Information Sciences 339 (2016) 369–394

Contents lists available at ScienceDirect

Information Sciences

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

Efficient pattern matching on big uncertain graphs

Ye Yuan[a,*], Guoren Wang[a], Lei Chen[b], Bo Ning[c]

[a] Department of Computer Science, Northeastern University, Shenyang 110816, China
[b] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China
[c] School of Information Science and Technology, Dalian Maritime University, Dalian, China

ARTICLE INFO

Article history:
Received 3 December 2014
Revised 28 November 2015
Accepted 26 December 2015
Available online 8 January 2016

Keywords:
Uncertain data
Graph data

ABSTRACT

A significant amount of research has been devoted to seeking efficient solutions to the problem of pattern matching over graphs. This interest is largely due to the many applications that require such efficient solutions, including protein complex prediction, social network analysis, and structural pattern recognition. However, in many real applications, the graph data are often noisy, incomplete, and inaccurate. In other words, there exist many uncertain graphs. Therefore, in this paper, we study pattern matching in the context of large uncertain graphs. Specifically, we want to retrieve all qualified matches of a query pattern in the uncertain graph. Though pattern matching over uncertain graphs is NP-hard, we employ a *filtering-and-verification* framework to speed up the search. In the filtering phase, we propose a *probabilistic matching tree* (PM-tree) built from match cuts obtained by a cut selection process. Based on the PM-tree, we devise a *collective pruning* strategy to prune a large number of unqualified matches. During the verification phase, we develop an efficient sampling algorithm to validate the remaining candidates. Extensive experimental results demonstrate the effectiveness and efficiency of the proposed algorithms. Finally, we show how our solution can be applied to querying knowledge graphs.

1. Introduction

Graphs constitute a generic data model with wide applicability in numerous domains, such as social networks, biological networks, and the World Wide Web. Indeed, it is often less complex for users to shoehorn semi-structured or sparse data into a vertex-edge-vertex data model than into a relational data model. Furthermore, it is also most natural for users to reason about an increasing number of popular datasets, such as the underlying networks of Twitter, Facebook, or LinkedIn, within a graph paradigm. Various types of queries over graph data have been investigated, such as subgraph search queries [62,69,73], shortest-path queries [6,23], reachability queries [34,57], and pattern matching queries [18,42]. Reachability, or shortest-path, queries focus on the relation between two vertices in a graph. On the other hand, pattern matching queries are concerned with the connectivity among sets of vertices. Thus, a pattern matching query is more informative than a simple shortest-path, or reachability, query. Furthermore, a pattern matching query can be answered in polynomial time [20], while processing a subgraph query is # P-complete [25]. Therefore, the database community has devoted considerable effort to the study of the pattern matching query problem [18–20,42,74].

* Corresponding author. Tel.: +8613940102867.
E-mail address: linuxyy@gmail.com, yuanye@ise.neu.edu.cn (Y. Yuan).

http://dx.doi.org/10.1016/j.ins.2015.12.034
0020-0255/

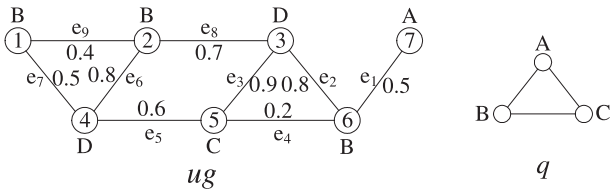

Fig. 1. Uncertain graph ug and query pattern q .

Interestingly, all of the aforementioned studies focus exclusively on applications where the edges of the graph are deterministic. Yet, in most applications, there is inherent uncertainty about the presence of edges due to often inevitable noise, incompleteness, and delays during data collection. For example, in protein–protein interaction (PPI) network, the proteins obtained from experiments may contain non-existing protein interactions, or on the contrary miss existing ones [10,28,52,54]; in social networks, graphs are often used to represent communities of users, where probabilities can be assigned to edges to model the degree of influence among users [1,40,46]; in communication or road networks, edge probabilities are used to quantify the connectivity between nodes, or to take traffic uncertainty into consideration [9,30]; finally, the uncertainty in an Resource Description Framework (RDF) graph is caused by data errors or semantic extraction inaccuracy in the data integration process [13,31,39].

