Sviluppo di un modello per la simulazione della rete Twitter

Un approccio basato sull'omofilia

Comi, Marco
Università degli Studi di Milano - Bicocca
Dipartimento di Informatica Sistemistica e Comunicazione
email m.comi7@campus.unimib.it

Gravina, Marco
Università degli Studi di Milano - Bicocca
Dipartimento di Informatica Sistemistica e Comunicazione
email m.gravina1@campus.unimib.it

21 febbraio 2017

Sommario

Tra le numerose definizioni di sistema complesso che sono state fornite troviamo quella dei fisici Nigel Goldenfeld e Leo Kadanoff che definiscono un sistema complesso "a highly structured system, which shows structure with variations". Accanto alla loro, i chimici Whitesides e Ismagilov ce ne forniscono un'altra: "a complex system is one whose evolution is very sensitive to initial conditions or to small perturbations, one in which the number of independent interacting components is large, or one in which there are multiple pathways by which the system can evolve"; infine David Rind definisce un sistema complesso "one in which there are multiple interactions between many different components". Partendo da queste definizioni, possiamo definire i social network dei sistemi complessi a tutti gli effetti; abbiamo infatti differenti e numerosi componenti (gli attori o utenti del social network) che interagiscono in modo indipendente facendo "evolvere" il sistema in maniera ogni volta potenzialmente diversa a seconda delle condizioni iniziali o delle "perturbazioni" subite dal sistema. Lo scopo della seguente relazione è quello di presentare un modello a grafo sviluppato per la simulazione delle dinamiche in un social network. Il modello presentato è da considerarsi come un "primo passo" (e non come una soluzione definitiva) verso una rappresentazione più completa e aderente al sistema reale.

Indice

1 Glossario **3**

2 Introduzione **4**

3 Riferimenti Matematici **5**

4 Sviluppo **9**
4.1 Strumenti utilizzati 9
4.2 Inizializzazione della rete 9
4.3 Il concetto di"clock" 10
4.4 Esecuzione del modello 11
4.4.1 Generazione tweet 11
4.4.2 Generazione di retweet 12
4.4.3 Aggiornamento probabilità 13
4.4.4 Aggiornamento archi 13

5 Simulazioni **14**
5.1 Introduzione alle simulazioni 14

6 Analisi dei dati **15**
6.1 Simulazione 1 15
6.1.1 Modifiche per simulazione successiva 22
6.2 Simulazione 2 23
6.2.1 Modifiche per simulazione successiva 27
6.3 Simulazione 3 28
6.3.1 Modifiche per simulazione successiva 31
6.4 Simulazione 4 32

7 Problemi riscontrati **40**
7.1 Problema 1: Clock e saturazione della rete 40
7.2 Problema 2: Lentezza dell'interfaccia grafica dinamica 40
7.3 Problema 3: Tempi di esecuzione degli algoritmi su grafi 41

8 Sviluppi Futuri **42**
8.1 Sviluppo 1: Caratterizzazione di agente 42
8.2 Sviluppo 2: Differenziazione del numero e del tipo interessi . . . 42
8.3 Sviluppo 3: Caratterizzazione di tweet 42
8.4 Sviluppo 4: Dinamicità della rete 43
8.5 Sviluppo 5: Inserimento utenti inattivi 43
8.6 Sviluppo 6: Assegnazione di un tempo ai tweet 43
8.7 Osservazioni 43

