

UNIVERSITÀ DEGLI STUDI DI ROMA TOR VERGATA



MACROAREA DI SCIENZE MM. FF. NN.

Corso di laurea in informatica

TESI DI LAUREA TRIENNALE

**“Pesare grafi casuali con applicazione a tecniche
di diffusione”**

Relatore:

**Chiar.ma Prof.ssa
Miriam Di Ianni**

Laureando:

Lorenzo Cristofori

**Sessione invernale
Anno Accademico 2016/2017**

Indice

1	Introduzione	1
1.1	Struttura della tesi	3
2	Modelli per la generazione di grafi casuali	4
2.1	Il modello Erdős–Rényi	5
2.2	Il modello Rich Get Richer	6
3	Generazione di grafi casuali pesati	8
3.1	Erdős–Rényi per grafi pesati	8
3.2	Rich Get Richer per grafi pesati	10
4	Processi di diffusione	12
4.1	Diffusione nelle reti	12
4.2	Best Friend Target Set Selection	14
5	Analisi sperimentale dei processi di diffusione	16
6	Conclusioni	21
6.1	Sviluppi futuri	22

Sommario

Con l'avvento del Web lo studio delle reti sociali è divenuto di enorme importanza. D'altro canto queste reti sono difficili da reperire, ecco allora che viene in aiuto la teoria dei grafi aleatori tramite la quale si sono creati diversi modelli di generazione di grafi casuali come il modello Erdős–Rényi e Rich-Ger-Richer. In questa tesi verranno descritti e analizzati questi due modelli e verranno proposte delle modifiche degli stessi in modo da generare grafi aleatori pesati. Sarà poi presentato il problema della diffusione in una rete, approfondendo il modello Best-Friend Target Set Selection. Infine verranno fatti studi sperimentali sulla cascata applicata ai grafi aleatori ottenuti e ne saranno studiati i risultati.

Capitolo 1

Introduzione

Una rete sociale o social network rappresenta un qualsivoglia gruppo di individui connessi tra loro tramite diversi tipi di legami che si possono instaurare attraverso una conoscenza casuale o rapporti diretti quali quelli lavorativi o familiari. Le reti sociali vengono utilizzate per studi culturali in discipline come ad esempio la sociologia, l'antropologia e l'etologia. La diffusione del web e del termine social network ha creato, negli ultimi anni, alcune ambiguità di significato. La rete sociale è infatti storicamente, in primo luogo, una rete fisica. Esempi di reti sociali possono essere comunità di lavoratori, associazioni, circoli sportivi, mentre le versioni di Internet delle reti sociali sono di diverso tipo.

Queste infatti risultano essere reti molto più ampie di quelle fisiche. Basti pensare che se il numero di Dunbar o regola dei 150, afferma che le dimensioni di una rete sociale in grado di sostenere relazioni stabili sono limitate a circa 150 membri, ecco che nelle reti sociali del web questo non è più vero.

In generale, la teoria e i modelli usati per lo studio delle reti sociali sono compresi nella social network analysis. L'analisi delle reti sociali può essere condotta con un formalismo matematico usando la teoria dei grafi. Un grafo è un insieme di elementi detti nodi collegati tra loro da archi, più formalmente una rete è una coppia di elementi ordinati $G = (V, E)$ con V insieme dei nodi ed E insieme degli archi, dove E è un insieme di coppie di nodi ovvero $E \subseteq V \times V$. Un grafo può essere orientato o non orientato. Un grafo diretto è caratterizzato dalla presenza di un verso associato ad un arco; in altri termini in un grafo orientato (i, j) e (j, i) rappresentano due archi differenti, il primo esce dal nodo i ed entra nel nodo j nel se-

condo caso vale l'opposto. Ad ogni arco può essere inoltre associato un peso.

Grazie alla potenza di questo semplice modello è possibile rappresentare ogni tipo di rete sociale. In particolare, in una rete, ogni nodo rappresenta una persona ed esiste un arco dalla persona i alla persona j se i è amica di j . Infine il peso collegato all'arco può rappresentare il livello di amicizia tra due persone o nodi della rete.

Le social network sono fonti di enormi quantità di informazioni, basti pensare a Facebook, Instagram, Twitter, Google ecc.. Tutte queste società hanno accesso ad un'enorme quantità di dati, ma proprio a causa del valore associato a queste reti sociali diventa, nella maggior parte dei casi, difficile se non impossibile adoperare questi dati per studi scientifici di qualsiasi tipo. Diventa necessario quindi avere dei modelli con i quali generare delle reti che rispecchino quelle reali così da poterle sfruttare a pieno. Un punto fondamentale nello studio delle social network risulta essere valutare le azioni degli individui non come isolate ma con la consapevolezza che queste possono essere causate da qualche altro evento nella rete e abbiano un impatto sul resto degli individui.

Ad esempio gli attuali motori di ricerca come Google fanno un ampio uso della struttura della rete per valutare la qualità e la rilevanza delle pagine web. Per produrre dei buoni risultati di ricerca, questi siti valutano la rilevanza di una pagina web non semplicemente basandosi sul numero di link che puntano verso quest'ultima, ma anche da aspetti più sottili. Ad esempio, una pagina può essere considerata più prominente se riceve link da pagine che sono a loro volta prominenti. Quindi diventa importante conoscere non solo la quantità di collegamenti tra gli individui ma anche la tipologia di questi collegamenti.

Nelle reti sociali un cambiamento, un prodotto inserito nel mercato ad esempio, può portare a modifiche nella struttura della rete che non erano state previste inizialmente. Diventa quindi importante studiare come questi cambiamenti, che possono essere informazioni, prodotti, innovazioni, idee si diffondono nella rete e sotto quali condizioni un individuo decide di aderire al cambiamento proposto. Osservando una grande popolazione è possibile individuare dei pattern ricorrenti in costante evoluzione mentre nella società risulta evidente come alcune pratiche sociali possano diventare popolari o rimanere sconosciute.

Abbiamo quindi bisogno di avere delle reti che possano rappresentare al meglio i tipi di legami presenti tra gli individui della stessa, in quanto è proprio grazie a questi collegamenti che un

individuo sceglie o meno di aderire al cambiamento proposto.

