Chapter 11

OUTLIER DETECTION IN GRAPHS AND NETWORKS

“In nature, we never see anything isolated, but everything in connection with something else which is before it, beside it, under it and over it.” – Johann Wolfgang von Goethe

1. Introduction

Graphs represent one of the most powerful and general forms of data representation, which can express diverse data, ranging from multi-dimensional entity-relation graphs, the web, social networks, communication networks, and biological and chemical compounds. Broadly speaking, two kinds of graphs arise often in real domains:

- The data may contain many small graphs, drawn over a small base domain of labeled nodes. Some examples of this scenario include chemical and biological compounds. The labels correspond to the chemical elements. Individual graph objects are defined as outliers based on the model of normal graph objects in the database.

- The data may be represented as a single large graph. Examples include the web, social and information networks. In some cases, such as the web and social networks, the nodes may correspond to distinct identifiers such as URLs, actors, or IP addresses. In other cases, the node identifiers are not unique. For example, if the nodes on the web are annotated by their subject category rather than their URL, the node labels are not unique. In other cases, no node labels may be available at all. These scenarios are somewhat different, which may allow the definition of different kinds of node, linkage or subgraph outliers.
In addition, other cases exist, which may be closer to one of the afore-mentioned scenarios. For example, multiple small graphs may be extracted from a larger network. An example of such a scenario is a bibliographic network, in which a publication may be represented as a small graph over a larger bibliographic co-authorship network. In this problem, individual small graphs may be defined as outliers based on their linkage relationships. As discussed in section 3.2.1, the methods for outlier analysis in a single large graph can be easily generalized to this case.

These different kinds of data require different methods for outlier analysis. For example, in the case of small graphs, a single object may be represented as an outlier. However, in large graphs, the outliers are defined as portions of the network, which may be nodes, edges or even subgraphs.

In temporal graphs, the structure of the network may change over time. Such scenarios typically occur in large scale web, social or communication networks, and therefore belong to the second or third cases discussed above. Outliers in temporal data may correspond to significant changes in specific structural aspects of the network such as the communities, shortest paths, or other local structural properties. Temporal graphs represent one of the most challenging cases for outlier analysis, because of the many different ways in which outliers may be defined. The following phrase from Chapter 1 is restated here: “The more complex the data, the more the analyst has to make prior inferences of what is considered normal for modeling purposes.” For example, in a temporal network, an outlier node could be a node with unusually high degree, unusual connectivity structure, unusually changing degree, unusually changing community structure, unusually changing distances to other nodes, or unusual relationships of node content to linkage structure. There are virtually an unlimited number of ways that outliers could be defined. Even within the context of a specific outlier type such as a node outlier, the appropriate model of regularity could be based on its degree, neighbor set, edge weight distribution etc. Therefore, it is somewhat perplexing in an unsupervised application, how one should define an outlier.

In such scenarios, it is important to define change analysis and outlier detection problems from an application-specific perspective, since there is no uniform definition of an outlier. Specific applications may provide better guidance about outliers. For example, in a spam detection application, the degree distributions of nodes can provide insights about outliers. In a network de-noising application, the linkage connectivity structure can be used to determine outlier links. Therefore, application-specific definition of outliers is needed.