9 Conclusioni **45**

Bibliografia **46**

1 Glossario

- **Tweet**: breve messaggio di testo non superiore a 140 caratteri pubblicato sul social network Twitter; con il termine *twittare* si intende l'azione di pubblicazione di un tweet
- **Retweet**: messaggio di lunghezza non superiore a 140 caratteri il cui testo riproduce quello di un altro messaggio con l'aggiunta del nome dell'autore e di un eventuale breve commento; con il termine *retwittare* si intende l'azione di pubblicazione di un retweet
- **Follower**: persona che segue un altro utente su un social media
- **Followee**: persona che viene seguita su un social media
- **Omofilia**: termine coniato dai sociologi Lazarsfeld e Merton che indica la tendenza insita nelle persone ad associarsi ad altre persone con caratteristiche simili alle proprie In [1] viene definita come *the principle that a contact between similar people occurs at a higher rate than among dissimilar people*
- **Omofilia per interesse**: con questo termine, nella trattazione, identificheremo la tendenza di un agente ad associarsi ad altri aventi interessi simili ai propri
- **Omofilia di gruppo**: tendenza ad associarsi a persone appartenenti al proprio gruppo sociale, etnico, a persone dello stesso genere o che svolgono la stessa professione
- **Vip**: nella trattazione, il termine vip è utilizzato con il significato che gli è comunemente attribuito
- **Rete a invarianza di scala**: rete nella quale la probabilità con cui un nuovo nodo si connette ai vertici esistenti non è uniforme, ma è maggiore se la connessione è con un vertice che ha già un grande numero di link [3]

2 Introduzione

Lo scopo della seguente relazione è quello di presentare il modello da noi sviluppato per simulare il comportamento di un social network. Il social network preso in considerazione è Twitter.
La simulazione prevede la rappresentazione di Twitter attraverso un grafo orientato in cui ogni utente è identificato da un nodo e la relazione di "following" è rappresentata mediante un arco orientato che esce dal nodo seguace ed entra nel nodo seguito.
Oltre a generare una simulazione di una rete Twitter il più possibile aderente alla realtà, il modello si propone, in particolare, di simulare il fenomeno per cui un utente scrive un tweet particolarmente interessante che viene notato e retweettato da un utente vip facendo così crescere in maniera esponenziale i follower del primo.
Questo fenomeno, per quanto non così comune, è stato riscontrato in più di un'occasione: è il caso, per esempio, della disputa tra Daniele Termite e l'onorevole Maurizio Gasparri[1], che ha visto i follower di Termite crescere esponenzialmente nel giro di poche ore.

[1] `http://www.corriere.it/politica/12_ottobre_01/gasparri-twitter-guerra-cruccu_901d2214-0beb-11e2-a626-17c468fbd3dd.shtml`

3 Riferimenti Matematici

Nello sviluppo del modello sono stati utilizzati diversi concetti matematici che, per completezza, riportiamo di seguito.

Distribuzione di Poisson
In teoria delle probabilità la distribuzione di Poisson (o poissoniana) è una distribuzione di probabilità discreta che esprime le probabilità per il numero di eventi che si verificano successivamente ed indipendentemente in un dato intervallo di tempo, sapendo che mediamente se ne verifica un numero λ. Questa distribuzione è anche nota come legge degli eventi rari [2]. (Figura 1)

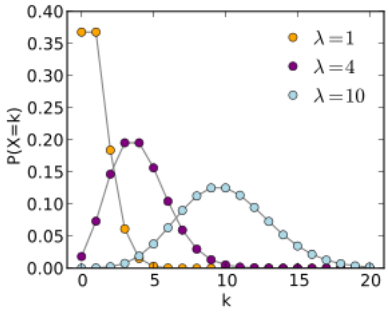


Figura 1: Grafico della distribuzione poissoniana

Distribuzione uniforme
In teoria delle probabilità la distribuzione continua uniforme è una distribuzione di probabilità continua che è uniforme su un insieme, ovvero che attribuisce la stessa probabilità a tutti i punti appartenenti ad un dato intervallo [a,b] contenuto nell'insieme.[3](Figura 2)

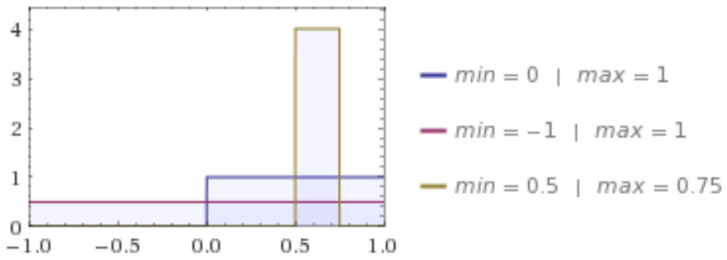


Figura 2: Grafico della distribuzione uniforme

[2] https://it.wikipedia.org/wiki/Distribuzione_di_Poisson
[3] https://it.wikipedia.org/wiki/Distribuzione_continua_uniforme