1.1 Struttura della tesi

In questo lavoro di tesi saranno studiati alcuni tra i più importanti modelli di generazione di grafi, come il modello Rich-Get-Richer o ancora il modello Erdős–Rényi, ma dato che questi modelli sono pensati per generare solo grafi non pesati, saranno presentate delle modifiche a questi algoritmi affinché possano generare grafi capaci di mantenere informazioni sulla forza dei collegamenti presenti tra gli individui della rete.

Saranno poi introdotti i processi di diffusione, verrà in modo particolare presentata la diffusione dell'informazione in una rete e saranno studiati i modelli di diffusione. In particolare, verrà approfondito il Best Friend Target Set Selection. Verranno implementati tutti questi algoritmi, sarà fatta un'analisi sperimentale del modello di diffusione del BF-TSS sui grafi costruiti tramite i due algoritmi precedentemente accennati. Infine verranno mostrati e discussi i risultati ottenuti dalle analisi sperimentali.

Capitolo 2

Modelli per la generazione di grafi casuali

Potendo schematizzare una comunità di persone attraverso l'utilizzo di grafi, è possibile studiarla sotto differenti aspetti.

Si potrebbe pensare di usare reti reali presenti nel Web, ma questo è pressoché impossibile in quanto, se pensiamo alle reti di Facebook o Google e al loro impatto sulla società, risulta chiara la difficoltà nell'ottenerle. Inoltre reti di questo tipo risulterebbero troppo grandi da gestire e quindi complicate da usare nella fase di sperimentazione. Da ciò nasce l'esigenza di disporre di modelli di grafi capaci di descrivere particolari reti reali e tramite i quali è possibile simulare i nostri algoritmi. Dalla necessità di avere reti che rispecchino i modelli reali e che siano semplici da usare, nasce la teoria dei grafi aleatori.

La teoria dei grafi aleatori, che venne sviluppata la prima volta nel 1959 dai matematici Paul Erdos, Alfréd Rényi e Edgar Gilbert, può essere riassunta come un punto di incontro tra la teoria dei grafi e la probabilità. In questa teoria vengono studiati grafi generati casualmente secondo una qualche distribuzione di probabilità e ne vengono studiate le proprietà.

Nei prossimi due capitoli verranno presentati i modelli Erdős–Rényi e Rich-Get-richer per grafi aleatori.

2.1 Il modello Erdős–Rényi

Il modello Erdős–Rényi (ER) è uno dei modelli base per la generazione di grafi aleatori. Il grafo generato è composto da n nodi, quindi $V = v_1, v_2, \dots, v_n$, ogni coppia dei quali ha probabilità p di essere collegata da un arco e quindi, $\forall v_i, v_j P[(v_i, v_j) \in E] = p$.

Definizione 1. *$G(n, p)$ è un grafo aleatorio con n vertici, dove ogni possibile arco esiste con probabilità p .*

Da ciò risulta essere $p = 1/2$ il massimo grado di aleatorietà, in quanto, l'esistenza o l'assenza di un arco, sono equiprobabili. Con p piccolo si ha, invece, un grafo sparso e con p alto si ottiene un grafo denso. Analizzando questo modello, se indichiamo con d_i la variabile aleatoria che definisce il grado uscente del nodo i , allora si ha che il valore atteso $E[d_i] = \sum_{j=1, j \neq i}^n P((v_i, v_j) \in E) = p(n - 1)$.

Vogliamo ora calcolare la probabilità che il grado sia esattamente k , ossia $P(d_i = k) \quad \forall k = 0, 1, \dots, n - 1$. Il grado di un nodo è k quando esistono k vicini per cui esiste l'arco e per gli altri $n - 1 - k$ l'arco non c'è, quindi per k volte è stato creato l'arco e per $n - 1 - k$ volte no.

$$P(d_i = k) = \binom{n-1}{k} p^k (1-p)^{n-k-1} = \frac{(n-1)(n-2)\dots(n-k)}{k!} p^k (1-p)^{n-k-1}$$

Prendere p costante non risulta essere una buona scelta, infatti, in reti di grandi dimensioni, una probabilità costante diventa irrealistica. Usando frazioni indipendenti dalla grandezza della rete, si generano grafi con valori dei gradi attesi dei nodi troppo grandi per rispecchiare quelli reali. Basti pensare a Facebook, quanta può essere la probabilità che esista un individuo che conosce $1/2$ o $1/3$ o una frazione costante di tutti gli utenti di Facebook? Conviene quindi scegliere p in modo che sia inversamente proporzionale alla grandezza della rete e quindi avremo $p = \frac{\lambda}{n}$ per una certa costante λ . Sostituendo il nuovo valore di p nella formula precedente otteniamo:

$$P(d_i = k) = \frac{(n-1)(n-2)\dots(n-k)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k-1} \approx \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k-1} \approx \frac{\lambda^k}{k!} e^{\frac{-\lambda}{n}(n-k-1)} \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

da cui utilizzando l'approssimazione di Stirling per la quale $x! \simeq \sqrt{2\pi x} \left(\frac{x}{e}\right)^x$ otteniamo:

$$P(d_i = k) \simeq \frac{\lambda^k}{k!} e^{-\lambda} \simeq \frac{1}{\sqrt{2\pi k}} \frac{e^k}{k^k} \lambda^k e^{-\lambda} = \frac{e^{-\lambda}}{\sqrt{2\pi k}} \left(\frac{e\lambda}{k}\right)^k \simeq \frac{1}{k^k}$$

Questo risultato ci porta ad affermare che la probabilità che esista un nodo con grado k decresce esponenzialmente in k . Siamo quindi interessati a sapere qual è la frazione di nodi la

cui popolarità è proprio k . Per fare ciò definiamo una variabile aleatoria x_i che assume valore

1 se $d_i = k$ altrimenti è uguale a 0.

$$X = \sum_{i=1}^n x_i = \text{numero di nodi con popolarità } k$$

$$E[X] = E\left[\sum_{i=1}^n x_i\right] = \sum_{i=1}^n E[x_i] \simeq \sum_{i=1}^n [1(\frac{1}{k^k}) + 0(1 - \frac{1}{k^k})] = n \frac{1}{k^k}$$

Quindi la frazione di nodi con popolarità k è:

$$\frac{E[X]}{n} = \frac{n}{n} \frac{1}{k^k} = \frac{1}{k^k}$$

2.2 Il modello Rich Get Richer

Il comportamento o le decisioni di ogni persona sono influenzati dalle scelte fatte da altri. Questo perché il guadagno di alcuni dipende dalle scelte prese da altri, o anche perché queste scelte possono trasmettere informazioni in grado di aiutare terze persone a prendere determinate decisioni.

