ABSONES: Agent Based SOcial NEtwork Simulator

Simone Ciccolella
s.ciccolella@campus.unimib.it
Mat. 762234

Daniele Bellani
d.bellani1@campus.unimib.it
Mat. 780675

Sommario. In questo progetto proponiamo un modello multi-agente per la simulazione di social network, in particolare la piattaforma di microblogging Twitter. Il nostro approccio prevede la rappresentazione del sistema come un grafo diretto e l'identificazione degli utenti come agenti di tale sistema complesso, che si attivano e compiono azioni sulla base di probabilità ed estrazioni di tipo Monte Carlo. Gli obiettivi che ci siamo posti sono stati la prevenzione del *completamento* della rete e il mantenimento del principio di *invarianza di scala*, così da simulare, in modo quanto più fedele, il comportamento della vera rete sociale. Con il modello così costruito, è possibile studiare fenomeni sociologici usando scenari simulati nel modo più attinente possibile alla realtà. In particolare, abbiamo provato a simulare il "salto" di popolarità di un individuo poco conosciuto in seguito all'interazione con un utente più "in vista", traendo ispirazione da quanto avvenuto nel 2012 tra l'On. Maurizio Gasparri e l'utente Daniele Termite.

Indice

1 Introduzione **3**

2 Letteratura correlata **4**

3 Riferimenti matematici **7**
3.1 Cosine similarity 7
3.2 Teoria dei grafi 7
3.2.1 Coefficiente di clustering 7
3.2.2 Assortatività 8
3.2.3 Rete a invarianza di scala 8

4 Descrizione modello **9**

5 Visione multi-agente **9**

6 Esperimenti **9**

7 Conclusione **10**

8 Sviluppi futuri **10**

1 Introduzione

I social network sono uno strumento che è diventato parte integrante della vita di tutti. Nel giugno 2017 la piattaforma Facebook ha passato la soglia dei due miliardi di utenti attivi su base mensile [1]; nel frattempo, Twitter si è imposto come mezzo di comunicazione principale tra i cosiddetti "influencer", con un grosso impatto sull'opinione pubblica (per esempio, giocando un ruolo fondamentale nell'elezione di Donald Trump [2]). L'aspetto principale di questa forma di comunicazione è la produzione, da parte degli utenti, di un'enorme mole di dati: le acquisizioni di YouTube da parte di Google [3] e di LinkedIn da parte di Microsoft [4] (e le cifre in gioco) testimoniano l'interesse delle grandi aziende per queste sorgenti di dati. L'analisi di questi ultimi aiuta a determinare strategie di mercato, a personalizzare raccomandazioni di prodotti, studiare e prevedere il sentimento su un evento o un prodotto, oppure condurre studi di stampo sociologico.

Sorgono però alcune difficoltà. In primo luogo, raramente questi dati vengono rilasciati in formato aperto (Open Data). Inoltre, le loro enormi proporzioni ne rendono difficile l'analisi e la gestione. Da qui la necessità di svolgere delle simulazioni *in-silico* con modelli sviluppati in modo da essere il più possibile veritieri. L'attendibilità di queste simulazioni dipende dal modello su cui vengono eseguite e dalle inevitabili assunzioni che sono state fatte durante la sua costruzione. L'obiettivo è quindi produrre sistemi artificiali che siano repliche fedeli di sistemi complessi reali. Una possibile strada è quella dei *sistemi multi-agente*. Questi si basano sulla definizione di *agente* (vedi sezione 5), un entità virtuale [8] capace di compiere azioni ed interagire con altri agenti all'interno di un ambiente (*environment*). Si può ottenere una rappresentazione multi-agente di una rete sociale rappresentando l'ambiente come un grafo (connesso o non connesso) e gli agenti come nodi di questo grafo. Gli archi rappresentano le interazioni tra i vari agenti, siano esse relazioni di "amicizia" (es. Facebook), oppure di "subscription" (es. Twitter). Il tempo viene spesso rappresentato in modo discreto dalle *iterazioni* o *step*. Ad ogni step, i nodi decidono individualmente le azioni da compiere, per esempio se stabilire o meno un nuovo collegamento con un altro nodo. Ciò rende il sistema *dinamico* ed in continua evoluzione. Un modello è considerato attendibile se conserva le caratteristiche proprie di una rete sociale per tutta la durata della simulazione. In questo modo è possibile simulare scenari utili alla previsione di fenomeni sociali nel modo più fedele possibile.