Distribuzione normale
Nella teoria della probabilità la distribuzione normale o gaussiana, è una distribuzione di probabilità continua che è spesso usata per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valor medio. Il grafico della funzione di densità di probabilità associata è simmetrico e ha una forma a campana (Figura 3), nota come campana di Gauss. La distribuzione normale è considerata il caso base delle distribuzioni di probabilità continue: più specificamente, assumendo certe condizioni, la somma di n variabili casuali con media e varianza finite tende a una distribuzione normale al tendere di n all'infinito. Grazie a questo teorema, la distribuzione normale si incontra spesso nelle applicazioni pratiche, venendo usata come un semplice modello per fenomeni complessi.[4]

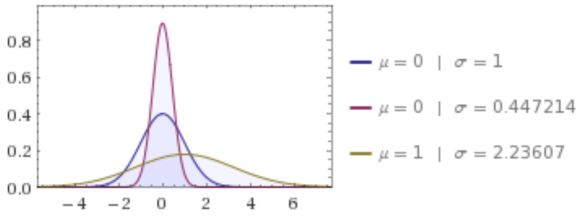


Figura 3: Grafico della distribuzione normale

Funzione dell'omofilia di gruppo
In [4] definiscono il concetto di *omofilia*. Questo concetto risulta essere differente da quello di omofilia tra due agenti; abbiamo quindi deciso di chiamarlo *omofilia di gruppo*.
Currarini et al. suddividono la popolazione N secondo le tipologie i e calcolano il numero medio di relazioni tra agenti dello stesso tipo (s_i) ed il numero medio di relazioni con agenti di tipi differenti (d_i). Formalizzano, quindi, l'omofilia (H_i) come

$$H_i = \frac{s_i}{s_i + d_i}.$$

Abbiamo tradotto questo concetto nel nostro modello identificando due principali gruppi: il gruppo dei *VIP* e quello dei *NON VIP*. Per ognuno dei gruppi identificati è stata calcolata l'omofilia di gruppo dividendo il numero medio di relazioni tra agenti dello stesso tipo per il numero medio totale di relazioni.

Omofilia per interesse
Prendendo spunto dal concetto di omofilia (vedi glossario) abbiamo definito una funzione che calcolasse la probabilità del legame tra due agenti sulla base dei loro interessi. Abbiamo associato ad ogni agente una lista di interessi con

[4] `https://it.wikipedia.org/wiki/Distribuzione_normale`

il relativo peso; l'idea è quella di permettere ad ogni agente di esprimere una preferenza maggiore o minore rispetto a ciascun argomento e di aumentare la probabilità del legame (che chiameremo *omofilia per interesse*) all'aumentare della corrispondenza tra le suddette preferenze.
La formalizzazione della funzione ottenuta è la seguente: siano n il numero di interessi che caratterizzano ciascun agente, i e j rispettivamente l'i-esimo e il j-esimo agente presi in considerazione e w_{ik} il peso attribuito dall'i-esimo agente al k-esimo interesse; l'omofilia per interesse tra l'agente i e l'agente j è

$$omofilia(i,j) = 1 - \sum_{k=1}^{n} [\frac{(w_{ik} - \bar{W})^2 + (w_{jk} - \bar{W})^2}{2}] * w_{ik}.$$

Il valore all'interno della sommatoria rappresenta l'eterofilia tra l'agente i e l'agente j. Per calcolare tale quantità, definiamo il valore $\bar{W}$ come

$$\bar{W} = \frac{w_{ik} + w_{jk}}{2}$$

ovvero il valore medio attribuito dagli agenti i e j all'interesse k. Possiamo ora definire il grado di scostamento tra due agenti con la formula (che richiama quella della varianza)

$$\frac{(w_{ik} - \bar{W})^2 + (w_{jk} - \bar{W})^2}{2}.$$

Per differenziare l'omofilia tra l'agente i e l'agente j da quella in senso opposto, moltiplichiamo questo valore per il peso attribuito dal primo agente all'interesse in questione. Abbiamo quindi ottenuto l'eterofilia tra i e j; sottraendo all'unità questo valore otteniamo la probabilità da noi desiderata.

