UNIVERSITÀ DEGLI STUDI DELL'AQUILA

DIPARTIMENTO DI INGEGNERIA E SCIENZE DELL'INFORMAZIONE E MATEMATICA

CORSO DI LAUREA IN INFORMATICA

Stima efficiente del numero di Network Motifs tramite Color-Coding e decomposizioni bilanciate

Relatore:

Dott. Stefano Leucci

Candidato:

Giulia Scoccia

Correlatore:

Prof. Guido Proietti

Matricola:

249503

ANNO ACCADEMICO 2019–2020

INDICE

1 Introduzione **2**

1.1 Contributo della tesi 3

1.2 Organizzazione del testo 4

CAPITOLO 1

INTRODUZIONE

I Motif, anche chiamati Graphlet o Pattern, sono piccoli sottografi indotti di un grafo, la conta dei motif è un problema ben noto del graph mining e dell'analisi dei social network. Dato un grafo in input, il problema richiede di contare la frequenza di tutti i graphlet di una certa taglia. Comprendere la distribuzione dei motif permette di avere una chiave della conoscenza delle interazioni tra le proprietà strutturali e i nodi del grafo. Fa luce sul tipo di strutture locali presenti nel grafo, che possono essere usate per una miriade di analisi. Poichè il conteggio dei graphlet può risultare computazionalmente impegnativo, di solito ci si accontenta di obiettivi meno ambiziosi. Uno di questi è la stima della frequenza: per ogni sottografo vogliamo stimare, nel modo più accurato possibile, la sua frequenza relativa rispetto a tutti i sottografi della stessa dimensione. Ancora meno ambiziosamente, visto che il numero di sottografi di una data dimensione cresce in modo esponenziale, si limita l'attenzione al problema della stima della frequenza relativa solo ai sottografi più ripetuti, ossia quelli che compaiono almeno una certa frazione di tempo. Ci sono due approcci per ottenere tali stime. Il primo è l'uso delle catene di Markov Monte Carlo, mentre il secondo è quello dell'uso del Color Coding, studi recenti mettono in luce e studiano le differenze tra i due approcci [1]. In questa tesi andremo a vedere il Color Coding, un'elegante tecnica randomizzata introdotta in [2], per determinare in maniera probabilistica percorsi e alberi in un grafo e più nel dettaglio una sua estensione molto interessante, l'algoritmo CC. Questo algoritmo fornisce

garanzie statistiche, forti e dimostrabili, per il problema dell'approssimazione dei conteggi esatti di graphlet, da cui le frequenze possono essere facilmente derivate, il suo utilizzo è molto utile applicato su grandi reti sociali [1, 3] CC si basa su due osservazioni chiave. La prima è che il Color Coding può essere usato per costruire "un'urna" astratta che contiene una sotto-popolazione di tutti i k-alberi di G. La seconda osservazione è che il compito di campionare k-graphlet, ossia graphlet con k nodi, può essere ridotto, con un overhead minimale, a campionare k-alberi, alberi con k nodi, dall'urna. Si può così stimare la conta dei motif in due step: la "fase costruttiva", in cui si crea l'urna da G e la "fase di campionamento", dove si campionano i graphlet dall'urna.

1.1 Contributo della tesi

In questo lavoro di tesi, l'attenzione è stata concentrata sulla ricerca della frequenza di k-treelet all'interno di grafi più o meno grandi. Per k-treelet , si intendono alberi indotti in un grafo con non k nodi. Per effettuare tale ricerca è stato necessario concentrarsi sulla fase costruttiva dell'algoritmo CC. Infatti, è stato visto in uno studio su una rete PPI (Protein-Protein Interaction) l'efficacia di questa fase per la ricerca della frequenza di particolari strutture biomolecolari (unicellulari e pluricelluri)[4]. La fase costruttiva, è descritta mediante una programmazione dinamica di tipo "top-down", è un processo inevitabile, che però richiede un grande impiego di tempo e spazio. Il lavoro svolto ha portato, per prima cosa ad un'implementazione, in Java. Il programma permette la ricerca delle occorrenze dei diversi k-treelet colorati, all'interno del grafo. Ogni albero con k nodi è rappresentato con una stringa binaria nella quale sono incluse tutte le informazioni, tra cui: colorazione, forma e fattore di bilanciamento (β). I colori devono essere necessariamente k colori differenti. Ogni albero è costruito in modo tale che i figli del nodo radice siano disposti in ordine non crescente. Il fattore di bilanciamento, β, garantisce l'unicità degli alberi. Il programma mantiene tutti questi k-treelet all'interno di una tabella, insieme alle proprie occorrenze. L'approccio alla costruzione della tabella dei treelet, a differenza dell'algoritmo originarale, è di tipo "bottom-up". Per rendere più efficiente il codice si è ricorsi all'utilizzo di Thread. In un secondo momento si è cercato

di ottimizzare l'approccio utilizzato e viene introdotto il concetto di decomposizioni bilanciate di un albero. Sfruttando queste decomposizioni si è riscritto il codice ottimizzato. Anche in questo caso, in fase implementativa viene adottato un approccio "bottom-up", contrariamente a quello "top-down" adottato in fase teorica.

1.2 Organizzazione del testo

La descrizione del lavoro è strutturata nel seguente modo. Nel capitolo 2 viene descritta la tecnica del color coding e il suo utilizzo per il conteggio degli alberi. Si vedrà l'algoritmo CC e la sua formulazione "top-down". Si discuterÃă la scelta di adottare un approccio "bottom-up" per l'implementazione e i suoi vantaggi. Nel capitolo 3 si vedranno le decomposizioni bilanciate e perchè aiutano per rendere l'algoritmo CC più efficiente. Anche in questo caso si discuterÃă sulle scelte effettuate in fase implementativa. Nel capitolo 4 si osserveranno i dati ottenuti dalle sperimentazioni dei diversi codici. Mentre il capitolo 5, tratterà le conclusione e i possibili approcci futuri.

BIBLIOGRAFIA

[1] Marco Bressan, Flavio Chierichetti, Ravi Kumar, Stefano Leucci, and Alessandro Panconesi. Motif counting beyond five nodes. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 12(4):1–25, 2018.

[2] Noga Alon, Raphael Yuster, and Uri Zwick. Color-coding. *Journal of the ACM (JACM)*, 42(4):844–856, 1995.

[3] Marco Bressan, Stefano Leucci, and Alessandro Panconesi. Motivo: fast motif counting via succinct color coding and adaptive sampling. *Proceedings of the VLDB Endowment*, 12(11):1651–1663, 2019.

[4] Noga Alon, Phuong Dao, Iman Hajirasouliha, Fereydoun Hormozdiari, and S Cenk Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(13):i241–i249, 2008.