Negli ultimi anni sono stati proposti alcuni modelli per la simulazione multi-agente di social network, con risultati alterni e non definitivi, dovuti alla complessità del tema. Da questi abbiamo tratto ispirazione per il nostro modello. Nella prima sezione esamineremo alcune delle metodologie proposte; successivamente, dopo aver fissato alcune definizioni di carattere matematico utili per le sezioni seguenti, passeremo all'illustrazione della nostra soluzione e della visione agent-based dello stesso. Nelle ultime due sezioni esporremo gli esperimenti condotti e ne commenteremo i risultati.

2 Letteratura correlata

La letteratura riguardante modellazione multi-agente di social network è relativamente scarsa. Nonostante la pubblicazione in formato open di dati provenienti da alcune piattaforme, risulta comunque molto difficile avere una visione generale di ciò che accade in sistemi così ampi e complessi. Una situazione migliore si riscontra nella letteratura riguardante l'analisi a posteriori di queste reti, da tempo molto studiate. Non mancano quindi metriche e parametri per la valutazione di reti sociali, anche se le visioni in merito sono molte e a volte piuttosto discordanti.

Un primo lavoro degno di nota è quello pubblicato da Hamill e Gilbert [5]. In primo luogo, gli autori stabiliscono quali sono le caratteristiche che una rete sociale simulata dovrebbe avere, tra cui le più importanti sono:

Bassa densità di rete La densità di una rete [5] (*network density*) è definita come il rapporto tra il numero di archi esistenti e il numero massimo di archi possibili. Un utente medio è collegato con un numero di utenti dell'ordine delle centinaia o poche migliaia, numero che, se confrontato con le centinaia di milioni di utenti (se non miliardi) di tutto il sistema, risulta essere piuttosto basso.

Assortatività positiva Con questo termine, gli autori indicano la tendenza dei nodi con più connessioni ad essere collegati con altri nodi molto connessi (vedi 3.2).

Presenza di comunità Ovvero, la tendenza a formare *clusters*, gruppi di nodi fortemente connessi tra di loro ma debolmente connessi con il resto del sistema. Viene introdotto, a questo proposito, il *coefficiente di clustering* (*clustering coefficient*, vedi 3.2)

Lunghezza ridotta dei cammini Secondo gli autori, in media si può raggiungere un utente di un social network partendo da un qualsiasi altro nodo compiendo solo pochi passi, ovvero percorrendo un cammino ridotto. La lunghezza dipende dalle proporzioni della rete. Questo è un effetto molto noto in letteratura, e prende il nome di *small-world effect* [6][18].

Vengono esposti inoltre diversi tipi di rete, emersi nel corso degli anni in letteratura (vedi fig. 1):

Regular lattice Ogni nodo è collegato ad un numero fisso di suoi vicini

Random network Ogni nodo è collegato in media ad un certo numero di altri nodi

Small world network Basato sul modello *regular lattice*, aggiunge o riarrangia collegamenti in modo casuale

Scale-free network Descritta per la prima volta da Barabási & Bonabeau [7], prevede che pochi nodi abbiano molti collegamenti (vedi sezione 3.2).

Gli autori indicano la costruzione *scale-free* come la migliore tra le quattro, in quanto presenta tutte le caratteristiche elencate in precedenza, con l'eccezione dell'*assortatività*, non particolarmente riflessa nel modello. Passano quindi all'esposizione della loro proposta, un modello ad agenti basato sul concetto di *social circles* [5]: ogni agente può stabilire un "link" con un altro agente solo se quest'ultimo può fare altrettanto. Quest'idea di *reciprocità* si adatta bene alla modellazione di alcune piattaforme (es. Facebook), mentre si adatta meno su altre: un esempio è Twitter, dove per stabilire un collegamento, il fatto che due utenti si conoscano direttamente è poco rilevante.

