Information Sciences 610 (2022) 549–570

Contents lists available at ScienceDirect

Information Sciences

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

Distributed graph pattern matching via bounded dual simulation

Sarra Bouhenni [a,b,c,*], Saïd Yahiaoui [b], Nadia Nouali-Taboudjemat [b], Hamamache Kheddouci [c]

[a] Ecole nationale Supérieure d'Informatique, BP M68, Oued Smar 16309, Algeria
[b] CERIST, Centre de Recherche sur l'Information Scientifique et Technique, Ben Aknoun 16030, Algeria
[c] Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France

ARTICLE INFO

Article history:
Received 21 January 2021
Received in revised form 19 May 2022
Accepted 7 August 2022
Available online 11 August 2022

Keywords:
Graph pattern matching
Subgraph matching
Massive graphs
Graph simulation
Distributed graph processing

ABSTRACT

Graph Pattern Matching (GPM) finds subgraphs of a large data graph that are similar to an input query graph. It has many applications, such as pattern recognition, detecting plagiarism, and finding communities in social networks. Current real-world applications generate massive amounts of data that cannot be stored on the memory of a single machine, which raises the need for distributed storage and processing. Recent relaxed GPM models, although of polynomial time complexity, are nevertheless not distributed by nature. Moreover, the existing relaxed GPM algorithms are limited in terms of scalability. In this paper, we propose Bounded Dual Simulation (BDSim) as a new relaxed model for a scalable evaluation of GPM in massive graphs. BDSim captures more semantic similarities compared to graph simulation, dual simulation, and even strong simulation. It preserves the vertices' proximity by eliminating cycles of unbounded length from the resulting match graph. Furthermore, we propose distributed vertex-centric algorithms to evaluate BDSim. We prove their effectiveness and efficiency through detailed theoretical validation and extensive experiments conducted on real-world and synthetic datasets. To the best of our knowledge, BDSim is the first relaxed GPM model that captures the cyclic structure of the query graph while being feasible in cubic time.

1. Introduction

Graph Pattern Matching (GPM) is the process of identifying subgraphs of a large data graph that answer a relatively small query graph. It is defined in terms of several models notably subgraph isomorphism, a well-studied problem but known to be NP-Complete [11]. Subgraph isomorphism requires strict constraints on the data graph to be met, which makes it impractical for the current real-world applications such as plagiarism detection an community detection in social networks. In contrast, relaxed GPM defines a new category of models answering GPM queries by relaxing the matching constraints imposed by subgraph isomorphism. In addition to the topological structure of the graph, these GPM models also consider the semantic information represented by labels on both vertices and edges. We find in this category graph simulation and its different extensions.

* Corresponding author at: Ecole nationale Supérieure d'Informatique, BP M68, Oued Smar 16309, Algerie.
E-mail addresses: cs_bouhenni@esi.dz (S. Bouhenni), syahiaoui@cerist.dz (S. Yahiaoui), nnouali@cerist.dz (N. Nouali-Taboudjemat), hamamache.kheddouci@univ-lyon1.fr (H. Kheddouci).

https://doi.org/10.1016/j.ins.2022.08.038
0020-0255/