Mining frequent subgraphs in multigraphs

Vijay Ingalalli [a,b,*], Dino Ienco [b], Pascal Poncelet [a]

[a] *LIRMM - University of Montpellier, France*
[b] *UMR TETIS, Irstea, Montpellier, France*

A B S T R A C T

For more than a decade, extracting frequent patterns from single large graphs has been one of the research focuses. However, in this era of data eruption, rich and complex data is being generated at an unprecedented rate. This complex data can be represented as a multigraph structure - a generic and rich graph representation. In this paper, we propose a novel frequent subgraph mining approach MuGraM that can be applied to multigraphs. MuGraM is a generic frequent subgraph mining algorithm that discovers frequent multigraph patterns. MuGraM efficiently performs the task of subgraph matching, which is crucial for support measure, and further leverages several optimization techniques for swift discovery of frequent subgraphs. Our experiments reveal two things: MuGraM discovers multigraph patterns, where other existing approaches are unable to do so; MuGraM, when applied to simple graphs, outperforms the state of the art approaches by at least one order of magnitude.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

Real world data can be easily modelled as a graph where entities are represented as nodes, and interactions between entities are represented as edges. When only one edge type is allowed between a pair of nodes, we refer to this graph structure as *single edge graph*; when more than one edge type is allowed between a pair of nodes, we refer to it as *multigraph*. Multigraph structure enables us to represent multiple relations between a pair of nodes [2,3].

Many real world datasets can be modelled as a network where a set of nodes are interconnected by multiple relations. Various domains are abound with multigraphs: social networks spanning over the same set of people, but with different life aspects (e.g., social relationships such as Facebook, Twitter, LinkedIn, etc.); protein-protein interaction multigraphs, where the protein pairs have direct interactions/physical associations or they are co-localised [1]; gene multigraphs, where genes are connected by different pathway interactions that belong to different pathways; *Resource Description Framework* knowledge graphs, where a subject/object node pair is connected by different predicates [17].

Since multigraphs allow more than one relation between a pair of nodes, we can represent real world data more succinctly, which in turn helps in mining patterns that cannot be discovered in the otherwise simple graphs. For example, a recent work in the field of bioinformatics [16] creates multigraphs by merging heterogeneous genomic and phenotype data, in order to identify the disease genes. Many such applications can be catalysed in order to mine interesting and useful patterns.

---

* Corresponding author at: LIRMM - University of Montpellier, France.
 E-mail addresses: vijay@lirmm.fr, vijay.ingalalli@inria.fr (V. Ingalalli), dino.ienco@irstea.fr (D. Ienco), pascal.poncelet@lirmm.fr (P. Poncelet).

(a) A sample data multigraph   (b) A set of frequent multigraph patterns
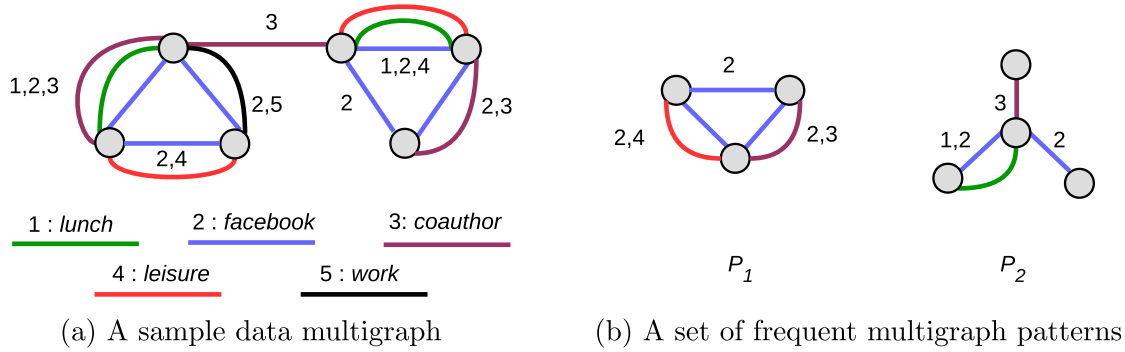
Fig. 1. (a) A data multigraph with unlabeled nodes; (b) a set of multigraph patterns with frequency $\delta = 2$.