É importante sottolineare come, in virtù di quanto appena descritto, nel nostro modello consideriamo la relazione di omofilia per interesse una relazione non simmetrica; ci è sembrato logico fare in modo che il valore dell'omofilia dal nodo i al nodo j fosse diverso da quello calcolato nella direzione opposta in quanto, anche nella realtà, se due persone hanno gli stessi interessi, non è detto che attribuiscano lo stesso peso a ciascuno di essi.
Prendiamo come esempio due persone A e B che possiedono uno stesso interesse per lo sport; può accadere che A sia una persona molto sportiva e che ritenga la sportività un fattore rilevante nello stabilire una relazione. B potrebbe essere invece un tifoso che considera la sportività un fattore meno predominante e più come un "valore aggiunto" nell'equazione della instaurazione di una relazione. Sebbene entrambi considerino la sportività come una variabile rilevante per stabilire una relazione, le attribuiscono un valore differente. Ecco quindi la motivazione che ci ha spinto a considerare l'omofilia da A a B è differente da quella da B ad A.

Invarianza di scala
Il termine *rete a invarianza di scala* fu coniato da Albert-Làzlò Barabasi, scienziato ungherese, nel 1998. Questo tipo di rete da lui teorizzato, che poi ha

trovato fondamento empirico, si distingue dal modello precedentemente dominante introdotto da Erdős–Rényi, anche noto come modello dei grafi casuali, in cui il collegamento tra due vertici di un grafo avveniva in maniera totalmente casuale. In una rete ad invarianza di scala, invece, la formazione di nuovi link avviene secondo quella che viene definita legge di potenza. In generale, quindi, l'invarianza di scala è quel fenomeno secondo cui un nodo tende a stabilire una relazione con nodi che ne hanno già molte.

4 Sviluppo

In questa sezione descriveremo nel dettaglio il metodo con cui abbiamo sviluppato il modello e le principali funzionalità che lo caratterizzano.

4.1 Strumenti utilizzati

Per lo sviluppo del modello e per la realizzazione dei grafici utili a questa relazione sono stati relizzati strumenti differenti.
Il codice è stato scritto in linguaggio Python, utilizzando la libreria NetworkX[5], appartenente al progetto PyCX[6], che permette di simulare vari sistemi complessi utilizzando una GUI già implementata.
Visti i problemi di performance riscontrati, per la visualizzazione grafica[7] della rete generata abbiamo utilizzato la piattaforma Gephi[8] che consente di visualizzare ed esplorare tutti i tipi di grafi e di reti.
Infine, per la realizzazone dei grafici abbiamo utilizzato un semplice foglio di calcolo.

4.2 Inizializzazione della rete

All'avvio del programma viene inizlializzata la rete, con il numero di nodi prestabilito e con la creazione di archi tra vari nodi. Ogni nodo è caratterizzato da una serie di parametri:

- identificativo del nodo
- flag per un nodo vip: impostato a 0 se il nodo non è vip, a 1 altrimenti
- array di interessi del nodo con relativo peso: ciascun interesse (nelle nostre simulazioni abbiamo tenuto conto di 5 interessi) è identificato da un valore numerico univoco e da un peso (da 1 a 10); a differenza degli interessi, che sono comuni a tutti i nodi della rete, i pesi associati a ciascuno di essi variano da nodo a nodo, così come ogni persona può trovare più o meno interessante un determinato argomento
- numero di nodi seguiti (followers)
- numero di nodi da cui si è seguiti (followees)
- probabilità del nodo di seguirne un altro: è un parametro intrinseco del nodo e differente per i nodi vip ed i nodi non vip. In base alla tipologia dell'agente (vip o non vip), questo parametro assume valori in un diverso intervallo definito in una variabile del modello. Attualmente, in quanto caratteristica propria del nodo, è un valore statico

[5] https://networkx.github.io
[6] http://pycx.sourceforge.net
[7] Per ulteriori dettagli, si veda il problema 2 della sezione "Problemi riscontrati"
[8] https://gephi.org

