://http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm (2004) Accessed 4th March 2004.