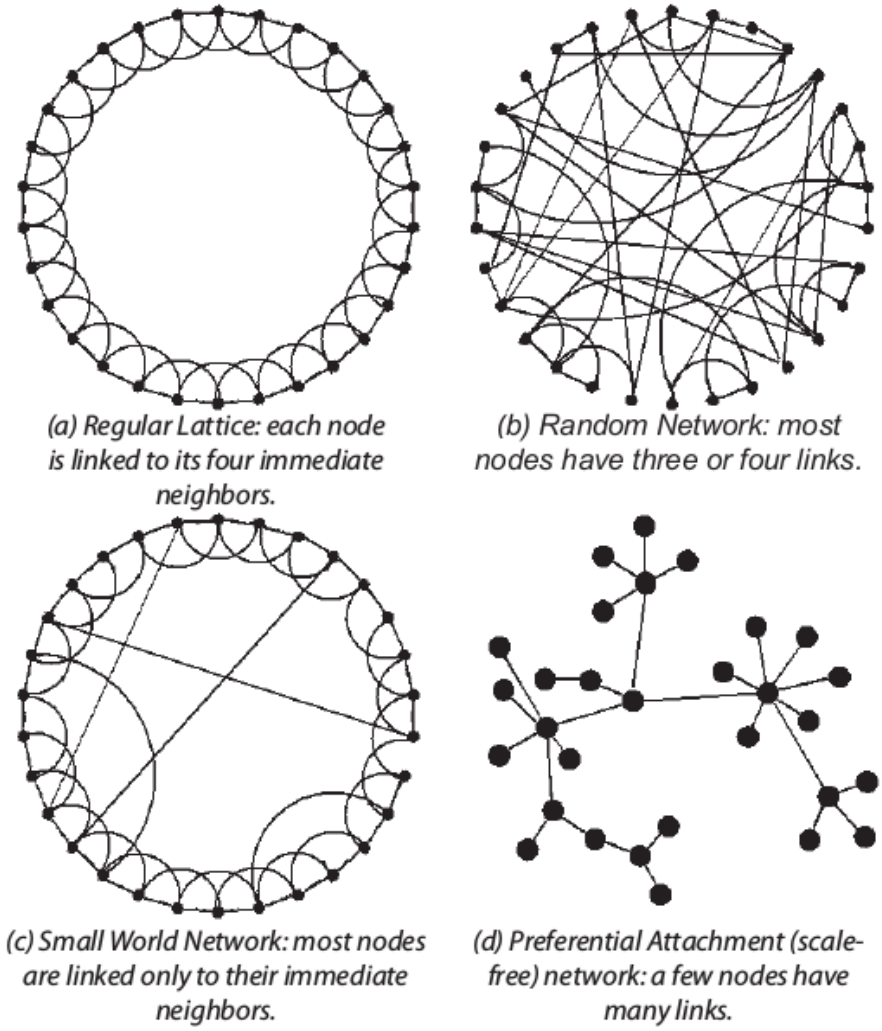

Figura 1: I quattro modelli di rete descritti in [5]

L'idea che una rete sociale sia assimilabile a una rete ad invarianza di scala (ovvero i cui nodi seguono una *power law*, vedi 3.2) è largamente supportata in letteratura: testi importanti e fondamentali come [6] e [18] asseriscono che i social network, così come il World Wide Web stesso, siano organizzati in modo tale che pochi nodi abbiano molti collegamenti e siano collegati tra di loro, mentre molti nodi abbiano pochi collegamenti e siano collegati ad alcuni di questi "hub". Entrambi i testi, così come [5], ricordano però l'assunzione *small-world*, che, specialmente per i social network, debba essere verificata.

Un altro lavoro che abbiamo recuperato è quello di Liben-Nowell & Kleinberg [9]. Il problema da loro affrontato è quello della previsione di collegamenti (*link prediction*) nei social network, ovvero prevedere, dato lo stato del sistema in un dato istante di tempo, la formazione di nuovi collegamenti negli istanti immediatamente successivi. In una simulazione *dinamica* di una rete è fondamentale stabilire un metodo per cui nuovi collegamenti vengono stabiliti tra i nodi. Un aspetto cruciale in questo caso è il criterio con cui viene valutata la *similarità* tra i nodi del grafo, ovvero come viene assegnato uno *score* a una coppia di nodi (arco) in modo che ne misuri la *"distanza"* rispetto ad una particolare proprietà o caratteristica. Nodi simili ma non ancora collegati avranno infatti un'alta probabilità di stabilire un collegamento negli istanti di tempo più prossimi.
Nell'articolo viene fatta una rassegna dei metodi principali; l'obiettivo di questi metodi è costruire la matrice di similarità :

Neighborhood-based methods Di questa categoria fanno parte tutti quei metodi che si basano sul *vicinato* (*neighborhood*) dei nodi di cui si vuole calcolare la similarità. La misura di distanza può essere il numero di nodi adiacenti in comune, oppure la probabilità che i due nodi in esame abbiano un vicino in comune, per esempio usando la *distanza di Jaccard*. In alternativa, questa probabilità può essere calcolata in modo proporzionale alla dimensione dei vicinati dei due nodi candidati, per esempio moltiplicandone la cardinalità.