- probabilità del nodo di essere seguito da un altro: è un parametro inizialmente stabilito in maniera casuale all'interno di un range di valori differente per i nodi vip ed i nodi non vip. È un valore dinamico, viene perciò modificato durante la simulazione
- probabilità del nodo di fare un retweet: rappresenta la probabilità intrinseca del nodo di effettuare un retweet e varia a seconda della tipologia del nodo. Questo parametro è statico durante l'esecuzione del modello
- probabilità del nodo di essere retweettato: rappresenta la probabilità intrinseca del nodo di essere retweettato e varia a seconda della tipologia del nodo. Questo parametro viene modificato, temporaneamente, per simulare l'effetto di un tweet particolarmente interessante da parte di un agente
- array di tweet generati: attualmente contiene l'identificativo dell'ultimo tweet generato dall'agente[9]
- array contenente gli id dei nodi che hanno retwittato il nodo in questione

All'inizializzazione un parametro stabilisce il numero di nodi etichettati come "vip" assegnando loro i valori di probabilità di essere seguiti e di essere retwittati conformi alla loro categoria. La creazione di un arco da un nodo i ad un nodo j avviene sulla base di una probabilità calcolata come la congiunta tra l'omofilia per interesse del nodo i nei confronti del nodo j e la probabilità del nodo j di essere seguito.
Per evitare la rapida saturazione della rete, abbiamo introdotto un filtro che stabilisse se effettuare il calcolo appena descritto. Il meccanismo di creazione degli archi basato sull'omofilia ha permesso di avere una topologia della rete il più possibile conforme ai parametri di inizializzazione.
Una serie di altre funzioni permettono di adattare i rimanenti parametri di ciascun nodo alla nuova topologia della rete.
Nel seguito di questa sezione daremo spiegazione della logica utilizzata nello sviluppo delle funzionalità che costituiscono il nucleo della simulazione.

4.3 Il concetto di"clock"

Prima di procedere con la descrizione dell'esecuzione del modello, è necessario introdurre il concetto di "clock".

Il clock costituisce uno step della nostra simulazione al quale abbiamo attribuito un valore temporale. Nello specifico abbiamo associato ad ogni clock un valore pari a sei ore in modo da avere così una giornata suddivisa in quattro parti; le quattro fasce orarie da noi identificate sono:

- **fascia 1:** 00:00 - 06:00
- **fascia 2:** 06:00 - 12:00

[9]Per suggerimenti rimandiamo alla sezione Sviluppi futuri

- **fascia 3:** 12:00 - 18:00
- **fascia 4:** 18:00 - 00:00

Abbiamo optato per questa suddivisione in seguito ad una serie di problemi riscontrati durante lo sviluppo e le prime simulazioni[10].

4.4 Esecuzione del modello

Definito il concetto di tempo per il nostro modello e descritto il processo di inizializzazione della rete, possiamo dettagliare la logica con cui abbiamo sviluppato il comportamento del modello.
L'esecuzione passa attraverso le funzioni *step()* e *draw()* che eseguono, rispettivamente, l'aggiornamento dei parametri della rete ed il rendering grafico della rete stessa. Le operazioni eseguite ad ogni step sono le seguenti:

- generazione di tweet
- generazione di retweet
- aggiornamento probabilità
- aggiornamento archi
- aggiornamento di followees, followers
- ricalcolo dei nodi vip
- rendering grafico

4.4.1 Generazione tweet

La generazione di tweet avviene utilizzando parametri diversi (i parametri lambda della distribuzione di Poisson) per i nodi vip e non vip. L'idea è quella di distinguere il modo di interagire con il social network, oltre che in base alla fascia oraria, anche in base al tipo del nodo.
Per ogni nodo viene generato un numero di "tweet che è possibile fare" secondo una distribuzione poissoniana con parametro λ associato al tipo di nodo ed alla fascia oraria del tweet. I tweet generati, identificati da un valore numerico, univoco e progressivo, sono associati al nodo che li ha generati.
Abbiamo inoltre inserito un fattore "tweet di qualità" per aggiungere alla simulazione il caso in cui una persona aumenta la propria probabilità di essere retwittata in seguito al post di un tweet particolarmente sagace o interessante. Questo tipo di fenomeno è stato reso aumentando, per un singolo step, la probabilità di essere retwittato dell' agente che ha effettuato il tweet, in quanto il momento più "virale" è quello immediatamente successivo all'evento.