Questo approccio è stato utilizzato per analizzare il concetto di popolarità nelle reti. Mentre la maggior parte degli individui conosce solo persone nelle proprie cerchie sociali, alcuni individui raggiungono una più ampia visibilità. Con lo scopo di individuare con quale probabilità avviene ciò, è stato analizzato il web, in particolare i collegamenti tra le pagine.

Verrebbe naturale pensare che il numero di pagine "famose" sia una frazione molto piccola rispetto al numero totale di pagine e quindi che le stesse possano essere considerate rare. Da ciò è nata l'ipotesi della distribuzione normale. Nel caso in cui la distribuzione delle pagine segua effettivamente la distribuzione normale, le pagine con k in-links dovrebbero decrescere esponenzialmente in k al crescere di k .

Quando si è analizzato il Web si è però visto che le pagine con k in-links sono in effetti più comuni di come la distribuzione normale lasciasse credere. In particolare, le pagine con k in-links, sono approssimativamente proporzionale a $1/k^2$. Una funzione che decresce in k per una qualche potenza, come $1/k^2$, è detta *PowerLaw*. Per cogliere questa diffusione di nodi popolari in un grafo sociale è stato introdotto il Rich-Get-Richer Model. Questo modello spiega come poter generare dei grafi random la cui distribuzione di nodi "famosi" segua la Power Law precedentemente introdotta.

La regola che guida la creazione di un grafo secondo il Rich Get Richer Model può essere riassunta in questo modo:

1. I nodi sono creati in ordine, e chiamati 1,2,...,n
2. Durante la creazione di un collegamento uscente da un nodo u viene scelto, con probabilità P , se eseguire il passo a) o il passo b)
 - (a) Un nodo u sceglie con probabilità uniforme un nodo v e viene creato l'arco diretto (u, v) .
 - (b) Un nodo u sceglie con probabilità uniforme un nodo v e viene creato un arco uscente da u diretto ad un nodo z tale che esiste l'arco (v, z) .

Il procedimento sopra descritto può essere iterato per creare più archi uscenti da uno stesso nodo. Se ci troviamo ad una certa iterazione i e scegliamo il punto 2), una volta scelto in modo casuale un nodo u avremo a disposizione $i - 1$ archi uscenti da u tra cui scegliere con probabilità uniforme.

Se indichiamo con $j \rightarrow i$ l'arco uscente dal nodo j ed entrante in i , la probabilità che venga creato può essere così calcolata.

Definiamo una variabile a_{ji} che assume il valore 1 se esiste l'arco $j \rightarrow i$ e 0 altrimenti.

$$\begin{aligned} P(j \rightarrow i) &= \frac{p}{j} + (1-p)P\left[\bigcup_{h=0}^{j-1} (j \text{ sceglie } h \wedge h \rightarrow i)\right] = \\ &= \frac{p}{j} + 1 - p \sum_{h=0}^{j-1} (P(j \text{ sceglie } h))(P(h \rightarrow i | j \text{ ha scelto } h)) = \frac{p}{j} + \frac{(1-p)}{j} \sum_{h=0}^{j-1} a_{ji} \end{aligned}$$

Nell'ultimo termine dell'equazione possiamo vedere che la probabilità dell'esistenza di un arco è data dalla somma delle due componenti $\frac{p}{j}$ e da $\frac{(1-p)}{j} \sum_{h=0}^{j-1} a_{ji}$. La prima componente è quella che viene chiamata componente Erdős-Rényi mentre la seconda componente è la componente Rich-Get-Richer. Da ciò segue che se p è molto vicino ad 1 vince la componente Erdős-Rényi mentre quando p è prossimo a 0 ci sarà predominanza della seconda componente.

Capitolo 3

Generazione di grafi casuali pesati

Si vogliono generare dei grafi casuali che rappresentino delle possibili reti sociali reali ma che, oltre alla struttura della rete, diano un peso ad ogni arco in modo da diversificare i collegamenti. In effetti, in una rete non è importante solo studiare la quantità di collegamenti tra i nodi, ma anche la tipologia di questi ultimi.

Se pensiamo ad una rete sociale, ogni individuo avrà diversi tipi di legami con i propri vicini, ad esempio una persona avrà dei migliori amici, famigliari, conoscenti, colleghi di lavoro; è utile quindi, nella generazione dei grafi, trovare un modo per assegnare dei pesi agli archi. Di seguito sono riportate delle modifiche ai modelli precedentemente mostrati in modo che, durante la generazione dei grafi, venga assegnato un peso ad ogni arco.

3.1 Erdős–Rényi per grafi pesati

La caratteristica principale del modello Erdős–Rényi è l'indipendenza delle scelte. Infatti, nella creazione di un qualunque arco, non vengono guardati gli archi già creati, né i propri vicini. Nella scelta del metodo per assegnare i pesi agli archi durante la creazione del grafo, si è scelto di seguire la metodologia usata dal modello Erdős–Rényi nella creazione dei link. Nella struttura della rete sociale si è deciso di valutare gli archi con peso maggiore, che rappresentano i legami più forti, come quelli più rari mentre viceversa quelli con peso minore i più probabili.

Per replicare questa proprietà si è scelto di assegnare i pesi con probabilità proporzionale al valore del peso. Il modello così modificato prende in input il numero di nodi n , la probabilità

p con cui crearli e il massimo peso che si vuole assegnare durante la creazione degli archi k . Dato $G = (V, E)$ il grafo creato con il modello Erdős–Rényi e $W = \{1, 2, \dots, k\}$ l’insieme dei pesi da assegnare durante la creazione degli archi, $\forall e \in E \wedge \forall h \in W$, se indichiamo con $w(e)$ il peso dell’arco e allora $P_W(w(e) = h) = \frac{1}{h} \frac{1}{\sum_{j=1}^k \frac{1}{j}}$.