Paths-based methods Anche la distanza intesa come lunghezza di un cammino tra due nodi può essere intesa come misura di similarità; per esempio, in [10] viene descritta una metrica che prende in considerazione la lunghezza tutti i cammini esistenti tra due nodi per quantificarne la similarità. Anche alcuni celebri algoritmi, come il *PageRank* di Google [11], fanno parte di questa categoria.

Altre tecniche possono essere usate in congiunzione con i metodi sopra elencati, per semplificarne la computazione oppure per irrobustirne la previsione. Procedure di *clustering* o *matrix factorization* possono aiutare ad eliminare i collegamenti meno significativi prima dell'effettivo calcolo delle metriche.

Essendo però il nostro obiettivo la simulazione di una *social* network, non si può ignorare la (forte) componente sociale del sistema considerato. Ogni nodo rappresenta un utente, e dunque ne eredita le caratteristiche personali come interessi e carattere. La letteratura specializzata in *social sciences* ha da tempo definito il concetto di *omofilia* (*homophily*) [12] [18], inteso come la tendenza di ogni individuo a stringere legami con altri individui dalle caratteristiche simili. Nel computo della misura di similarità (o dissimilarità) bisognerebbe quindi tenere conto di tali caratteristiche e trovare quindi il modo di rappresentarle nel sistema. In [13], per esempio, gli autori hanno tentato di tradurre in linguaggio matematico l'idea di omofilia, in particolare per quanto riguarda l'aggregazione di individui in base al gruppo di appartenenza. Dopo aver effettuato un'etichettatura degli individui in base a questi raggruppamenti, hanno definito un indice H_i per l'omofilia:

$$H_i = \frac{s_i}{s_i + d_i} \tag{1}$$

dove s_i indica il numero medio di *amicizie* (*friendships*) che un individuo facente parte del gruppo i ha con i membri dello stesso gruppo, mentre d_i indica il numero medio di *amicizie* che un membro del gruppo i ha con gli individui appartenenti ad altri gruppi. Questa misura, se confrontata con la frequenza relativa w_i degli individui apparteneti al gruppo i, fornisce un indicazione del comportamento degli agenti all'interno del sistema, ovvero se tendono a stabilire conllegamenti fra individui dello stesso gruppo (*inbreeding homophily*) oppure con individui apparteneti ad altri gruppi (*heterophily*).
Occorre notare come, nell'ambito dei social network, il termine *omofilia* e il termine *assortatività* siano spesso indicati come sinonimi. [20] e [21] per esempio, forniscono rispettivamente una misura di assortatività per attributo e per grado (vedi 3.2), ovvero un indice di quanto i vertici di una rete siano legati a nodi con simili valori per l'attributo considerato oppure con simili gradi di archi entranti o uscenti.

Esistono modelli agent-based oltre a quello esposto in [5], per esempio quelli descritti in [14] e [15]; entrambi usano una rete *scale-free* come ambiente per gli agenti (il primo orientata, la seconda non orientata), ed entrambi regolano gli eventi che accadono nel sistema e le azioni degli agenti secondo degli spazi di probabilità, che si aggiornano dinamicamente. Queste probabilità sono in funzione dell'*in-degree* di ogni nodo e della *similarità* tra ogni coppia di nodi. In particolare, [14] contempla anche l'aggiunta e la delezione di nodi durante la simulazione, operazioni anch'esse regolate da una specifica distribuzione di probabilità.

Un'ultima risorsa da noi sfruttata è stata la relazione stilata da Marco Comi e Marco Gravina che, come noi e prima di noi, si sono cimentati nella simulazione del social network Twitter. Abbiamo preso spunto dalle loro soluzioni e dalle criticità del loro modello per sviluppare la nostra proposta.

3 Riferimenti matematici

Di seguito descriviamo brevemente alcuni concetti matematici fondamentali per la comprensione delle sezioni successive.