[10] Per chiarimenti rimandiamo al "problema 1" della sezione "Problemi riscontrati"

4.4.2 Generazione di retweet

Il meccanismo di retweet implementato segue, invece, una logica simile ma leggermente più complessa rispetto a quella dei tweet, volta a ritardare la saturazione della rete e, soprattutto, a fornire al modello un comportamento più realistico.
Anche in questo caso viene generato un numero secondo la distribuzione di Poisson con le stesse modalità con cui viene fatto per i tweet. Chiamiamo questo numero "numero di coins" e consideriamolo come un insieme di gettoni, ovvero il numero di possibilità che un agente ha di effettuare un retweet. Ad esempio, un agente i che ha un quattro coins, avrà quattro possibilità di effettuare un retweet (in altri termini effettuerà un massimo di quattro retweet per quella fascia oraria).
Essendo il numero di coins generato secondo una distribuzione Possoniana, abbiamo la garanzia che il comportamento del nodo i sia conforme a quello degli altri nodi dello stesso tipo per la fascia oraria in questione. Per stabilire quale nodo sarà retwittato, viene eseguita una scansione sequenziale della lista contenente tutti i nodi della rete.

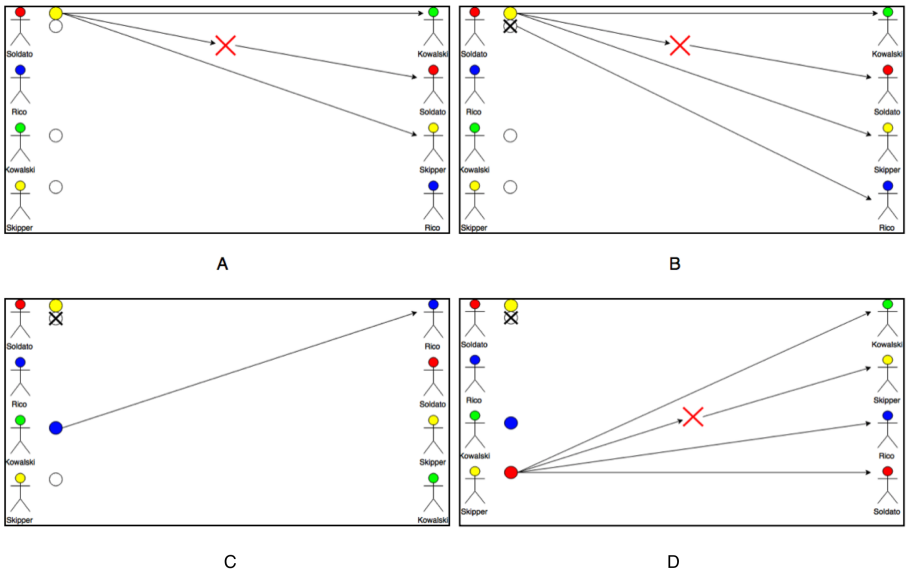


Figura 4: A) L'agente denominato Soldato assegna il primo retweet all'agente chiamato Skipper non considerando se stesso; B) Il secondo coin non viene assegnato; C) L'agente Kowalski assegna il coin all'agente Rico; D) L'agente Skipper assegna il retweet all'agente Soldato non considerando se stesso.

Per garantire una distribuzione omogenea dei retweet sui nodi, evitando di favorire con questo metodo i primi nodi presenti nella lista (in quanto più "scansionati"), abbiamo utilizzato un meccanismo di randomizzazione della posizione

dei nodi della rete nella lista.
Il meccanismo dei coins appena introdotto è rappresentato in Figura 4.
Abbiamo infine introdotto un fattore *penalità* che riduce la prbabilità di retwittare un nodo e che è maggiore se il nodo retwittato non è un vicino (non è seguito) del nodo retwittante; abbiamo deciso di introdurre questo fattore, perchè, facendo riferimento al funzionamento di Twitter, ogni utente può retwittare persone che non segue, ma è più probabile che retwitti qualcosa che viene pubblicato da una persona da lui seguita. Nel codice relativo alla generazione di retweet, abbiamo inoltre implementato un meccanismo per riconoscere e memorizzare in un file di testo i retweet effettuati da un agente vip dei tweet di un agente non vip.

