Exploiting the Scale-Free Structure of the WWW

Niina Päivinen

with Tapio Grönfors and Seppo Lammi

Department of Computer Science

University of Kuopio

Contents

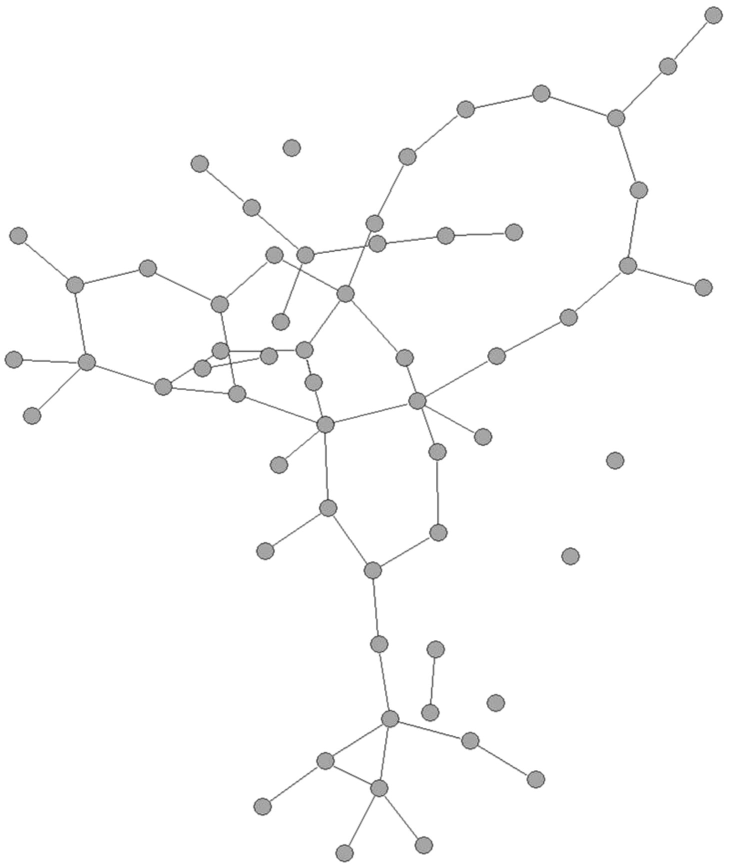
- motivation
- network models
  - random
  - small-world
  - scale-free
- an application of scale-free structure to clustering: a scale-free minimum spanning tree (SFMST)
- discussion

Motivation

- the size of the WWW is estimated to be at least $8*10^8$ documents
- it has been shown that two randomly chosen documents on the web are, on the average, only 19 clicks from each other
- What kind of topology does the network of the documents and the links between them have?

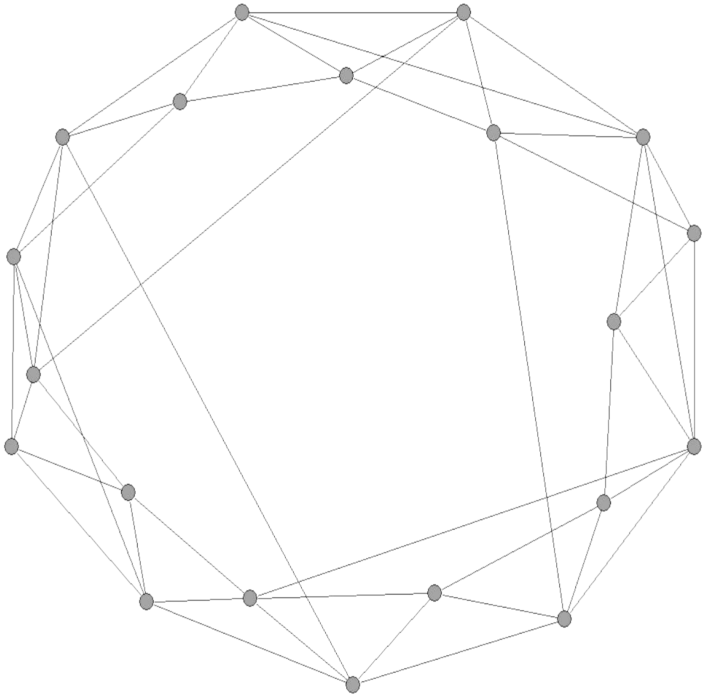
Random Network Model

- nodes are randomly connected to each other
- on average, every node has the same amount of links