3.1 Cosine similarity

Il *coseno di similitudine* (*cosine similarity*) [16] è una misura di similarità tra due vettori v_1 e v_2, definita in questo modo:

$$sim(v_1, v_2) = \frac{v_1 v_2}{|v_1||v_2|} \tag{2}$$

dove il numeratore rappresenta il *prodotto scalare* tra i due vettori, mentra il denominatore rappresenta il prodotto dei moduli. I valori possibili ricandono nell'intervallo $[-1, 1]$, dove 1 si ottiene in caso di vettori identici, mentre -1 in caso di completa dissimilarità (vettori opposti).

3.2 Teoria dei grafi

Un *grafo orientato* G [17] è una coppia (V, E), dove V (insieme dei *vertici*) è un insieme finito ed E è una relazione binaria in V.
Se (u, v) è un arco [17] di un grafo $G = (V, E)$ diciamo che il vertice v è *adiacente* al vertice u.
Dato un grafo G orientato [17], il *grado uscente* (*out-degree*) di un vertice è il numero di archi che escono dal vertice; il *grado entrante* (*in-degree*) è il numero i archi che entrano nel vertice.
Un cammino (*path*) [6] da un vertice v_0 a un vertice v_n è una lista ordinata di archi $P = \{(v_0, v_1), (v_1, v_2), ..., (v_{n-1}, v_n)\}$, e n corrisponde alla lunghezza di questo cammino.

Un grafo orientato G si dice *completo* [6] quando ogni coppia di vertici è collegata da una coppia simmetrica di archi. La definizione è analoga al caso in cui il grafo sia non orientato, con la differenza che, in quest'ultimo, ogni coppia di archi opposti situata tra due nodi è sostituita da un solo arco non orientato. Il numero di archi [6] in un grafo *non orientato completo* è pari a $\frac{N(N-1)}{2}$, dove N è il numero di vertici del grafo. Se si escludono i *cappi* (*self-loops*), allora un grafo *orientato completo* è composto da $N(N-1)$ archi.

3.2.1 Coefficiente di clustering

Il *coefficiente di clustering* (*clustering coefficient*) [6], in un grafo *non orientato*, cerca di esprimere il grado con cui i vicini di un dato nodo sono collegati tra di loro. Questo fornisce un'idea della compattezza del gruppo in cui è inserito quello specifico nodo. E' possibile inoltre avere una misura globale della tendenza dei nodi a "raggrupparsi" calcolando il *coefficiente di clustering medio* (*average clustering coefficient*). Nei grafi orientati [18] è disponibile una misura analoga, chiamata *coefficiente di clustering globale* (*global clustering coefficient*, a volte anche chiamata *ratio of transitive triplets* [6]); la sua definizione è la seguente:

$$C_\Delta = \frac{3 \times NumeroDiTriangoli}{NumberoDelleTripleConnesse} \tag{3}$$

dove per *triangolo* si intende un insieme di 3 vertici, ognuno dei quali connesso agli altri due da archi, mentre per *tripla connessa* si intende tre vertici uvw connessi dagli archi (u, v) e (v, w). I valori possibili ricadono nell'intervallo $[0, 1]$.

3.2.2 Assortatività

Una rete si dice *assortativa* (*assortative*) [18] se una frazione significativa dei suoi archi collegano vertici simili tra di loro. La similitudine può essere calcolata rispetto a un particolare attributo dei nodi, o, in alternativa, rispetto al *grado* dei vertici. In quest'ultimo caso, in un grafo diretto [20] la *in-assortativity* e la *out-assortativity* misurano rispettivamente la tendenza dei nodi a legarsi con altri nodi che hanno *in-degree* o *out-degree* identico al loro. Sia $r()$ la funzione che calcola la assortatività, e siano indicati i due tipi di assortatività appena descritti come $r(in, in)$ e $r(out, out)$, è possibile estendere la portata della definizione definendo anche $r(in, out)$ e $r(out, in)$. Supponendo di voler calcolare $r(in, out)$, si ricorre alla seguente formula [21]

