Majority cascade in cost networks

Moscato Gabriele, Cacciapuoti Antonio

1 Problema

Il progetto affronta un problema classico della teoria dei grafi legato alla diffusione dell'influenza in reti sociali o tecniche. Un grafo G = (V, E) è una struttura composta da un insieme V di nodi e un insieme E di archi che connettono coppie di nodi. In questo contesto, ogni nodo $v \in V$ può rappresentare un'entità come un individuo in una rete sociale o un server in una rete informatica, mentre gli archi rappresentano le connessioni o le interazioni tra queste entità. Il focus del progetto è un processo di Majority Dynamical Influence Diffusion, che modella la diffusione dell'influenza attraverso la rete. Dato un sottoinsieme di nodi, detto Seed Set $S \subseteq V$, il processo si sviluppa iterativamente attraverso una sequenza di sottoinsiemi $\text{Inf}[S, 0], \text{Inf}[S, 1], \ldots, \text{Inf}[S, r]$, dove ogni $\text{Inf}[S, i]$ rappresenta l'insieme dei nodi attivati al passo i. In particolare, il processo parte con $\text{Inf}[S, 0] = S$ e prosegue attivando i nodi i cui vicini hanno già raggiunto una soglia di influenza. Formalmente, un nodo $v \in V$ viene attivato al passo r se soddisfa la condizione $|N(v) \cap \text{Inf}[S, r-1]| \geq \frac{d(v)}{2}$, dove N(v) è l'insieme dei vicini del nodo v e d(v) è il suo grado, ovvero il numero di archi incidenti su v.

L'obiettivo principale del progetto è quello di trovare un Seed Set ottimale da cui far partire il processo di diffusione, soggetto a un vincolo di costo. Il problema è dunque definito come segue: dato un budget K, si vuole individuare un sottoinsieme di nodi $S \subseteq V$, il cui costo totale C(S) sia minore o uguale a K. Il costo di un Seed Set è definito come la somma dei costi associati ai singoli nodi che lo compongono: $C(S) = \sum_{v \in S} cost(v)$. La funzione obiettivo è massimizzare il numero di nodi influenzati alla fine del processo, ovvero massimizzare la cardinalità di $\text{Inf}[S, r]$, il più grande sottoinsieme di nodi attivati alla conclusione del processo di diffusione.
Poiché il problema è NP-hard, trovare una soluzione esatta risulta computazionalmente intrattabile per reti di grandi dimensioni. Per questo motivo, è necessario procedere mediante l'uso di approcci euristici, che offrono soluzioni approssimative bilanciando efficacia e complessità computazionale.
Infine, un secondo obiettivo del progetto è quello di analizzare e confrontare le prestazioni dell'algoritmo sviluppato rispetto a due algoritmi predeterminati, noti nella letteratura. Questo confronto verrà effettuato in termini di efficacia nel massimizzare l'influenza nel rispetto del budget.

2 Rete

La rete sulla quale sono stati eseguiti gli algoritmi implementati e analizzati i relativi risultati è la rete di collaborazione Arxiv GR-QC (General Relativity and Quantum Cosmology) [1]. Essa proviene dall'archivio e-print arXiv e rappresenta le collaborazioni scientifiche tra autori di articoli presentati nella categoria Relatività Generale

e Cosmologia Quantistica. Se un autore i ha co-firmato un articolo con un autore j, il grafo contiene un arco non orientato che collega i a j. Nel caso in cui l'articolo sia co-firmato da k autori, si genera un sottografo completamente connesso su k nodi. I dati coprono gli articoli pubblicati nel periodo compreso tra gennaio 1993 e aprile 2003 (124 mesi). La raccolta inizia a pochi mesi dalla nascita dell'archivio arXiv, rappresentando quindi quasi l'intera storia della sezione GR-QC (al 2003). La rete conta 5242 nodi e circa 15.000 archi. Il diametro della rete è pari a 17, e la più grande componente connessa contiene 4158 nodi.
Si è valutata tale rete come particolarmente adatta per simulare un processo di Influence Diffusion. Infatti, si tratta di una rete di collaborazioni scientifiche, in cui i collegamenti tra gli autori indicano la co-autorship di articoli accademici. Essa quindi riflette dinamiche concrete e reali di come l'influenza si diffonde all'interno di una comunità collaborativa. Anche la dimensione della rete è stata un fattore preso in considerazione. Essa risulta sufficientemente complessa (essendo formata da più di 5000 nodi), ma non è tanto grande da essere computazionalmente difficile da trattare. Questa caratteristica ha permesso di sviluppare e testare i diversi algoritmi di diffusione dell'influenza con uno sforzo computazionale ed una complessità temporale ragguardevoli ma comunque sopportabili e gestibili.