Algorithm 1 Erdős–Rényi weighted graph

```

1: procedure ER WEIGHTED GRAPH( $n, p, k_{max}$ )
2:    $V \leftarrow \{1, 2, \dots, n\}$ 
3:    $E \leftarrow \{\}$ 
4:    $G \leftarrow (V, E)$ 
5:    $W \leftarrow \{1, 2, \dots, k_{max}\}$ 
6:   for all  $j \in V$  do
7:     for all  $i \in V$  do
8:        $tolink \leftarrow True$  (with probability  $p$ )
9:       if  $tolink$  then
10:         $E \leftarrow E \cup \{(j, i)\}$ 
11:         $h \leftarrow$  extraction  $w$  from  $W$  with probability  $P_w(h)$ 
12:         $w(j, i) \leftarrow h$ 

return  $G$ 

```

3.2 Rich Get Richer per grafi pesati

Si vuole creare un algoritmo che generi un grafo G pesato tale che segua il Rich Get Richer Model. Dato E insieme degli archi $\forall(u, v) \in E$ allora $w(u, v)$ denota la forza del legame tra il nodo u e il nodo v .

1. I nodi sono creati in ordine, e chiamati 1,2,...,n
2. Durante la creazione di un collegamento uscente da un nodo u viene scelto, con probabilità P , se eseguire il passo a) o il passo b)
 - (a) Un nodo u sceglie con probabilità uniforme un nodo v e viene creato l'arco diretto (u, v) . Viene assegnato $w(u, v) = 1$.
 - (b) Un nodo u sceglie con probabilità uniforme un nodo v e viene creato un arco uscente da u diretto ad un nodo z tale che esiste l'arco (v, z) . Il peso $w(u, z) = w(v, z) + 1$.

Il procedimento sopra descritto può essere iterato per creare più archi uscenti da uno stesso nodo. Se ci troviamo ad una certa iterazione i e scegliamo il punto 2) una volta scelto in modo casuale un nodo u avremo a disposizione $i-1$ archi uscenti da u tra cui scegliere con probabilità uniforme.

Per come è stato costruito il grafo, gli archi creati scegliendo il punto a) sono quelli di peso 1 e quindi minimo, ma sono anche gli archi che collegano "direttamente" un nodo ad un altro. Vogliamo quindi che gli archi di tipo a) siano quelli che rappresentano il legame più forte.

Per pesare gli archi in modo ragionevole, una volta terminata l'esecuzione dell'algoritmo, possiamo riassegnare i pesi agli archi in questo modo:

sia w_{max} il peso maggiore assegnato ad un qualunque arco durante l'esecuzione dell'algoritmo. $\forall e \in E : w(e)$ è il peso dell'arco e , il nuovo peso $w(e) = (w_{max} + 1) - w(e)$.

Il grafo G risultante sarà quindi un grafo che rispetta il Rich-Get-Richer Model e che, inoltre, tiene traccia della forza dei legami di amicizia tra i nodi.

Algorithm 2 Rich Get Richer weighted graph

```
1: procedure RGR WEIGHTED GRAPH(V, probability, iterations)
2:    $G = (V, E)$ 
3:    $a \leftarrow probability * 100$ 
4:   for  $j \leftarrow 1$ , iterations do
5:     for all  $i \in V$  do
6:        $r \leftarrow$  random int between 1 and 100
7:       let  $v \in V$  a node chosen uniformly at random
8:       if  $r \leq a$  then                                 $\triangleright$  step A
9:          $E \leftarrow E \cup \{(i, v)\}$ 
10:         $w(i, v) \leftarrow 1$ 
11:       else                                      $\triangleright$  step B
12:          $u \in V : (v, u) \in E$ 
13:          $E \leftarrow E \cup \{(i, u)\}$ 
14:          $w(i, u) \leftarrow w(v, u) + 1$ 
15:      $maxWeight \leftarrow max(w(u, v)) : (u, v) \in E$        $\triangleright$  maximum weight of an edge
16:     for all  $u, v \in V : (u, v) \in E$  do
17:        $w(u, v) \leftarrow (maxWeight + 1) - w(u, v)$ 

return  $G$ 
```

Capitolo 4

Processi di diffusione

I processi di diffusione si collocano in quella che viene chiamata in sociologia la diffusione delle innovazione, la quale si occupa di studiare l'avanzare di una nuova idea, tecnologia o prodotto all'interno di una rete sociale. La diffusione del nuovo in una rete non avviene a livello globale bensì locale. Ad esempio, in un ambito lavorativo, possiamo trovarci a scegliere tecnologie che siano compatibili con i nostri colleghi di lavoro piuttosto che quelle più popolari nella società.

Queste osservazioni sono ben colte dal concetto di omofilia, per il quale si tende ad uniformarsi con i pensieri e le scelte delle persone a noi più vicine. I processi di diffusione modellano questi comportamenti e studiano come riuscire ad avviare una cascata possibilmente completa nella rete. Tra tutti questi processi si è scelto di analizzare il Best Friend Target Set Selection che coglie bene il concetto di omofilia.

4.1 Diffusione nelle reti

Prendendo in considerazione i nuovi comportamenti, le opinioni e le tecnologie e come tutte queste si diffondono da persona a persona attraverso una rete sociale si è visto come le persone influenzino i loro amici e siano influenzati da questi nell'adottare nuove idee. La comprensione di come funziona questo processo è basata su una lunga storia di lavoro empirico in sociologia, conosciuta come diffusione dell'innovazione.