One of the most important tasks in graph data management is *frequent subgraph mining* [6,11,13,14,22] where the problem is to discover patterns that occur *frequently* in a graph database. Although plenty of approaches exist to mine frequent patterns in single edge graph, to the best of our knowledge, no approach exists to mine frequent patterns in multigraphs.

Considering FSM in *single edge graph* data, the existing approaches can be categorized into two main families: (i) FSM for transactional graph setting, where the graph database consists of a set of relatively small sized graphs called transactions, and (ii) FSM for single large graph setting, where the graph database consists of a single large graph.

In *transactional graph* databases, a subgraph is frequent if it appears in at least δ transactions, where δ is a user defined frequency threshold value. Several works have been proposed to address FSM in transactional graph databases [11,13,22]. However, since the task of FSM in single large graph setting is more challenging than the transactional one, several approaches [6,14] have been proposed by considering various frequency (or support) evaluation measures [4,20].

This work is motivated by the fact that the existing FSM approaches cannot be applied to multigraph data. That is, when the graph data contains multiple relations between a pair of entities, the existing FSM approaches cannot discover frequent patterns that contain a subset of multiedges. Thus, whenever multiple relations (multiedges) exist between a pair of nodes, in order to use the existing FSM approaches, one has to map the multiple relations (multiedges) to a unique value (distinct edge label) and then perform FSM, which however, does not yield desirable results, thereby making the existing approaches rather incomplete. To the best of our knowledge, no existing work can discover frequent subgraphs in single large multigraphs. It is to be noted that one of the recent works called GRAMI [6] claims to handle multi-labeled graphs (which we refer to as multigraphs). However, neither they provide any details about managing multigraphs in their paper, nor their latest code is capable of handling multigraph data.

Let us consider a typical scenario of performing FSM on a multigraph as depicted in Fig. 1. The data multigraph in Fig. 1(a) is an extract of the real world AUCS dataset [12] that has five different relations (edge types) namely, *lunch, Facebook, coauthor, leisure, and work* which are defined among a set of university employees (nodes). If we perform FSM on this dataset by setting a frequency threshold $\delta = 2$, the existing FSM approaches output *no* patterns, since they treat a set of relations between a pair of nodes as a unique identifier, rather than treating it as a set of multiple relations. And thus, they are unable to discover those frequent patterns that are spanned from a subset of the relations, as depicted in Fig. 1(b).

The objective of the proposed work is to fill the gap in the field of FSM by proposing an approach to extract frequent patterns from multigraph data by considering patterns that can span over a subset of the multigraph relations. Thus, we propose MUGRAM (*Frequent **MultiGra**ph **M**iner*) - an algorithm that enumerates all frequent subgraph patterns in a single large multigraph. The major contributions of this work are:

- a set of efficient pruning rules to swiftly traverse the search space for multigraph pattern extraction;
- an efficient method to quickly evaluate the pattern support;
- a quantitative and qualitative evaluation of MUGRAM on real world graph data.

The experimental evaluation reveals that MUGRAM is not only an approach that can handle *multigraph* data efficiently but it also outperforms the state-of-the-art approaches in extracting frequent subgraphs in *single edge graph* data.

The rest of the paper is organized as follows. In Section 2, we discuss the related works. In Section 3, we introduce some basic definitions and formalize the problem. In Section 4, we discuss the proposed multigraph mining algorithm MUGRAM along with several optimization strategies. Detailed experimental evaluations are conducted in Section 5, followed by the conclusion in Section 6.

2. Related work

Several existing works address the problem of FSM for both transactional graph databases and single graph databases. For the transactional graph database setting, the work of Inokuchi et al. [11] has shaped the foundation for many later works. This work proposed an approach called AGM to efficiently mine the association rules among the frequently appearing