3 Algoritmi Implementati

L'algoritmo implementato mira a trovare un Seed Set ottimale, cioè un insieme di nodi iniziali per avviare un processo di Influence Diffusion, rispettando un vincolo di budget. Il suo funzionamento si basa su una delle due seguenti misure di centralità:

- Between Centrality: $g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$
  dove σ_{st} è il numero totale di shortest path dal nodo s al nodo t e $\sigma_{st}(v)$ è il numero di tali path che passano per v;
- Degree Centrality: $C_D(v) = \deg(v)$
  definita come il numero di link incidenti sul nodo.

Esse rappresentano l'importanza dei nodi all'interno della rete. L'obiettivo è selezionare i nodi più influenti entro il limite di budget disponibile. Di seguito le varie fasi del funzionamento dell'algoritmo:

1. Filtraggio iniziale: il primo passo consiste nel generare una lista di nodi disponibili, escludendo quelli il cui costo supera il budget. La lista iniziale, chiamata 'available_nodes', contiene solo i nodi per cui il costo non eccede il budget. Questa operazione garantisce che solo i nodi accessibili in termini di budget vengano presi in considerazione.

2. Ordinamento dei nodi: l'algoritmo offre due modalità per ordinare i nodi:

   - Senza rapporto centralità/costo: Se l'opzione 'use_ratios' non è abilitata, i nodi vengono ordinati semplicemente in base alla misura di centralità selezionata. Questo significa che i nodi con il valore di centralità più alto vengono considerati prioritari, poiché hanno un'importanza maggiore nella diffusione dell'influenza.
   - Con rapporto centralità/costo: Se l'opzione 'use_ratios' è abilitata, i nodi vengono ordinati in base al rapporto tra la centralità e il costo. Questo metodo tiene conto non solo dell'importanza del nodo (attraverso la centralità), ma anche del costo associato, privilegiando i nodi che offrono un buon compromesso tra alta centralità e basso costo.

3. Generazione del Seed Set: dopo aver ordinato i nodi, l'algoritmo procede in modo iterativo per costruire il Seed Set. Inizia con un Seed Set vuoto e un costo totale pari a zero. Per ciascun nodo nella lista ordinata, l'algoritmo verifica se il suo costo può essere aggiunto al Seed Set senza superare il budget totale. Se l'aggiunta del nodo non eccede il budget, il nodo viene incluso nel Seed Set e il costo totale viene aggiornato. Questo processo continua fino a quando il budget viene raggiunto o superato.

4. Termine dell'algoritmo: l'algoritmo si interrompe quando non è più possibile aggiungere nuovi nodi senza eccedere il budget. A questo punto, il Seed Set generato viene restituito come output.

Si tenga in considerazione che nei risultati che seguitano con 'Algoritmo 3' si fa riferimento alla versione dell'algoritmo che utilizza la Degree Centrality come misura di centralità, mentre 'Algoritmo 4' utilizza la Betweeness Centrality. I risultati relativi ad entrambi gli algoritmi sono stati ottenuti tenendo disabilitata l'opzione 'use_ratios'.

4 Risultati

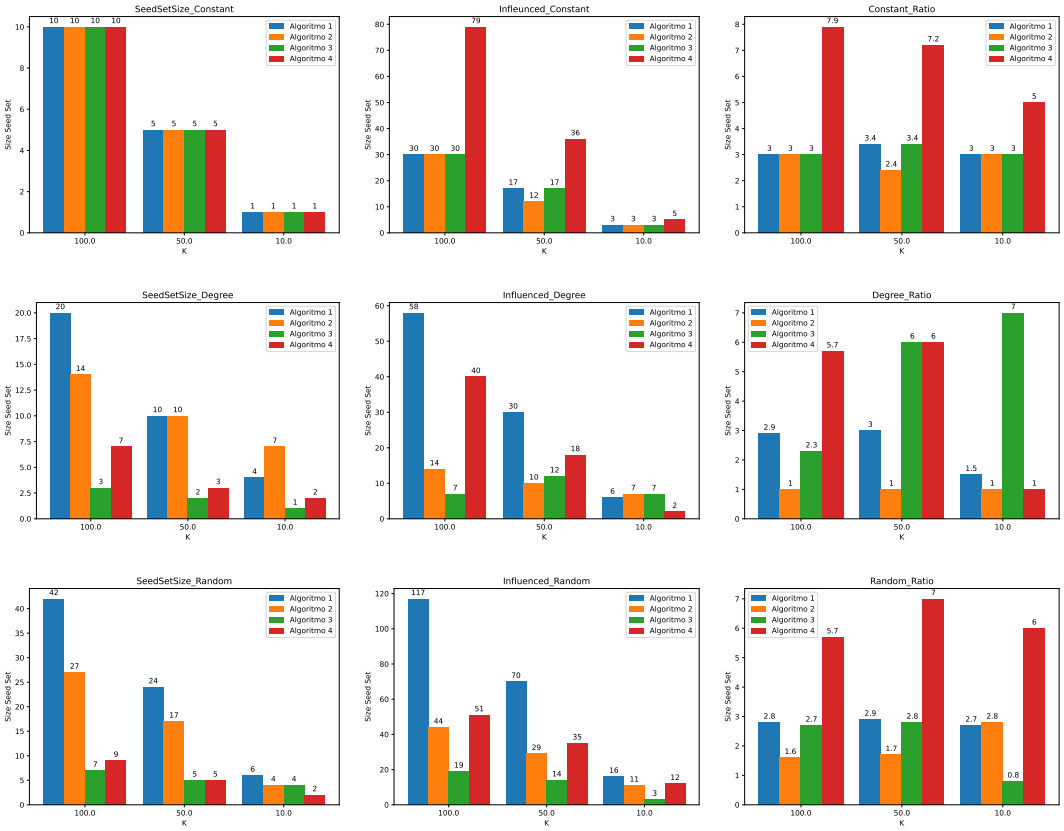
I risultati ottenuti dai test, riportati in nove grafici suddivisi in tre triplette, mostrano chiaramente che l'Algoritmo 4 è il migliore in termini di efficienza nell'influenzare nodi rispetto alla dimensione del Seed Set. Ogni tupla di grafici è relativa a una delle tre funzioni di costo utilizzate per definire il costo dei nodi: costo random, costo costante (pari a 10) e costo proporzionale al grado del nodo. Ogni tupla contiene un grafico che rappresenta la dimensione del Seed Set trovato per ciascun algoritmo, un grafico che mostra la cardinalità dell'insieme di nodi influenzati dal Seed Set e infine un grafico che raffigura il ratio nodi_influenzati/nodi_nel_seedset.
L'Algoritmo 4 ha ottenuto i risultati migliori nella maggior parte dei casi, registrando il ratio nodi_influenzati/nodi_nel_seedset più elevato per quasi tutti i budget testati. In particolare:

