CSC17103 - Graph Mining Summer 2024 Homework 02
Teaching Assistant: Le Nhut Nam

Trong phần này, bạn sẽ lần lượt giải quyết các **Problem**. Bạn cần trình bày nó và submit lên hệ thống Moodle dưới dạng .pdf với tên MSSV.pdf, ví dụ: 12345678.pdf

Problem 1. *Thế nào là frequent patterns? Cho ví dụ? Tại sao chúng ta cần tìm frequent patterns trong dữ liệu?*

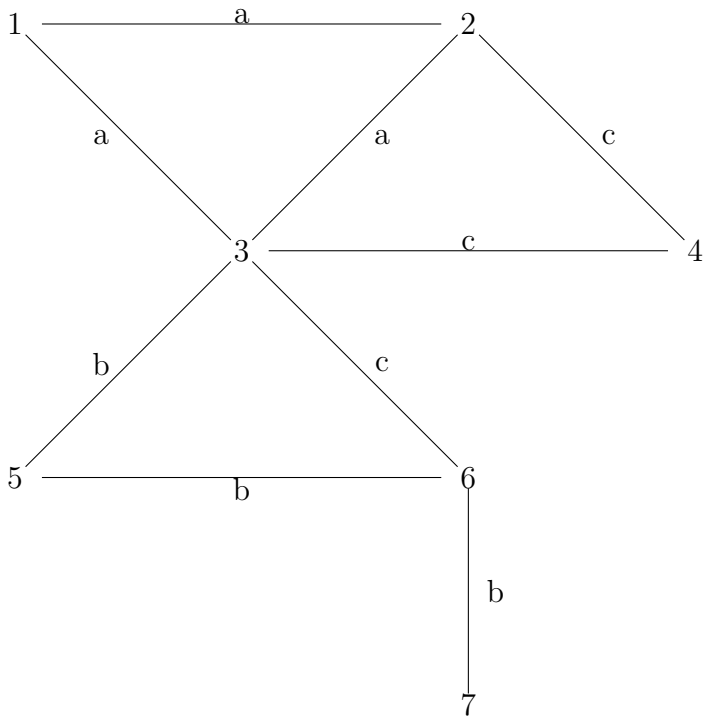
Problem 2. *Trình bày nguyên lý Apriori.*

Definition 1 (Frequent subgraphs)**.** *Một đồ thị (con) được gọi là frequent nếu support của nó (occurence frequency) trong một tập dữ liệu cho trước không nhỏ hơn một ngưỡng support bé nhất.*

Và chúng ta sẽ phát biểu bài toán *frequent subgraph mining* như sau:

Definition 2 (Bài toán Frequent Subgraph Mining (FSM))**.** *Tìm tất cả các đồ thị con của G mà xuất hiện ít nhất σ lần.*

Ví dụ: Ta xem xét dữ liệu đồ thị dưới đây.



Giả sử rằng $\sigma = 2$, frequent subgraph là (chỉ tính nhãn cạnh):

- a, b, c
- a-a, a-c, b-c, c-c
- a-c-a ...

Một vấn đề, số lượng patterns cấp số nhân!!!.

Như vậy, làm sao để khai thác frequent subgraphs? Chúng ta chia thành hai tiếp cận chính như sau:

1. Tiếp cận dựa trên Apriori
2. Tiếp cận dựa trên Pattern-growth

Đối với tiếp cận Apriori, chúng ta một số cách tiếp cận như sau:

1. AGM/AcGM [3]
2. FSG [4]
3. PATH#

Đối với tiếp cận Pattern-growth, chúng ta có một số cách tiếp cận như sau:

1. MoFa [9]
2. gSpan [10]
3. Gaston [9]
4. FFSM [9]
5. SPIN [2]

Problem 3. *Trình bày phương pháp của hai hướng tiếp cận Apriori và Pattern-growth. Mỗi hướng tiếp cận có những khó khăn gì?*

Bên cạnh khai thác mẫu phổ biến đồ thị con, chúng ta có một số loại mining khác như sau:

- Maximal frequent subgraph mining (MFSM): Một đồ thị được gọi maximal nếu không có các siêu đồ thị của nó là frequent.
- Closed frequent subgraph mining (CFSM): Một frequent subgraph được gọi là đóng, nếu tất cả các siêu đồ thị của nó nhỏ hơn (nghiêm ngặt) một frequency. Một số thuật toán cho hướng này: CloseGraph [11], SPIN [2], MARGIN [8].
- Significant subgraph mining (SSM): Đây là loại khai thác đồ thị con sử dụng một số kiểm định ý nghĩa (ví dụ như p-value). Một số thuật toán tiêu biểu: gBoost [7], gPLS [6], GraphSig [5].
- Representative orthogonal graphs mining (ROGM): Đây là loại khai thác đồ thị con với độ tương đồng bị chặn và chồng chéo tương ứng với các mẫu khác. Một số thuật toán tiêu tiểu: ORIGAMI [1].