4.4.3 Aggiornamento probabilità

Dopo la generazione di tweet e retweet, viene effettuato un aggiornamento delle probabilità sulla base del numero di volte che un agente è stato retwittato. Nello specifico viene calcolato un "bonus" utilizzato per aumentare in percentuale le probabilità di essere retwittato e di essere seguito del nodo in questione. Questo bonus è pesato in base alla natura dell'agente retwittante; il retweet da parte di un agente vip è considerato più "importante" (e dunque ha un peso maggiore) di quello effettuato da uno non vip. In seguito, viene effettuato un controllo sulle probabilità aggiornate per evitare che superino il valore di 0.9 ed è svuotata la lista contenente gli identificativi degli agenti che hanno retwittato il nodo preso in analisi; questo serve ad evitare che, negli step successivi, retweet già considerati influiscano nuovamente sul ricalcolo delle probabilità.

4.4.4 Aggiornamento archi

La prima operazione effettuata nel procedimento di aggiornamento degli archi è il ricalcolo dell'omofilia di gruppo (che potrebbe essere variata se è cambiata la proporzione tra vip e non vip). Sono calcolate l'omofilia per interesse di ogni agente con ogni altro agente e la possibilità di creare un arco tra essi. Per ogni coppia di agenti i e j, questa possibilità è calcolata utilizzando la probabilità congiunta tra l'omofilia del gruppo dell'agente i, l'omofilia per interesse da i a j, la probabilità del nodo i di seguire e la probabilità dell'agente j di essere seguito. Abbiamo infine introdotto un filtro per limitare l'"esplosione" del numero di nuovi archi creati.

Dopo aver effettuato le modifiche elencate, sono aggiornati il numero di followers e di followees di ciascun nodo. L'ultima funzione eseguita ad ogni clock è il *draw()*: in questa funzione si valuta il cambiamento di stato (da non vip a vip) di ogni agente, è ricalcolato il numero di agenti per ciascuna categoria e sono salvati, nei formati opportuni, i dati utili alla fase di analisi. Generalmente la funzione *draw* si occupa di "ridisegnare" la rete avvalendosi dell'interfaccia

grafica dinamica della libreria; il peso computazionale di questa operazione è stato uno dei problemi che abbiamo dovuto affrontare (approfondito nella sezione "Problemi riscontrati") e che ci ha spinto a non utilizzare il rendering grafico offerto da NetworkX preferendogli l'utilizzo di Gephi.

5 Simulazioni

In questa sezione spiegheremo come abbiamo effettuato le simulazioni, con quali criteri e parametri e perchè abbiamo deciso di testare due modelli differenti.

5.1 Introduzione alle simulazioni

Uno dei problemi riscontrati durante la fase di esecuzione delle nostre simulazioni è stato quello dell'"espolosione" del numero di vip all'interno della rete.
Dopo un certo periodo di tempo (non troppo lungo) riscontravamo un numero di nodi vip eccessivo per poter considerare la simulazione anche solo vicina ad un caso reale.
Partendo da questo problema abbiamo pensato che potesse essere interessante indagare le cause di questo fenomeno ed analizzare il comportamento della rete in due casi differenti; abbiamo dunque distinto le simulazioni effettuate in due categorie:

- **Simulazioni con l'AND**
- **Simulazioni con l'OR**

Le prime sono le simulazioni per cui il passaggio di un nodo dallo stato "non vip" a quello "vip" è stato calcolato verificando che fossero soddisfatte entrambe le condizioni *probabilità di essere seguito maggiore di un certo parametro* e *numero di follower maggiore di una certa percentuale*; le seconde sono, invece, quelle per cui le condizioni enunciate sono state poste in OR, ovvero è sufficiente che almeno una delle due condizioni sia valida per far diventare vip l'utente.

Per ciascuno dei due casi sono state effettuate quattro simulazioni con gli stessi parametri; ciò significa che la prima simulazione del "ramo AND" è stata effettuata con i parametri impostati agli stessi valori della prima simulazione del "ramo OR" e così via.
Questa scelta è stata fatta allo scopo di valutare quale delle due condizioni identificasse più fedelmente il comportamento reale della rete di Twitter.
Di seguito presentiamo i risultati per ciascuna simulazione di ognuno dei due casi esposti.

