11 Network Statistics

Michael Brinkmeier and Thomas Schank

Owing to the sheer size of large and complex networks, it is necessary to reduce the information to describe essential properties of vertices and edges, regions, or the whole graph. Usually this is done via *network statistics*, i.e., a single number, or a series of numbers, catching the relevant and needed information. In this chapter we will give a list of statistics which are not covered in other chapters of this book, like distance-based and clustering statistics. Based on this collection we are going to classify statistics used in the literature by their basic types, and describe ways of converting the different types into each other.

Up to this point *network statistic* has been a purely abstract concept. But one has quite good ideas what a statistic should do:

A network statistic . . .

. . . should describe essential properties of the network.
This is the main task of network statistics. A certain property should be described in a compact and handy form. We would like to forget the exact structure of the underlying graph and concentrate on a restricted set of statistics.

. . . should differentiate between certain classes of networks.
A quite common question in network analysis regards the type of the 'measured' network and how to generate models for it. This requires the decision whether a generated or measured graph is similar to another one. In many situations this may be done by identifying several statistics, which are invariant in the class of networks of interest. Using these statistics an arbitrary graph can be tested for membership in a specific class, by determining its statistics and comparing them with some references.

. . . may be useful in algorithms and applications.
Some network statistics may be used for algorithms or calculations on the graph. Or they might indicate which graph elements have certain properties regarding the application.

To which degree a certain statistic fulfills one or more of these tasks obviously depends on the application and the network. Therefore we will not go into detail about the interpretation, and restrict ourselves to the description of types of statistics, common constructions, and several examples.

U. Brandes and T. Erlebach (Eds.): Network Analysis, LNCS 3418, pp. 293–317, 2005.
© Springer-Verlag Berlin Heidelberg 2005

11.1 Degree Statistics

The most common and computationally easy statistic is the vertex degree. Depending on the underlying network and its application, it may be a simple measure for the strength of connection of a specific vertex to the graph, or – as in the case of indegrees – a measure for the relevance. But usually, instead of using this statistic directly, the main interest lies in the absolute number or the fraction of vertices of a given in-, out-, or total degree. It has been discovered that the distribution of degrees in many naturally occurring graphs significantly differs from that of classical random graphs.

In a classical undirected random graph $G_{n,p}$ the fraction of vertices of degree k is expected to be

$$\binom{n-1}{k} p^k (1-p)^{n-1-k} \qquad \text{(Binomial distribution)}$$

if the number of vertices n is small, or approximately

$$\frac{(np)^k}{k!} e^{-np} \qquad \text{(Poisson distribution)}$$

if n is large. But in many natural graphs the degrees seem to follow a *power law*, i.e.

$$c\,k^{-\gamma} \qquad \text{with } \gamma > 0 \text{ and } c > 0.$$

The power law is a good example for a parameter elimination

by means of a functional description, as described in Section 11.7.1. Since the degree distribution can be described as $ck^{-\gamma}$, it suffices to determine the constant exponent γ. This can easily be done with linear regression of the log-log-plot of the distribution. The scaling constant c is then dictated by the fact that the sum over all k is either the number of edges (in the absolute case) or 1 (in the relative case). Of course, the exponent is only meaningful if the degree distribution has the appropriate form.

Examples for graphs and networks whose degree distribution seems to follow a power law include

- the actor collaboration graph ($\gamma \approx 2.3$) [40],
- the World Wide Web ($\gamma_{\text{in}} \approx 2.1, \gamma_{\text{out}} \approx 2.45/2.72$)[1][16, 40, 102],
- the power grid of the United States of America ($\gamma \approx 4$) [40],
- the Internet (router and autonomous systems) ($\gamma \approx 2.2$) [197].

More details about the power law, and models for generating graphs satisfying it, can be found in Chapter 13.

Experiments and measurements indicate that the power law and the resulting exponents are good statistics for the classification of graphs. They are especially useful for the decision whether a specific model generates graphs which are close to the naturally occurring ones.

[1] In the examples shown in Figures 11.4 and 11.5 near the end of this chapter, we have $\gamma_{\text{in}} \approx 2$ and $\gamma_{\text{out}} \approx 3.25$.