Physica A 444 (2016) 360–372

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Summarizing scale-free networks based on virtual and real links

Yijun Bei [a], Zhen Lin [b,c,*], Deren Chen [c]

[a] College of Software Technology, Zhejiang University, Hangzhou 310027, China
[b] Technology Development Department, Shanghai Clearing House, Shanghai 200002, China
[c] College of Computer Science & Technology, Zhejiang University, Hangzhou 310027, China

HIGHLIGHTS

- Attributes and structural similarities are integrated into a unified framework.
- A novel summarizing graph based on virtual and real links is proposed.
- A centroid selecting approach is presented to partition large virtual graphs.
- HB-Graph is adopted to optimize grouping results when adjusting subgroups.
- SGVR allows users to view and analyze graphs from different granularities.

ARTICLE INFO

Article history:
Received 19 January 2015
Received in revised form 16 June 2015
Available online 20 October 2015

Keywords:
Graphs
Real relationships
Scale-free networks
Summarizing
Virtual relationships

ABSTRACT

Techniques to summarize and cluster graphs are indispensable to understand the internal characteristics of large complex networks. However, existing methods that analyze graphs mainly focus on aggregating strong-interaction vertices into the same group without considering the node properties, particularly multi-valued attributes. This study aims to develop a unified framework based on the concept of a virtual graph by integrating attributes and structural similarities. We propose a summarizing graph based on virtual and real links (SGVR) approach to aggregate similar nodes in a scale-free graph into k non-overlapping groups based on user-selected attributes considering both virtual links (attributes) and real links (graph structures). An effective data structure called HB-Graph is adopted to adjust the subgroups and optimize the grouping results. Extensive experiments are carried out on actual and synthetic datasets. Results indicate that our proposed method is both effective and efficient.

1. Introduction

In recent years, numerous scale-free networks, such as social networks, protein networks, and sensor networks, have increasingly emerged. The effective and efficient identification of the underlying characteristics of these complex networks has become a challenge in research. In many applications, these networks are very large and often include thousands or even millions of nodes and edges. It is almost impossible for users to understand the underlying information by merely visualizing or browsing the network itself. Thus, graph summarization or node aggregation approaches are indispensable to analyze these large and complex networks.

* Correspondence to: Technology Development Department, Shanghai Clearing House, No. 2 East Beijing Road, Shanghai 200002, China.
E-mail addresses: byj@zju.edu.cn (Y. Bei), linzhen@shclearing.com, nblinz@gmail.com (Z. Lin), drc@zju.edu.cn (D. Chen).

http://dx.doi.org/10.1016/j.physa.2015.08.048
0378-4371/

Clustering and partitioning are two important learning techniques that are widely used to analyze networks [1–3]. Summarization is another effective approach to solve the problem of graph information mining. Most existing graph summarization methods [4], such as degree distribution, clustering coefficient, and hop plots, consider structural information and disregard the fact that nodes may contain multi-valued attributes in numerous real applications. However, some approaches [5] aim to summarize graphs by using only attribute information. Node attributes and graph topological structures are equally important in aggregating nodes. A perfect model should summarize graphs where nodes inside the same group have similar attributes and are closely connected in structure by considering the attributes and structural similarities. Thus, the problem of summarizing graphs, which contain multi-valued attributes, is quite challenging because attributes and structural similarities seem to be independent goals.

This study aims to develop a unified framework based on the concept of a virtual graph by integrating the attributes and structural similarities. We also propose an approach called summarizing graph based on virtual and real links (SGVR) to aggregate similar nodes in a scale-free graph into k non-overlapping groups using user-selected attributes and considering both virtual links (attributes) and real links (structures). The proposed methods allow users to dynamically roll-up or drill-down graph summaries, and thus analyze and visualize graphs from different granularities.

The main contributions of this paper are as follows:

1. The SGVR approach is proposed to summarize groups based on the multi-valued attributes and graph structure. We also introduce two types of node relationships, namely, virtual link and real link, and build a virtual graph for nodes by using virtual links for convenience in grouping.
2. A method is presented for meaningful centroids that are chosen to effectively partition large virtual graphs into smaller ones. To optimize the grouping results, efficient subgroup adjustment methods are presented using an effective data structure called HB-Graph. We achieve multi-resolution summaries by utilizing user-selected attributes and stack-based recording technique.

The remainder of this paper is organized as follows. Section 2 discusses related works. Section 3 introduces the virtual graph model and describes node-aggregating approaches based on the virtual graph. Section 4 introduces the SGVR algorithm and the node adjustment strategies for group optimization. Section 5 presents an evaluation of the summarizing algorithm. Section 6 concludes the study.

2. Related works

Frequently occurring mining substructures can help users gain insight into graphs and understand the characteristics of large graphs [6]. However, the usability of patterns hinders effective analysis because of the overwhelming number of frequent patterns. Although a few studies have been carried out to relax the rigid structural requirement or to reduce output cardinality for improving the representative quality [7–11], complexity in retrieving direct information from the patterns still exists. Graph mining mainly focuses on searching the problem pattern from a large set of small graphs, whereas our approaches deal with a single large complex graph.

Clustering graphs is another effective technique to understand graphs [1,2,12–17]. However, these existing approaches mainly employ a topological structure in clustering evaluation, such that similar elements belong in the same group, whereas dissimilar elements are found in different groups. Newman et al. [18] introduce a quality measure called modularity, which is computed over the entire clustering at each iteration to estimate when clustering should stop. Xu et al. [2] propose an algorithm called SCAN that detects clusters, hubs, and outliers by using structural density. A hierarchical model based on a multilevel geodesic approximation is adopted [1] to understand scale-free networks. The clustering approaches mentioned earlier only consider the graph structure or node attributes and disregard other parameters. Zhou et al. [3,19] recommend a clustering algorithm called SA-Cluster, which considers both structural and attribute similarities through a unified distance measurement by inserting attributed nodes. The GraphScope method proposed in Ref. [15] can discover meaningful time-evolving communities in large and dynamic graphs through information theoretic principles. Sun et al. [16] suggest a clustering framework, called RankClus, that directly generates clusters integrated with ranking in a multi-typed information network. To further understand graph clustering approaches, a detailed introduction about the definitions and methods for graph clustering is presented in the survey by Schaeffer et al. [14]. Satuluri et al. introduce an efficient and effective localized graph sparsification method by retaining only a fraction of the original number of edges, which increases the speed of graph clustering algorithms without compromising quality [13].

Most existing methods to summarize graphs obtain summaries through statistical tests that contain limited information. However, these summarizing processes are difficult to control. OLAP-style aggregating methods are considered to carry out the multi-dimensional analysis of graph data [20,21,3]. Chen et al. [22,20] introduce an OLAP-based model and propose a novel graph framework. These approaches consider different semantics of OLAP operations for graphs and classify the framework into two types, namely, informational and topological frameworks. These methods only present the concept of the graph OLAP; and thus far, intensive studies have yet to be conducted. Tian et al. [21] also propose OLAP-style aggregation methods called SNAP and k-SNAP for graph analysis. Although the methods in literature [21] can control the resolutions of summaries, they can only summarize the graphs in one direction because of their selection strategy for group splitting. Cheng et al. [3] propose the SA-Cluster algorithm that uses a unified distance measure. However the efficiency of this algorithm is low because the algorithm is carried out iteratively.