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j^{in} q_k^{out})}{\sigma_{in}\sigma_{out}} \tag{4}$$

dove e_{jk} è la probabilità che un arco qualsiasi conduca da un nodo con *in-degree* j a un nodo con *out-degree* k, σ_{in} è la deviazione standard della distribuzione q^{in} e σ_{out} è la deviazione standard della distribuzione q^{out}. La distribuzione di probabilità q_k^{out} (e analogamente q_j^{in}) è calcolata come segue:

$$q_k^{out} = \frac{(k+1)p_{k+1}^{out}}{\sum_k k p_k^{out}} \tag{5}$$

dove p_k^{out} è la probabilità che un nodo abbia *out-degree* k. Al denominatore si trova l'out-degree medio della rete. L'assortatività per attributo è analoga, e si calcola in questo modo:

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b} \tag{6}$$

dove e_{xy} è la probabilità che un arco qualsiasi conduca da un nodo con valore dell'attributo x a un nodo con valore y, mentre a e b sono rispettivamente le frequenze degli archi che escono ed entrano in un nodo con valori x e y. In entrambi i casi il valore di r ricade nell'intervallo $[-1, 1]$, con 1 a indicare la perfetta assortatività e -1 la perfetta disassortatività.

3.2.3 Rete a invarianza di scala

Una rete viene indicata come *rete a invarianza di scala* (*scale free network*) se la distribuzione dei gradi dei nodi (probabilità $p(k_i)$ che un nodo scelto in modo random abbia grado k_i) segue una *power law* [7].
Una *power law* [6] è una funzione $y = f(x)$ in cui il valore y della funzione è proporzionale ad una potenza del valore in ingresso x:

$$p_x \sim x^{-\gamma} \tag{7}$$

L'equazione 4 è chiamata *power law distribution* e l'esponente $-\gamma$ è detto *degree exponent*. La rappresentazione con il logaritmo dell'equazione 4 diventa

$$\log p_x \sim -\gamma \log p \tag{8}$$

Se vale l'equazione 5, allora ci si aspetta che $\log p_x$ dipenda linearmente da $\log x$, con inclinazione della retta pari a γ, come indicato nella figura 2. Assumendo che il grado di un nodo sia una quantità discreta, si può adottare il formalismo per il discreto; in questo modo, l'equazione 4 diventa

$$p_x \sim C x^{-\gamma} \tag{9}$$

dove C è una costante.