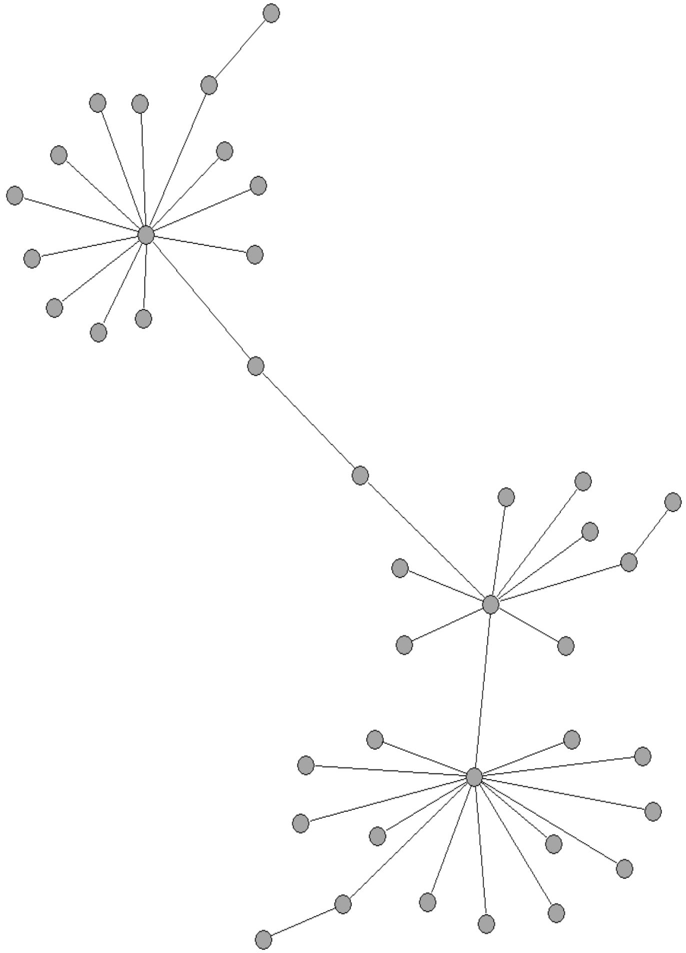
Small-World Network Model

- start with a lattice where each node is connected to its nearest and next-nearest neighbours
- add shortcut links between random nodes (or rewire existing connections)



Scale-Free Network Model

- add nodes to the network in such a way that linking probability is higher when the node is already well-connected
- a few nodes with many links, many nodes with only some links



The Scale-Free Structure

- pros
  - high error tolerance: with a high probability, random node failures do not cause much damage
- cons
  - vulnerability to (intentional) attacks: disabling the most well-connected nodes leads to several damage to the network performance

