EXperimental Learning

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Big Data and Social Analytics certificate course

MODULE 4 UNIT 1
Video 5 Transcript

MIT BDA Module 4 Unit 1 Video 5 Transcript

Speaker key

XD: Xiaowen Dong

HY: HapYak

XD: So, in the previous video we focused on the importance of individual vertices, in this video we will take a more global view of the structure of the networks. We will talk about three different network properties, namely degree distributions, diameters and average path lengths, and clustering coefficients and we will also introduce two important classes of networks.

HY: Degree distribution

XD: The first property of the network is the degree distribution, which intuitively measures the distribution of vertex degrees. Clearly this indicates the structure of the network, for example, in the so-called path graph the degrees of the vertices are similar to each other, while in the so-called star graph there is one vertex that has a significantly larger degree than all the rest.

Formally degree distribution is defined as the frequency distribution of vertex degrees. We can use Pk to represent the fraction of vertices that have a degree k. In this example we have one vertex with degree zero which means that it is basically isolated, we have three vertices with degree 1 and two vertices with degree 2, and the one vertex with degree 3. Therefore this graph has the following degree distribution.

00:01:27

We now discuss an important class of networks; here we see a synthetic graph on the left and its degree distributions on the right. Notice that in this graph most of the vertices have a relatively small degree where only a few have a substantially larger degree; such a behavior is reflected in the degree distribution shown on the right. We see that this distribution actually have a long tail which corresponds to the few vertices with large degrees. It turns out that many real-world networks have such a degree distribution, for example, in a social network it is usually the case that only a few people know an extremely large amount of people, this network would therefore have a long-tail degree distribution.

One important mathematical model for a long-tailed degree distribution is the power law distribution; in short, this means that the probability of having a vertex with degree k, namely Pk, varies as a power of k, as a consequence the logarithm of the degree distribution is a linear function of the logarithm of the degree. This can be verified by the figure on the right where we see that the degree distribution follows roughly a straight line in a log-log scale. A roughly straight line in the log-log scale is actually a signature of the power law distributions.

Power law distributions are important because many real-world networks have such a degree distribution, for example, a network of actors corroborating with each other, a network of web pages on the internet or a network of power transmission lines all have a degree distribution that follows a power law. Networks with power law distributions are called scale-free networks, and we will explain in more detail in the additional notes.

HY: What is your understanding of the following statement: The real world examples of actor collaboration, the world wide web and a power grid all have a long-tail degree distribution.

Very few in the network will have lots of connections to others in the network, but the majority (a large number) will have few connections to others in the network.

00:03:10

Diameter and average path length

XD: The second network properties are the diameter and average path lengths of networks. These two reveal the structure of the networks by measuring, in the worst case, or on average, respectively, how quickly one vertex can reach another vertex in the graph. For example, in the star graph every vertex can reach another vertex with at most two steps, where in a path graph it may take up to six steps to move from one vertex to the other.

Formally the diameter of a graph, dg is defined as the longest shortest path between any pair of vertices in the graph. The average path length lg is defined as the average length of shortest paths among all pairs of vertices. In this example we list all the shortest paths with their lengths between every pair of vertices in the graph; it is obvious to see that the diameter of this graph is 2, where the average path length is 4 over 3. Where diameters might be influenced by a particular pair of vertices, average path lengths in general tells how well the vertices are connected in the graph, so specifically the shorter the average path lengths the more connected the vertices are in the network. Together with clustering coefficients, which we will introduce next, they define two basic properties of the so-called small-world networks.

HY: The _ the average path length, the _ the vertices are connected.

A. Shorter; more
B. Longer; more
C. Shorter; less

00:04:35

Clustering distribution

XD: The last property of the network that I will mention is clustering coefficient which basically measures how likely the two neighbors V and W over vertex U are themselves connected in the graph, forming a closed triangle. In the practical example of a social network this amounts to asking whether two of your friends are themselves friends; the higher the clustering coefficient, the more triangles in the network, which indicates that the vertices are more closely connected.

For a formal definition of clustering coefficients we first define triplets and the triangles. A triplet is a set of three vertices that are connected either by two edges, which means it is open, or by three edges, which means it is closed. In the example on the right we have an open triplet which is centered on the vertex U; a triangle can then be considered as consisting of three closed triplets, one centered on each of the three vertices.

The clustering coefficient of a network is then defined as the fraction of connected triplets that are closed, which can be computed by the following formula. Notice that we times the number of triangles by three because one triangle actually corresponds to three closed triplets. We compute clustering coefficients on the following two example graphs.

For the first graph we have one triangle which is formed by vertices 2, 3 and 4, and we have 5 triplets where the center of each triplet is the vertex listed in the middle, therefore the clustering coefficient of this network is 3 over 5. For the second graph with an added edge in red, we have 2 triangles and 8 triplets, the clustering coefficient of this network has increased to 3 over 4, which is consistent with our perception that vertices in this network tends to be more well connected.

HY: A triangle consists of _ triplets.

A. 1
B. 2
C. 3
D. 4

XD: As a final remark the clustering coefficients that we introduced here are called global clustering coefficients, or sometimes transitivity.

00:06:48

We will introduce other definitions of clustering coefficients in the additional notes if you are interested. Average path lengths and the clustering coefficients are two properties used to characterize the so-called small-world networks.

A small-world network tends to have a short average path length and a high clustering coefficient. Many real-world networks are small-world networks, for example, the theory of 6 degrees of separation suggests that the average path length in a social network could be as small as 6; this is initially supported by Stanley Milgram's famous experiment in 1967 where he found out that on average it took only 5.5 steps of acquaintances for a letter to be sent from Nebraska to Massachusetts. Recent studies suggest that such a number can be even smaller on an online social network such as a Facebook network. The C. Elegans Neural Network is another classical example of small-world networks.

Finally, the scale-free network, which has a degree distribution that follows a power law is also one type of small-world networks. As I recap in this video we have introduced three network properties and two important classes of networks. We can compute such properties on the reality common networks or the networks of your own interest.