Una serie di studi fatti nella metà del 20° secolo ha stabilito una strategia di ricerca di base per studiare la diffusione di un nuova tecnologia o idea attraverso un gruppo di persone e

ha analizzato i fattori che hanno facilitato o ostacolato il suo progresso. Questi studi hanno portato ad analizzare come le interazioni tra i nodi vicini possano portare alla diffusione dell’innovazione, possibilmente in tutta la rete, a partire da un insieme limitato di nodi iniziali denotati come iniziatori o semi. In particolare questo problema, conosciuto come Influence Massimization, è stato introdotto da Domingo e Richardson con l’obiettivo di realizzare una campagna di marketing virale utilizzando un numero piccolo di iniziatori. L’Influence Massimization Problem richiede, dato un intero k , l’insieme k di iniziatori che massimizzano la cascata. Per fare ciò è necessario assumere un meccanismo secondo il quale i nodi entrano a far parte della cascata.

Un primo meccanismo è basato sull’associazione di una soglia θ_v , per ogni nodo v , assumendo che per ogni coppia di nodi u, v ci sia un peso $W_{u,v}$, che indica l’influenza del nodo v sulla scelta del nodo u . Il nodo u è incluso nella cascata se e solo se $\sum_{v \in N(u)} W_{u,v} \geq \theta_u$, dove $N(u)$ è l’insieme dei vicini di u e $I_v = 1$ se v è nella cascata e 0 altrimenti. Quindi l’inclusione di u nella cascata dipende dalla somma dei pesi di tutti i suoi vicini che appartengono già alla cascata, la quale non deve essere minore del valore di soglia.

Un secondo meccanismo di diffusione è l’indipendent cascade model. In questo, caso quando un nodo u è incluso nella cascata, ogni vicino v che non appartiene ancora alla cascata viene incluso con probabilità $P_{u,v} = \frac{W_{u,v}}{\sum_{z \in N(u)} W_{u,z}}$. Mentre i due problemi precedenti cercano un insieme di iniziatori di una data cardinalità capace di diffondere una cascata completa, altri sono interessati a trovare un insieme di dimensione minima di iniziatori capace di diffondere la cascata in un insieme di almeno k nodi.

Uno dei problemi del secondo tipo è il Target Set Selection(TSS). Questo modello può essere così descritto: dato un grafo $G = (V, E)$, se indichiamo con $d(v)$ il grado di $v \in V$, per ogni nodo $v \in V$ è fornito un valore di soglia $t(v) \in N$, dove $1 \leq t(v) \leq d(v)$. Inizialmente tutti i nodi sono inattivi. Viene scelto un sotto insieme di nodi (target set) e viene reso attivo. Quindi, ad ogni passo, lo stato dei vertici viene modificato secondo la seguente regola: un vertice inattivo v diventa attivo se almeno $t(v)$ dei suoi vicini sono attivi. Il processo continua fino a che tutti i nodi sono attivi o nessun vertice può essere più convertito da inattivo ad attivo.

Tutti questi modelli però, assumono che un nodo diventi attivo in accordo con la quantità cumulativa dell’influenza che riceve dai suoi vicini. Questo viene modellato come la somma

del numero di archi uscenti da un nodo ed entranti in altri nodi già facenti parte della cascata. Questi modelli quindi non riescono a catturare le situazioni in cui i nodi entrano a fare parte della cascata non a causa del numero di vicini che ne fanno già parte ma grazie alla forza dei legami con questi ultimi. È per questo motivo che è stato studiato il problema Best Friend Target Set Selection.

4.2 Best Friend Target Set Selection

Il modello di diffusione del problema Best Friend Target Set Selection (BF-TSS) fa riferimento ad un grafo diretto fornito di una funzione peso r sugli archi, utilizzata come ranking dei vicini di ogni nodo e l'inserimento di un nodo u nella cascata è collegato alla quantità di ranking dei vicini di u .

Quindi u entra a far parte della cascata se i suoi vicini con ranking più alto ne fanno già parte. r rappresenta quindi il grado di amicizia di ogni nodo con ogni sua conoscenza. Secondo questo modello di diffusione, un nodo entra a far parte della cascata se i suoi θ migliori amici sono già nella cascata, dove θ è una soglia sulla funzione peso del nodo.

Un'istanza del problema BF-TSS è un grafo diretto $G = (V, E, r, \theta)$, con pesi sugli archi e sui nodi, dove $r : E \rightarrow N$ e $\theta : V \rightarrow N^+$, infine $V_0 \subseteq V$ è l'insieme dei nodi attivi. Si assume che per ogni $u \in V$, $\{r(u, v) | v \in N(u)\} = \{1, 2, \dots, K\}$ con $K \leq |N(u)|$: per ogni nodo $u \in V$ e per ogni $v, z \in N(u)$, $r(u, v) > r(u, z)$ significa che il legame (u, v) è più forte del legame (u, z) o che v è un amico migliore di z per u .

La regola di diffusione soggiacente il BF-TSS, con la quale un nodo decide se entrare nella cascata, specifica quale insieme $\theta(u) \leq |N(u)|$ dei vicini di u è necessario per contenere solo nodi attivi in modo da indurre u a diventare attivo.

Risulta interessante osservare due restrizioni in particolare della BF-rule, la prima per la quale $\theta(u) = 1$ (Follow one best friend) e la seconda dove $\theta(u) = |N_M(u)|$, con $N_M(u) = \{v \in N(u) : r(u, v) = \max\{r(u, z) : z \in N(u)\}\}$ (Follow all best friend). In effetti nel primo caso vogliamo che un nodo u entri a far parte della cascata se uno tra i suoi migliori amici ne fa parte, mentre nel secondo caso u aderisce alla cascata se tutti i suoi migliori amici vi appartengono. Sono riportati i due pseudo codici relativi alle restrizioni di cui sopra.

