1 Sommario

Questa tesi va a ripercorrere il lavoro svolto da Leonardi Stefano durante il tirocinio svolto all'interno dell'università, con la supervisione del dottorando Sheikh Nasrullah, per il professore Montresor Alberto.

1.1 Introduzione

L'intero tirocinio è basato sull'articolo intitolato "Community-enhanced Network Representation Learning for Network Analysis" che può essere abbreviato con la sigle CNRL. Questo articolo propone un innovativo metodo per individuare delle comunità all'interno dei grafi. Lo scopo di questo tirocinio è stato quindi quello di replicare i risultati ivi proposti, una volta compreso il codice alla base, e dove possibile cercare di migliorare le prestazioni dell'algoritmo proposto, anche sfruttando elementi magari non considerati dal metodo originale.
L'algoritmo proposto nell'articolo sfrutta diverse tecniche, alcune delle quali dettagliatamente spiegate negli articoli:

- "DeepWalk: Online Learning of Social Representations"
- "node2vec: Scalable Feature Learning for Networks"
- "How exactly does word2vec work?"

Di seguito la spiegazione del perché l'individuazione di comunità è una tecnica così importante.

1.2 Cosa sono i grafi e cosa le comunità

Alla base di tutto vi sono i grafi, una particolare struttura dati composta da nodi e archi. I primi possono rappresentare un entità eventualmente anche con l'aiuto di attributi. Diversamente gli archi sono dei collegamenti che legano due nodi, questi possono essere interamente orientati o non orientati. Questa struttura dati permette di rappresentare tantissime situazioni per tale motivo è così importante. Lo scopo dell'individuazione di comunità è per l'appunto come dice il nome, scovare delle comunità, ossia dei gruppi di nodi simili fra loro.
È importante notare che i nodi possono essere simili per molte diverse ragioni, possono avere le stesse caratteristiche strutturali in quanto sono dei catalizzatori che legano molti altri elementi, magari dei punti isolati da tutto il resto, o più semplicemente hanno valori degli attributi analoghi.
Trovare nodi simili permette di gestirli in maniera omogenea e di sfruttarne le peculiarità. Visto che con i grafi possono rappresentare molte situazioni anche le comunità risultano essere molto versatili, per tale motivo sono tanto importanti.

1.3 Come si valutano le comunità

È stato spiegato cosa sono i grafi e cosa sono le comunità che vi si possono costruire sopra, ma non è stato spiegato come si può decidere, date due partizioni differenti su un grafo, quale di queste due è la migliore. Intuitivamente se il grafo è molto piccolo tanto da poterlo comprendere guardando unicamente una sua rappresentazione grafica, e le comunità individuate sono poche allora si potrebbe decidere anche ad occhio qual è la migliore partizione.
Se però le comunità individuate sono tante e il grafo non è più piccolo, allora non è più così facile effettuare una scelta e tanto meno riuscire a giustificarla. Esistono perciò dei metodi formali per analizzare e decidere cosa andare a preferire, questi metodi prendono il nome di modularità.

Prima di parlare di ciò è però necessario formalizzare il concetto di partizione per comprendere di cosa si sta parlando:

- Una partizione è un insieme di comunità
- Non tutte le comunità hanno le stesse dimensioni, ossia lo stesso numero di nodi che vi appartengono
- Le comunità non hanno né una dimensione massima né una dimensione minima.
  In realtà si può considerare che la loro dimensione cade nell'intervallo $[1, n]$, dove il limite inferiore 1 è dato dal fatto che devono contenere almeno un elemento, mentre il limite superiore n è dato dal fatto che non può contenere più nodi di quelli esistenti nel grafo (che sono in numero di n)
- È possibile che un nodo non appartenga a nessuna comunità
- Un nodo del grafo può appartenere a più di una comunità, potenzialmente anche a tutte quelle presenti, questo significa che c'è una sovrapposizione di comunità

Come accennato la metrica di valutazione di una partizione è la modularità (il cui simbolo è Q). Questo nome può essere associato a diverse metriche infatti a seconda di cosa cerchiamo avremo interesse a valorizzare certi aspetti e penalizzarne altri, per tale motivo la modularità è solo il nome usato per indicare la valutazione di una partizione.
Durante il tirocinio abbiamo adottato tre metodi differenti: mod_withMax, mod_overlap e modified modularity. I primi due nomi non sono ufficiali in quanto li abbiamo scelti al solo scopo di riconoscerle, mentre l'ultimo modified modularity è quello utilizzato nell'articolo su CNRL.
I dettagli del funzionamento e del perché le abbiamo scelte sono all'interno dell'apposito capitolo sulle metriche di valutazione.

1.4 I cambiamenti apportati

Per calcolare la partizione il codice di CNRL utilizza un sistema di visite random e grazie a questo poi ricostruisce le possibili somiglianze fra i nodi. Non volendo toccare questa sezione le modifiche da noi riportate si inseriscono all'interno della creazione grazie a tali visite di cammini.
Una cammino è definito come una sequenza di nodi, preso un elemento qualsiasi fra questi deve esser possibile arrivare all'elemento successivo attraversando un solo arco, ossia i due elementi devono essere direttamente connessi, almeno nel senso di percorrenza.
L'algoritmo genera x cammini, ognuno di lunghezza massima l, sul grafo, dove $x = n * w$:

- n: è il numero di nodi del grafo
- w: è il numero di cammini che si fanno partire da ogni singolo nodo
- l: è la lunghezza massima di ogni cammino

Con questi dati l'algoritmo di CNRL calcola la suddivisione in comunità, si può notare come tutti questi dati vengono estratti esclusivamente dalla struttura del grafo, ossia da come nodi ed archi sono interconnessi.
L'intuizione sfruttata per modificare l'algoritmo sta nel fatto che c'erano altri dati non utilizzati, ossia gli attributi dei nodi. Tramite la creazione di nuovi particolari cammini è possibile far risultare due nodi in precedenza lontani in quanto non direttamente connessi vicini, in quanto questi due nodi condividono uno o più attributi. È facile intuire che l'introduzione di tutti questi nuovi dati porta ad un radicale cambiamento nella suddivisione in comunità.

1.5 Risultati

Parte temporaneamente mancante...