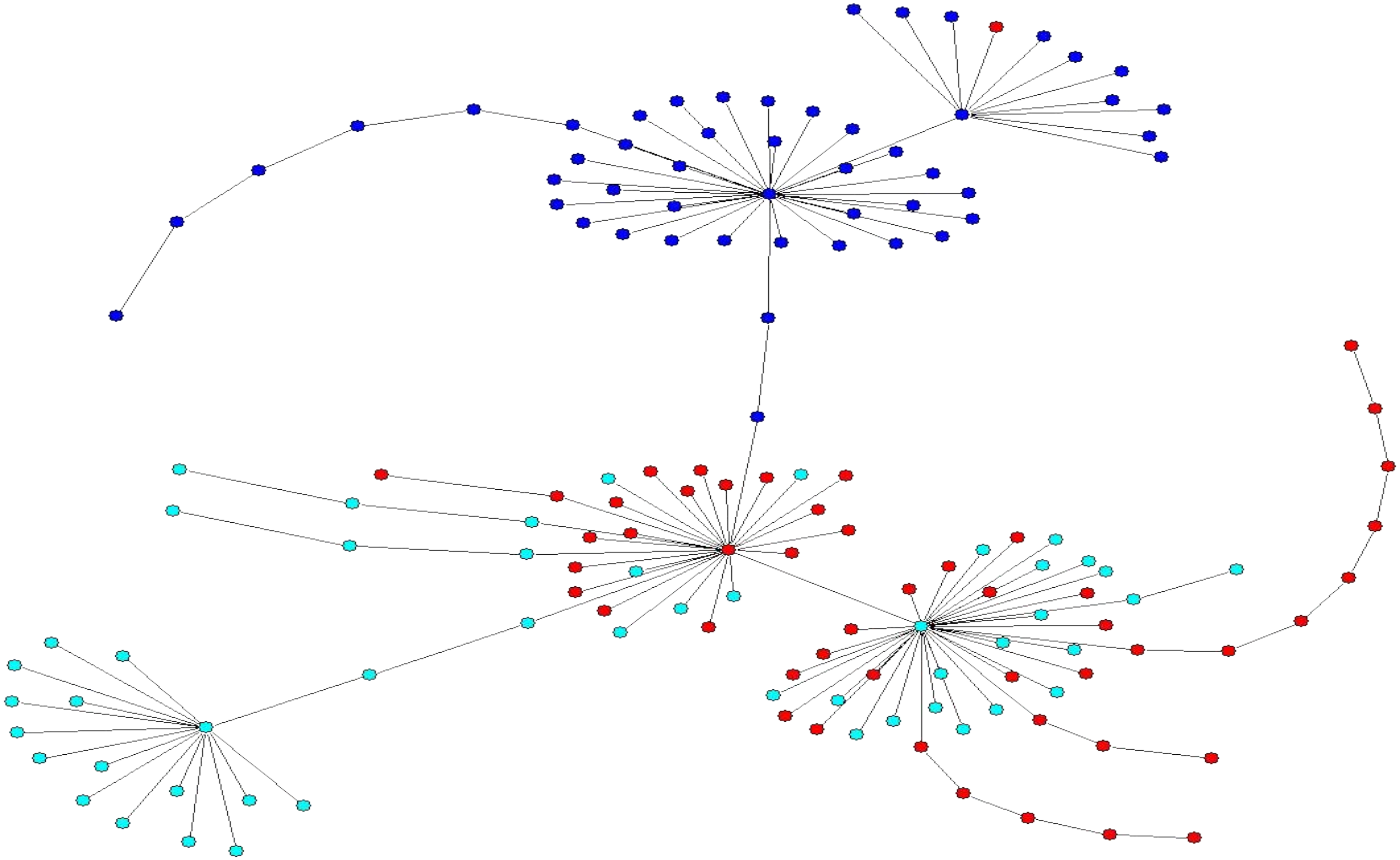
Application to Clustering

- a dataset containing some data points
- each point is represented as a feature vector (a set of measurements)
- for example, iris dataset: 150 data points, four measurements per each point, three different species of iris plants
- goal: cluster the data points in such a way that one cluster contains one species