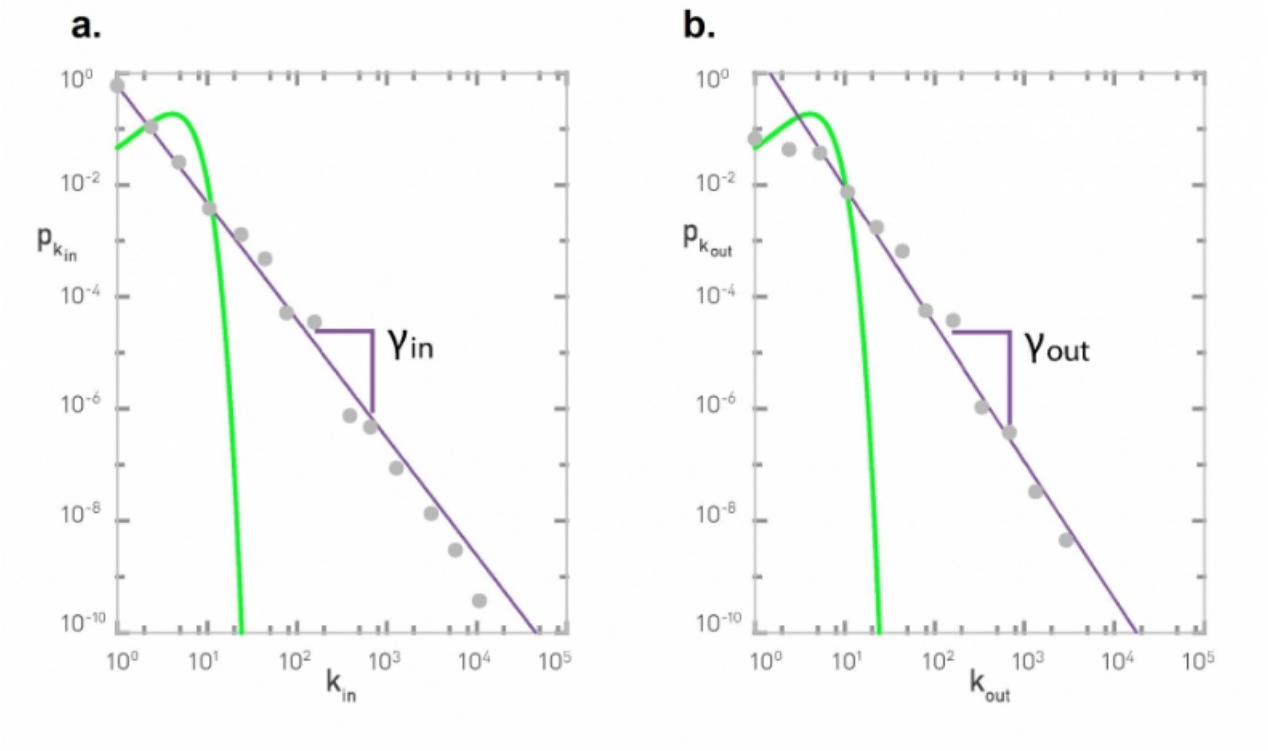

Figura 2: Distribuzione del WWW misurata nel 1999 [6]. La figura a. rappresenta la distribuzione dei nodi rispetto all'*in-degree*, mentre b. rispetto all'*out-degree*. Entrambe usano una scala logaritmica.

4 Descrizione modello

5 Visione multi-agente

6 Esperimenti

In questa sezione esponiamo gli esperimenti e le prove condotte con il nostro modello.

Il nostro obiettivo è stato quello di produrre un modello ad agenti che simulasse il comportamento del social network Twitter, rispettando il più possibile il vero comportamento del sistema. Più in dettaglio, anche a fronte di quanto espresso in letteratura (vedi sezione 2), abbiamo fissato due obiettivi principali:

1. Evitare il completamento della rete
2. Produrre una rete ad invarianza di scala (sezione 3)

Abbiamo individuato questi due punti come i più cruciali per la riuscita del modello,

7 Conclusione

8 Sviluppi futuri

Riferimenti bibliografici

[1] Welch, Chris (2017). "Facebook crosses 2 billion monthly users". *The Verge. Vox Media.* Retrieved June 27, 2017.

[2] Bulman, May (2016). "Donald Trump's 'celebrity-style' tweets helped him win US presidential election, says data scientist". *The Independent* Retrieved November 28, 2016.

[3] Sorkin, Andrew Ross & Peters, Jeremy, W. (2006). "Google to Acquire YouTube for $1.65 Billion". *The New York Times.* Retrieved October 9, 2006.

[4] Greene, Jay (2016). "Microsoft to Acquire LinkedIn for $26.2 Billion". *The Wall Street Journal.* Retrieved June 14, 2016.

[5] Hamill, L., & Gilbert, N. (2010). Simulating large social networks in agent-based models: A social circle model. Emergence: Complexity and Organization, 12(4), 78.

[6] Barabási, A. L. (2016). Network science. Cambridge university press.

[7] Barabási, B. A. L., & Bonabeau, E. (2003). Scale-free. Scientific American, 288(5), 50-59.

[8] Ferber, J. (1999). Multi-agent systems: an introduction to distributed artificial intelligence (Vol. 1). Reading: Addison-Wesley. ISO 690

[9] Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. journal of the Association for Information Science and Technology, 58(7), 1019-1031.

[10] Katz, L. (1953). A new status index derived from sociometric analysis. Psychometrika, 18(1), 39-43.

[11] Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks, 56(18), 3825-3833.

[12] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 27(1), 415-444.

[13] Currarini, S., Jackson, M. O., & Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. Econometrica, 77(4), 1003-1045.

[14] Zeng, R., Sheng, Q. Z., Yao, L., Xu, T., & Xie, D. (2013, January). A practical simulation method for social networks. In Proceedings of the First Australasian Web Conference-Volume 144 (pp. 27-34). Australian Computer Society, Inc..

[15] Singer, H. M., Singer, I., & Herrmann, H. J. (2009). Agent-based model for friendship in social networks. Physical Review E, 80(2), 026113.

[16] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

[17] Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). Introduction to algorithms second edition. McGraw-Hill.

[18] Newman, M. (2010). Networks: an introduction. Oxford university press.

[19] Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Laboratory (LANL). ISO 690

[20] Foster, J. G., Foster, D. V., Grassberger, P., & Paczuski, M. (2010). Edge direction and the structure of networks. Proceedings of the National Academy of Sciences, 107(24), 10815-10820.

[21] Newman, M. E. (2003). Mixing patterns in networks. Physical Review E, 67(2), 026126.