Chapter 26
Graph Mining Techniques: Focusing on Discriminating between Real and Synthetic Graphs

Ana Paula Appel
Federal University of Espirito Santo at São Mateus, Brazil

Christos Faloutsos
Carnegie Mellon University, USA

Caetano Traina Junior
University of São Paulo at São Carlos, Brazil

ABSTRACT

Graphs appear in several settings, like social networks, recommendation systems, computer communication networks, gene/protein biological networks, among others. A large amount of graph patterns, as well as graph generator models that mimic such patterns have been proposed over the last years. However, a deep and recurring question still remains: "What is a good pattern?" The answer is related to finding a pattern or a tool able to help distinguishing between actual real-world and fake graphs. Here we explore the ability of ShatterPlots, a simple and powerful algorithm to tease out patterns of real graphs, helping us to spot fake/masked graphs. The idea is to force a graph to reach a critical ("Shattering") point, randomly deleting edges, and study its properties at that point.

INTRODUCTION

Traditional data mining algorithms, such as association rule mining, market basket analysis, and cluster analysis, commonly attempt to find patterns in a single relation that stores a collection of independent instances. An emerging challenge for data mining is to tackle the problem of mining collections of inter-related instances, represented as graphs, usually spanning several relations.

Graphs are convenient representations of numerous settings, such as social networks, scientific publication networks, authors vs. conference participation, and others. Over the last decade, the number of data that should be represented as

DOI: 10.4018/978-1-4666-3604-0.ch026

graphs (e.g. social networks in LinkedIn, Facebook and Flickr) has increased exponentially (Newman, 2003). Therefore, graph mining has become an essential technique to extract knowledge from networks. Many fascinating non-intuitive network properties, such as small and shrinking diameter, degree distribution, triangles, eigenvalues, and community structures (Faloutsos et al., 1999, Leskovec et al., 2007b, Tsourakakis, 2008, Fortunato, 2010) have been discovered.

Network properties are important to understand the network behavior and formation. For example, if most of the nodes follow a specific pattern, the deviating ones can be outliers and should be more carefully studied. To develop graph analysis techniques, or to test new hypotheses and algorithms for graphs, it is often necessary to drill over synthetic graphs, where the target properties are known to occur within known parameters. Therefore, the generation of synthetic graphs is an important asset. Data generation to evaluate traditional data mining algorithms is a fairly well-understood technique, and several statistical methods have been long available. However, the development of synthetic graphs that resemble real ones requires great efforts and synthetic graph generation is still an open research issue (Chakrabarti et al., 2004). To improve graph generators, especially to generate well-suited graphs to help evaluate social network analysis tools, it is important to identify and describe properties that can distinguish real from synthetic networks.

Spotting synthetic data is another important research topic, since synthetic networks can be generated or mixed with real networks to hide or fake information. For instance, in a network for product recommendation or reliability, unethical participants might try to insert few, well-designed synthetic fake nodes, skewing the scores reliability.

In this context, this chapter presents techniques and properties, such as node degree distribution, number of triangles, adjacency matrix eigenvalues and others (Faloutsos et al., 1999, Tsourakakis, 2008, Leskovec et al., 2007b) developed to mine graphs. However, these traditional properties used alone do not allow distinguishing synthetic network from real ones. For example, if one compares the degree distribution of a real network and a synthetic network generated by preferential attachment, both will have a power law degree distribution.

ShatterPlots (Appel et al., 2009) is a technique that allows distinguishing between real and synthetic graphs. Its process is based on network resilience and randomly removes edges from a network until it reaches a state, known as Shattering point, where the network reaches its largest effective diameter, that is, the node reachability is at its worst point. At this point, it is possible to extract interesting patterns for the number of nodes and edges and the triangles distributions, as well as properties of connected components and adjacency matrix eigenvalues. We will show that the combination of all these properties helps us separate real networks from synthetic ones.

Traditionally, only the node degree, the distributions of the number of triangles and diameter properties have been used to distinguish between synthetic and real networks. Those measurements are generally enough to compare real and random graphs (Erdos and Renyi, 1960), since those characteristics can tell random graphs apart. However, more elaborate network generators, such as the Small World (Watts and Strogatz, 1998) and the Preferential Attachment (Barabasi and Albert, 1999) techniques can closely mimic these properties following a power law (Clauset et al., 2009), enforcing a large number of triangles and a small diameter and better resembling the real networks. Spotting real graphs from synthetic ones created by those newer generators is therefore a tough enterprise. Forcing a graph to a critical state, as the ShatterPlots technique does, and measuring its properties at that state reveals the inner nature of the graph, helping to identify its inherent constitution.