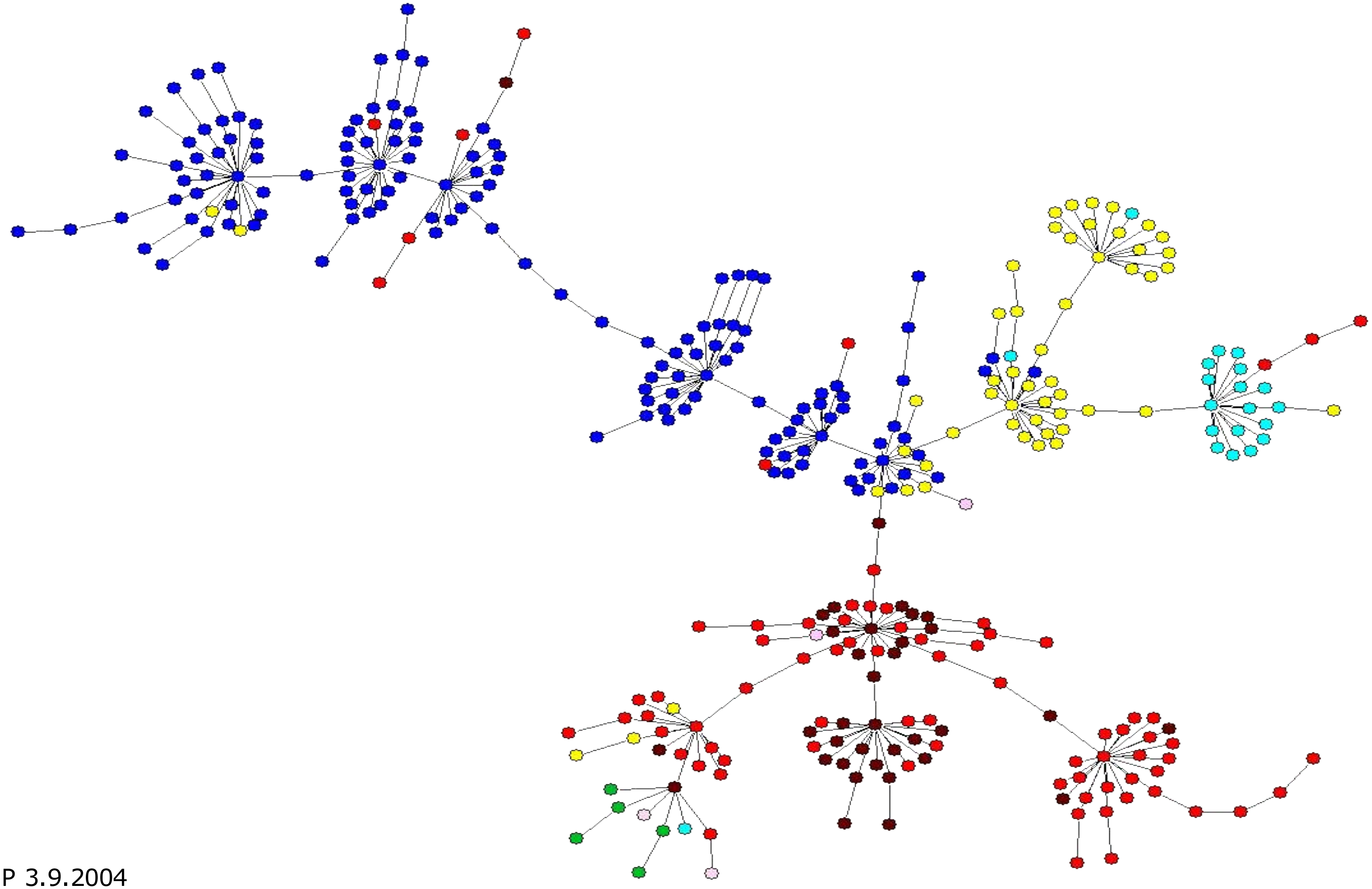
The Construction of an SFMST

- feature vectors = nodes
- calculate the distances between the nodes
- edge weights = reversed distances
- add the edge with the greatest weight
- repeat
  - select the edge with greatest weight in such a way that no cycles are formed and the tree stays connected
  - update weights if necessary
- until all the nodes are in the tree