Trong bài tập này, chúng ta quan tâm đến khai thác đồ thị con phổ biến đối với các đồ thị nhỏ. Ta phát biểu bài toán khai thác đồ thị con

Definition 3 (Khai thác đồ thị con).

- *Support: Cho trước một tập đồ thị có nhãn* $D = \{G_1, G_2, \ldots, G_n\}, G_i = \langle V_i, E_i, l_i \rangle$*, và một đồ thị con* G*, tập support của* G*, ký hiệu là* $D_G = \{G_i \mid G \sqsubseteq G_i, G_i \in D\}$*, trong đó* $G \sqsubseteq G_i$ *ám chỉ rằng* G *là đồ thị con đẳng cấu với* G_i*; Support được định nghĩa* $\sigma(G) = \dfrac{|D_G|}{|D|}$

- *Input: tập đồ thị có nhãn* $D = \{G_1, G_2, \ldots, G_n\}, G_i = \langle V_i, E_i, l_i \rangle$*, support nhỏ nhất* min_sup

- *Output: Một đồ thị con* G *là phổ biến nếu* $\sigma(G) \geq min\_sup$*; Mỗi đồ thị con là liên thông.*

Và bài toán khai thác mẫu đồ thị con phổ biến có đầu vào và đầu ra như sau:

- Đầu vào:
  - Tập hợp các đồ thị (gọi là cơ sở dữ liệu đồ thị).
  - Đồ thị đơn vô hướng (không khuyên, không đa cạnh).
  - Mỗi đồ thị có nhãn tương ứng với các đỉnh và cạnh của nó.
  - Các đồ thị có thể không liên thông.
  - Ngưỡng support nhỏ nhất min_sup
- Đầu ra:
  - Các đồ thị con phổ biến mà thỏa mãn ràng buộc support nhỏ nhất.
  - Mỗi đồ thị con phổ biến là liên thông.

Các tiếp cận mining thường gặp:

- Tiếp cận dựa trên Apriori: FSG [4]
- Tiếp cận dựa trên Pattern-growth: gSpan [10]
- Tiếp cận dựa trên tham lam: Subdue

Problem 4. *Dựa trên MSSV của bạn, nếu*

- *Chữ số tận cùng của MSSV là 0, 2, 4, 6, 8, bạn thực hiện trình bày lại FSG và cho ví dụ minh họa tính toán.*
- *Chữ số tận cùng của MSSV là 3, 5, 7, 9, bạn thực hiện trình bày lại gSpan và cho ví dụ minh họa tính toán.*

Và trình bày về tiếp cận tham lam Subdue. Phân biệt, đánh giá ưu điểm và nhược điểm của từng loại tiếp cận.

References

[1] M. Al Hasan, V. Chaoji, S. Salem, J. Besson, and M. J. Zaki. Origami: Mining representative orthogonal graph patterns. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 153–162. IEEE, 2007.

[2] J. Huan, W. Wang, J. Prins, and J. Yang. Spin: mining maximal frequent subgraphs from graph databases. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 581–586, 2004.

[3] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000 Lyon, France, September 13–16, 2000 Proceedings 4*, pages 13–23. Springer, 2000.

[4] M. Kuramochi and G. Karypis. An efficient algorithm for discovering frequent subgraphs. *IEEE transactions on Knowledge and Data Engineering*, 16(9):1038–1051, 2004.

[5] S. Ranu and A. K. Singh. Graphsig: A scalable approach to mining significant subgraphs in large graph databases. In *2009 IEEE 25th International Conference on Data Engineering*, pages 844–855. IEEE, 2009.

[6] H. Saigo, N. Krämer, and K. Tsuda. Partial least squares regression for graph mining. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 578–586, 2008.

[7] H. Saigo, S. Nowozin, T. Kadowaki, T. Kudo, and K. Tsuda. gboost: a mathematical programming approach to graph classification and regression. *Machine Learning*, 75:69–89, 2009.

[8] L. T. Thomas, S. R. Valluri, and K. Karlapalem. Margin: Maximal frequent subgraph mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):1–42, 2010.

[9] M. Wörlein, T. Meinl, I. Fischer, and M. Philippsen. A quantitative comparison of the subgraph miners mofa, gspan, ffsm, and gaston. In *Knowledge Discovery in Databases: PKDD 2005: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005. Proceedings 9*, pages 392–403. Springer, 2005.

[10] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 721–724. IEEE, 2002.

[11] X. Yan and J. Han. Closegraph: mining closed frequent graph patterns. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 286–295, 2003.