- Con costo costante, l'Algoritmo 4 ha raggiunto il ratio più alto (7.9) con un budget di 100.

- Con costo random, ha ottenuto il ratio maggiore (7.0) con un budget di 50.

- Con costo proporzionale al grado del nodo, ha avuto il miglior risultato con un ratio di 5.7 per un budget di 100, mentre l'Algoritmo 3 ha ottenuto il miglior ratio (7.0) con un budget di 10. Con un budget di 50, entrambi gli algoritmi (3 e 4) hanno raggiunto un ratio pari a 6.0.

È importante notare che l'Algoritmo 1 e l'Algoritmo 2 hanno sistematicamente generato i Seed Set più numerosi. Questo risultato potrebbe parzialmente spiegare il motivo per il quale hanno in seguito registrato i ratio più bassi. Ciò sottolinea che avere un Seed Set più grande non porta necessariamente a un risultato migliore, poiché l'efficacia di un Seed Set si misura anche in termini di efficienza, cioè la capacità di influenzare un gran numero di nodi con il minor numero possibile di nodi nel Seed Set.
Pertanto, sulla base dei risultati, l'Algoritmo 4 emerge come il più efficace nella diffusione dell'influenza, bilanciando al meglio la dimensione del Seed Set e la capacità di influenzare nodi nella rete.



5 Conclusioni

I risultati ottenuti dimostrano che, sebbene trovare soluzioni ottimali sia complesso, gli approcci euristici possono fornire soluzioni approssimative di qualità soddisfacente. Gli algoritmi utilizzati nel progetto hanno permesso di valutare l'efficacia del processo di diffusione di maggioranza seguendo diverse euristiche, in particolare valutandole in funzione del budget e dei costi associati ai nodi.
Il confronto tra diversi algoritmi ha mostrato come le varie soluzioni differiscono tra di loro, e quanto le diverse scelte in ambito di progettazione di un algoritmo possono influenzare in modo significativo l'efficacia di un algoritmo.
In generale, la scelta del seed set e il relativo costo giocano un ruolo fondamentale nella capacità di massimizzare la diffusione dell'influenza in una rete. Sebbene non esaminato in dettaglio in questo progetto, è evidente che la rete scelta e le sue caratteristiche influenzano in modo significativo l'efficacia degli algoritmi utilizzati.
In conclusione, possiamo dire che il problema della dominazione di maggioranza nelle reti è complesso e presenta molteplici variabili che influenzano la sua soluzione. La scelta dell'euristica più adatta dipende da diversi fattori, come il tipo di rete, il budget disponibile, i costi associati ai nodi e l'obiettivo specifico della diffusione. Pertanto, in fase implementativa, è cruciale valutare attentamente queste variabili per determinare l'approccio più efficace, bilanciando tra efficienza computazionale e qualità della soluzione.

Riferimenti bibliografici

[1] Jure Leskovec, Jon Kleinberg e Christos Faloutsos. “Graph evolution: Densification and shrinking diameters”. In: *ACM Trans. Knowl. Discov. Data* 1.1 (mar. 2007), 2–es. ISSN: 1556-4681. DOI: 10.1145/1217299.1217301. URL: https://doi.org/10.1145/1217299.1217301.