Based on the above discussion, in this paper, we study pattern matching queries over large uncertain graphs. In the following, we describe the problem of probabilistic graph pattern matching and outline our contributions.

1.1. Probabilistic graph pattern matching

We first introduce graph pattern matching on deterministic graphs, and then proceed to discuss uncertain graph pattern matching.

Given a graph pattern query q with n vertices $\{v_1, \ldots, v_n\}$ and a deterministic graph g^c, a deterministic pattern matching query retrieves all matches of q in g^c. For a given q and an n-vertex set $m = \{u_1, \ldots, u_n\}$ in g^c, m is a match for q in g^c, if (1) the n vertices $\{u_1, \ldots, u_n\}$ in g^c have the same labels as the corresponding vertices $\{v_1, \ldots, v_n\}$ in q; and (2) for any two adjacent vertices v_i and v_j in q, the shortest-path distance between the two corresponding vertices u_i and u_j in g^c is no larger than a given threshold γ [19,74].

Example 1. Consider the pattern query q and the deterministic graph ug^c in Fig. 1. For this example the probabilities of each edge can be ignored. Let the weight of each edge be 1 and the distance constraint γ be 3. Vertices {2, 5, 7} or {5, 6, 7} form a match for q in ug^c, since their vertex labels are same as those of q, namely, {*A*, *B*, *C*}, and the shortest-path distance between each pair of vertices is less than 3. Though the vertex set {1, 5, 7} also has labels {*A*, *B*, *C*}, it is not a match because the shortest-path distance between vertices 1 and 7 is 4, which violates the distance constraint.

The semantics of pattern matching queries have many real life applications [19,20,74]. For example, suppose that Fig. 1 is a graph model of LinkedIn, where vertices represent active users and edges indicate the friendship relations among users. Job attributes are used to label the vertices, e.g., {*A*, *B*, *C*} = {Scientist, Professor, Student}. The pattern matching query q looks for relations among scientists, professors and students. Finding such patterns may help social science researchers discover close connections (due to the distance constraint) between a successful scientist and his/her circle of students or professors.

For the uncertain graph pattern matching problem, we focus here on *threshold-based probabilistic pattern matching* (T-PM) over large uncertain graphs, where vertices are deterministic and edges are uncertain. Specifically, let g be an uncertain graph, let q be a graph pattern query, and let ϵ be a probability threshold. A T-PM query retrieves all vertex sets $m = \{u_1, \ldots, u_n\}$ in g (i.e., n vertices in g), such that the *pattern matching probability* (PMP) of m in g is at least ϵ. We will formally define PMP later.

We employ the *possible world semantics* [53], which has been widely used for modeling query processing over uncertain databases, to explain the semantics of PMP. A *possible world graph* (PWG) of an uncertain graph is a possible instance of the uncertain graph. It contains all of the vertices and a subset of the edges of the uncertain graph, and its weight is the product of all probabilities associated with the edges. Then, for a graph pattern query q with n vertices $\{v_1, \ldots, v_n\}$ and an n vertex set $m = \{u_1, \ldots, u_n\}$ in an uncertain graph g, the probability of m being a match for q is the sum of the weights of those PWGs g', of g, where m is a match for q in g'. For m to be a match for q in g', it must satisfy the two conditions of deterministic graph pattern matching defined above.

Example 2. Fig. 2 shows a couple of the PWGs of the uncertain graph ug of Fig. 1 and their respective weights. There are altogether $2^9 = 512$ PWGs for ug, and the sum of all weights is 1. To decide if a vertex set $m = \{5, 6, 7\}$ is a match for q in the uncertain graph ug, we first find all of ug's PWGs that contain m as a match for q. Again, recall that m is a match for q in g' if (1) vertices in m and q have the same labels, and (2) each pair of corresponding vertices in m has a shortest-path distance of at most 3 ($\gamma = 3$). Here, the result includes both of the PWGs depicted in Fig. 2, as well as many others. Next,