Algorithm 3 BFTSS-Follow One Best Friend

```
1: procedure BFTSS-FOLLOW ONE BEST FRIEND( $G$ , initiators)
2:    $G = (V, E)$ 
3:    $cascade \leftarrow \text{initiators}$ 
4:    $cascade_{old} \leftarrow \{\}$ 
5:   while  $cascade_{old} \neq cascade$  do
6:      $cascade_{old} \leftarrow cascade$ 
7:     for all  $u \in cascade$  do
8:       for all  $v \in V : (v, u) \in E$  do
9:         if  $w(v, u) \geq \max(w(v, i)) : (v, i) \in E$  then
10:           $cascade \leftarrow cascade \cup \{v\}$ 
return  $cascade$ 
```

Algorithm 4 BFTSS-Follow All Best Friend

```
1: procedure BFTSS-FOLLOW ALL BEST FRIEND( $G$ , initiators)
2:    $G = (V, E)$ 
3:    $cascade \leftarrow \text{initiators}$ 
4:    $cascade_{old} \leftarrow \{\}$ 
5:   while  $cascade_{old} \neq cascade$  do
6:      $cascade_{old} \leftarrow cascade$ 
7:     for all  $u \in cascade$  do
8:       for all  $v \in V : (v, u) \in E$  do
9:         if  $N_M(v) \in cascade$  then       $\triangleright N_M(v) = v$  neighbors with max weight
10:           $cascade \leftarrow cascade \cup \{v\}$ 
return  $cascade$ 
```

Capitolo 5

Analisi sperimentale dei processi di diffusione

L'implementazione degli algoritmi descritti nei precedenti capitoli è stata fatta utilizzando il linguaggio di programmazione Python. È possibile consultare l'implementazione degli algoritmi a questo link: <https://goo.gl/65fk5E>.

Ho scelto di non utilizzare alcuna libreria di creazione dei grafi in quanto un'implementazione personalizzata è risultata migliore in termini di prestazioni, soprattutto nel caso dell'algoritmo Follow Best Friend.

Sono stati creati circa 1000 grafi del tipo Erdős–Rényi e Rich Get Richer seguendo, rispettivamente, l'algoritmo 1 e 2. I grafi sono stati creati in modo random, passando valori casuali in input alle funzioni di creazioni dei grafi.

Come valori di input ho scelto, per il numero di nodi, 500, 800, 1000, 5000, 10000. Per la probabilità p valori compresi tra 0.2 e 0.8. Infine, per quanto riguarda la funzione di creazione di grafi del tipo Rich Get Richer, è richiesto un valore aggiuntivo corrispondente al numero di iterazioni dell'algoritmo, dipendente dalla cardinalità dei nodi.

Sono qui mostrate alcune immagini raffiguranti i risultati ottenuti dalla generazione dei grafi.

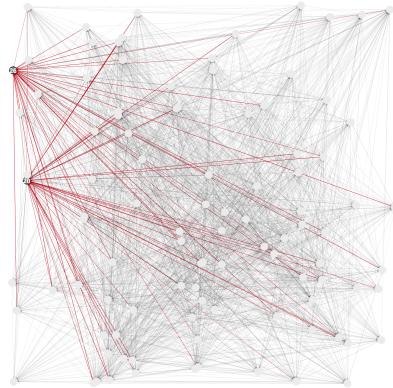


Figura 5.1: Grafo di 100 nodi creato con algoritmo 1. I nodi evidenziati sono quelli con in-degree maggiore.

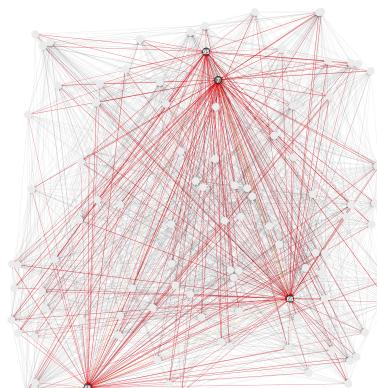


Figura 5.2: Grafo di 100 nodi creato con algoritmo 2. I nodi evidenziati sono quelli con in-degree maggiore.

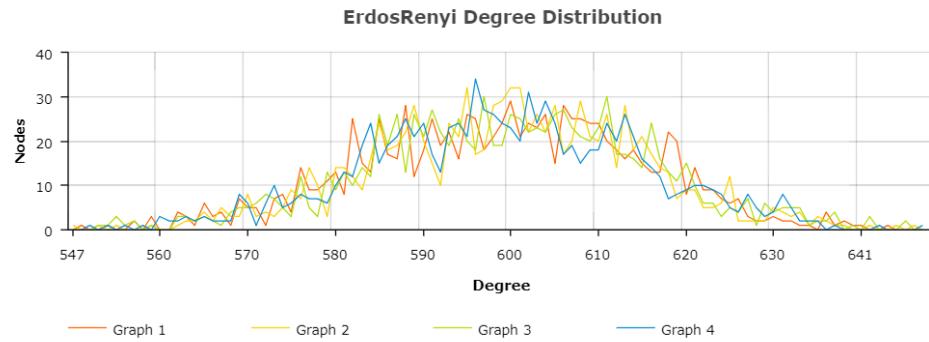


Figura 5.3: Distribuzione dei gradi dei nodi per 4 grafi creati secondo il modello Erdős–Rényi

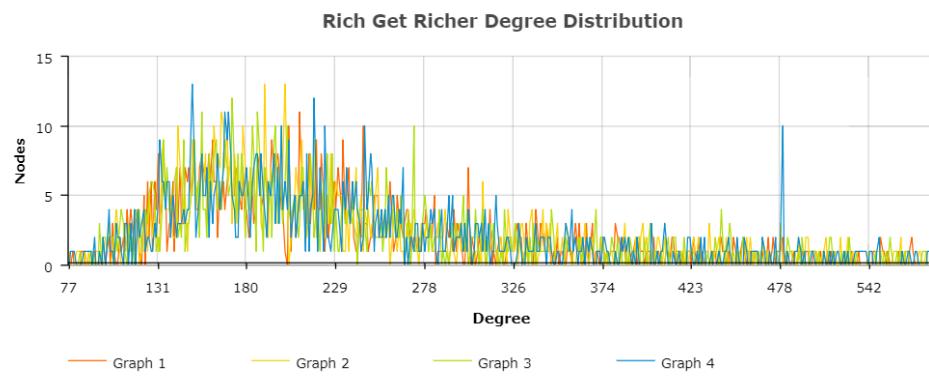


Figura 5.4: Distribuzione dei gradi dei nodi per 4 grafi creati secondo il modello Rich Get Richer

Come si evince dalle figure 5.1 e 5.2, il modello Rich Get Richer, come ci si aspettava, ha permesso di creare grafi con un numero di nodi famosi maggiori rispetto al modello Erdős–Rényi. La distribuzione di nodi mostrata in figura 5.3 e in figura 5.4 dei due modelli hanno quindi seguito il valore atteso in entrambi i casi.