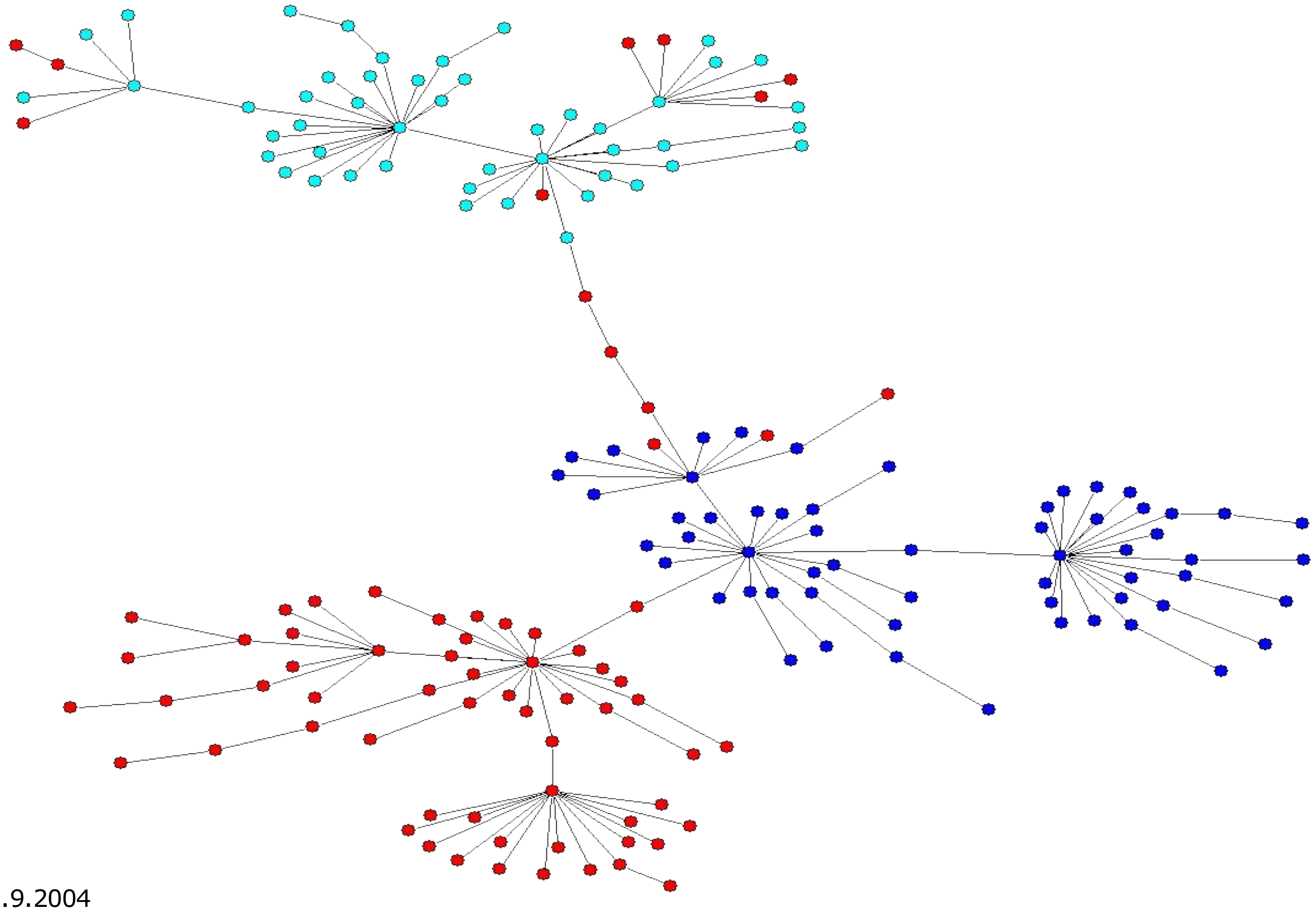
An SFMST, iris dataset



An SFMST, ecoli dataset



An SFMST, wine dataset



Discussion

- Could Web search engines take advantage of the scale-free structure of the WWW?
- Why does the scale-free structure seem to appear in many different circumstances and real-life situations?