Su tutti i grafi creati sono stati eseguiti gli algoritmi 3 e 4, rispettivamente con, $\forall u \in V$, valori di soglia $\theta(u) = 1$ e $\theta(u) = N(u)$. Inoltre è stato eseguito anche l'algoritmo Follow Best Friend usando valori di soglia casuali, del tipo $\theta = \frac{N(u)}{t}$ con $t \in N^+$.

Nei risultati ottenuti è emerso come, nei casi in cui la soglia θ è molto vicina ad 1, la cascata raggiunga tutta la rete in gran parte delle esecuzioni. Mentre, nei casi in cui la soglia è vicina ad essere uguale al numero di vicini dei nodi presi in considerazione, la cascata non ha inizio. Nella seguente tabella vengono mostrate le percentuali relative alle esecuzioni che hanno portato a cascate complete e quelle relative a cascate che hanno avuto inizio, ma che si sono bloccate. Inoltre verranno mostrati grafici relativi alle esecuzioni più significative dell'algoritmo Follow Best Friend, nei grafi costruiti con il modello Rich Get Richer e in quelli creati con il modello Erdős–Rényi.

$\theta(u)$	Cascade complete	Cascade bloccate
1	100%	100%
$\frac{1}{4}N(u)$	31.70%	36.58%
$\frac{1}{3}N(u)$	22.22%	29.62%
$\frac{4}{5}N(u)$	0%	13.15%
$N(u)$	0%	0%

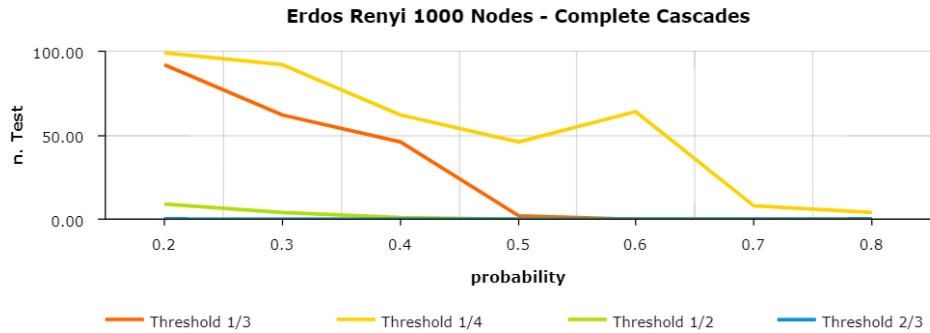


Figura 5.5: Grafico relativo a 700 prove su grafi di 1000 nodi costruiti secondo il modello Erdős–Rényi. In ascissa la probabilità usata nella creazione dei grafi. In ordinata il numero di prove che hanno portato ad una cascata completa sul totale. Ogni linea corrisponde a un valore di soglia dell’algoritmo Follow Best Friend differente.

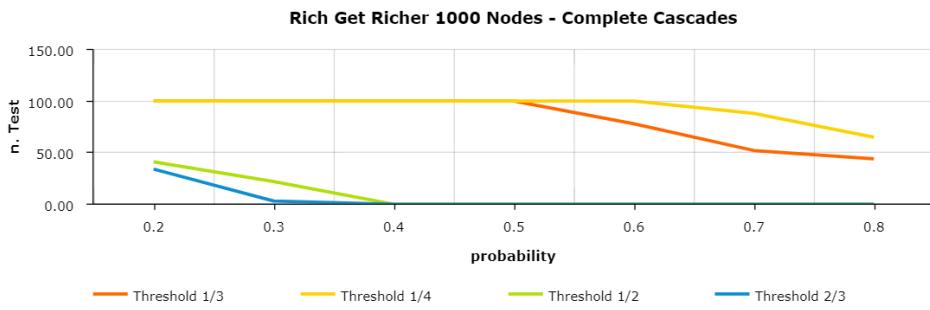


Figura 5.6: Grafico relativo a 700 prove su grafi di 1000 nodi costruiti secondo il modello Rich Get Richer. In ascissa la probabilità usata nella creazione dei grafi. In ordinata il numero di prove che hanno portato ad una cascata completa sul totale. Ogni linea corrisponde a un valore di soglia dell’algoritmo Follow Best Friend differente.

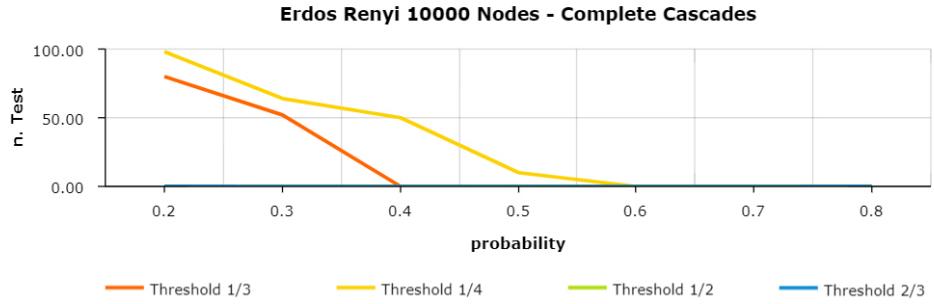


Figura 5.7: Grafico relativo a 700 prove su grafi di 10000 nodi costruiti secondo il modello Erdős–Rényi. In ascissa la probabilità usata nella creazione dei grafi. In ordinata il numero di prove che hanno portato ad una cascata completa sul totale. Ogni linea corrisponde a un valore di soglia dell’algoritmo Follow Best Friend differente.

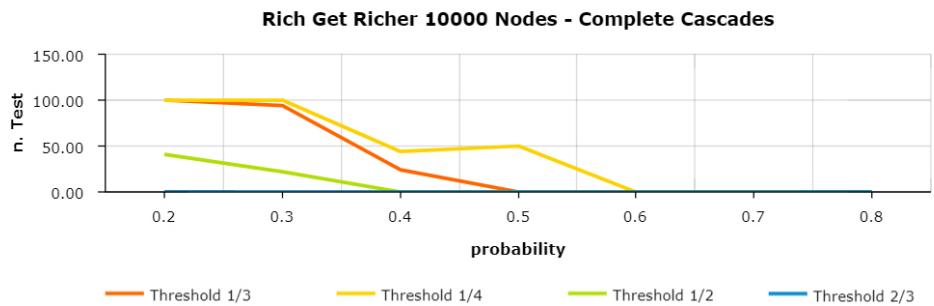


Figura 5.8: Grafico relativo a 700 prove su grafi di 10000 nodi costruiti secondo il modello Rich Get Richer. In ascissa la probabilità usata nella creazione dei grafi. In ordinata il numero di prove che hanno portato ad una cascata completa sul totale. Ogni linea corrisponde a un valore di soglia dell’algoritmo Follow Best Friend differente.

Capitolo 6

Conclusioni

Molti degli studi sociali che fanno uso della teoria dei grafi hanno bisogno di modelli per generare reti casuali, in modo da sostituire i grafi reali non di facile reperibilità. Questo studio di tesi ha confermato come il modello Rich Get Richer sia migliore nel descrivere reti reali rispetto al modello Erdős–Rényi, che risulta troppo probabilistico.

Per quanto riguarda il problema Best Friend Target Set Selection si è osservato che senza un modo adeguato per scegliere l'insieme di iniziatori diventa difficile riuscire ad arrivare ad una cascata completa in tutta la rete. Questo è vero nella maggior parte dei casi, con valori di soglia anche non troppo lontani da 1, mentre con valori molto vicini ad 1, anche con insiemi piccoli di iniziatori, la cascata ha inizio con molta facilità.

Infine si è visto come l'applicazione del modello Follow Best Friend a grafi costruiti secondo l'algoritmo 2 per la creazione di grafi Rich Get Richer pesati, abbia raggiunto cascate complete in casi maggiori rispetto all'applicazione dello stesso modello su grafi costruiti con l'algoritmo 1 per grafi pesati che seguono il modello Erdős–Rényi. In entrambi i modelli il numero di cascate complete diminuisce molto velocemente aumentando il numero di soglia. Inoltre il diminuire delle cascate risulta dipendere molto anche dalla probabilità con cui sono stati creati i grafi. Quest'ultimo risultato va osservato pensando ai due modelli trattati.

Nel modello Erdős–Rényi la probabilità guida la presenza o l'assenza di archi. Con probabilità prossime ad 1 il grafo risultante è molto denso. L'alta densità del grafo in relazione ai pesi assegnati durante la creazione, risultano limitare le possibilità di ottenere cascate complete.

Nel caso del modello Rich Get Richer, alte probabilità portano ad una predominanza della com-

ponente Erdős–Rényi sulla componente Rich Get Richer. Questo comporta un grafo molto più simile a quelli creati dal modello Erdős–Rényi, cosa che risulta diminuire le cascate complete raggiunte dell’algoritmo.

6.1 Sviluppi futuri

Possibili sviluppi futuri per questo lavoro di tesi sono l’individuazione di altri modelli casuali per la generazione di reti sociali, che abbiano, inoltre, la capacità di pesare gli archi adeguatamente. Diverrà dunque possibile l’implementazione di questi modelli con lo scopo di analizzare l’effetto del modello di cascata del problema Best Friend Target Set Selection su queste reti. Risulta interessante, infine, confrontare i risultati degli algoritmi Follow Best Friend tra i grafi costruiti con i modelli random presentati in questo lavoro di tesi e le reti reali. Questo però implica un accurato lavoro di ricerca di grafi, in quanto sono necessarie reti sociali con pesi sugli archi che sono tra quelle di più difficile reperibilità.

Bibliografia

- [1] David Easley, Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.
- [2] Miriam Di Ianni, Giorgio Gambosi. *Information diffusion under strong relations influence*. July 22, 2016.
- [3] Chen, N. *On the approximability of influence in social networks*. In Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Philadelphia, PA, USA, 2008).
- [4] Leskovec, J., Adamic, L. A., and Huberman, B. A. *The dynamics of viral marketing*. ACM Trans. Web 1, 1 (May 2007).
- [5] Goldenberg, J., Libai, B., and Muller, E. *the network: A complex systems look at the underlying process of word-of-mouth*. Marketing letters 12, 3 (2001), 211-223.
- [6] Kempe, D., Kleinberg, J., and Tardos, E. *Influential nodes in a diffusion model for social networks*. In *Proceedings of the 32nd International Conference on Automata, Languages and Programming* (Berlin, Heidelberg, 2005), ICALP'05, Springer-Verlag, pp. 1127-1138.
- [7] Goel, S., Watts, D. J., and Goldstein, D. G. *The structure of online diffusion networks*. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (New York, NY, USA, 2012), EC '12, ACM, pp. 623-638.
- [8] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. *Group formation in large social networks: Membership, growth, and evolution*. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2006), KDD '06, ACM, pp. 135-144.

national Conference on Knowledge Discovery and Data Mining (New York, NY, USA, 2006), KDD '06, ACM, pp. 44-54.

- [9] Ackerman, E., Ben-Zwi, O., and Wolfowitz, G. *Combinatorial model and bounds for target set selection*. *Theoretical Computer Science* 411 (2010), 4017-4022.
- [10] Bazgan, C., Chopin, M., Nichterlein, A., and Sikora, F. *Parameterized approximability of maximizing the spread of influence in networks*. *Journal of Discrete Algorithms* 27 (2014), 54-65.
- [11] Ben-Zwi, O., Hermelin, D., Lokshtanov, D., and Newman, I. *Treewidth governs the complexity of target set selection*. *Discrete Optimization* 8, 1 (2011), 87-96.
- [12] Chopin, M., Nichterlein, A., Niedermeier, R., and Weller, M. *Constant thresholds can make target set selection tractable*. In Proceedings of the First Mediterranean Conference on Algorithms, vol. 7659 of LNCS.