6 Analisi dei dati

In questa sezione sarà fatta un'analisi dei dati delle simulazioni effettuate; è utile ricordare che ogni simulazione è eseguita per 360 step che, in termini temporali, equivalgono a 90 giorni.

6.1 Simulazione 1

La prima simulazione è stata inizializzata con in seguenti parametri:

- numero di nodi della rete: **1000**
- percentuale di vip all'inizializzazione: **0.2**

INTERVALLI PROBABILITÀ VIP

	MIN	MAX
SEGUIRE	0.01	0.1
ESSERE SEGUITO	0.6	0.8
RETWITTARE	0.01	0.1
ESSERE RETWITTATO	0.6	0.7

INTERVALLI PROBABILITÀ NON VIP

	MIN	MAX
SEGUIRE	0.3	0.5
ESSERE SEGUITO	0.01	0.15
RETWITTARE	0.5	0.75
ESSERE RETWITTATO	0.05	0.15

Tabella 1: Intervalli di probabilità

LAMBDA TWEET

	00:00-06:00	06:00-12:00	12:00-18:00	18:00-24:00
VIP	1	0.5	1.5	1.7
NON VIP	0.3	0.7	0.7	1

LAMBDA RETWEET

	00:00-06:00	06:00-12:00	12:00-18:00	18:00-24:00
VIP	0.5	0.7	1	1.2
NON VIP	1	2	1.7	1

Tabella 2: Valori medi di retweet e tweet per fasce orarie

- numero interessi per ogni agente: **5**
- massimo peso possibile per ogni interesse: **10**

- probabilità di effetuare un tweet interessante: **0.01**
- bonus attribuito dopo aver effettuato un tweet interessante: **0.2**
- bonus attribuito al retweet da parte di un vip: **2**
- bonus attribuito al retweet da parte di un non vip: **1**

Di seguito, i risultati della prima simulazione per entrambi i modelli.
Caso AND

La Figura 5 rappresenta uno screen dello stato della rete al termine dell'esecuzione del modello.

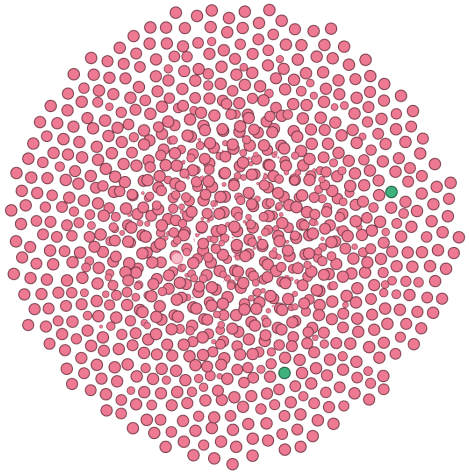


Figura 5: Stato della rete al termine della simulazione 1

I nodi colorati di verde sono i nodi vip e la dimensione di ciascun nodo è proporzionale al numero di archi entranti (e dunque ai suoi follower). Possiamo notare come, sebbene molti nodi abbiano un numero di follower molto elevato e prossimo a quello dei nodi vip, nessuno di loro sia diventato vip; il numero di vip è infatti rimasto invariato.
Il risultato ottenuto è coerente con quanto ci aspettavamo per questo modello in quanto la condizione che consente ad un nodo di diventare vip è più forte in questa simulazione piuttosto che nell'altra.
Riportiamo ora in Figura 6 i dati relativi ad alcuni nodi particolarmente significativi che abbiamo analizzato. Il primo è uno dei due nodi vip. Le probabilità di essere seguito e di essere retwittato del nodo vip crescono rapidamente nei primi step della simulazione assestandosi al valore massimo consentito dal modello di 0.9. Questo tipo di comportamento è plausibile in quanto abbiamo molti nodi non vip che, grazie all'invarianza di scala, tendono a seguire e retwittare i nodi più popolari. Un valore di probabilità così elevato ci permette inoltre di generare una rete con una topologia conforme ai parametri inseriti all'inizializzazione

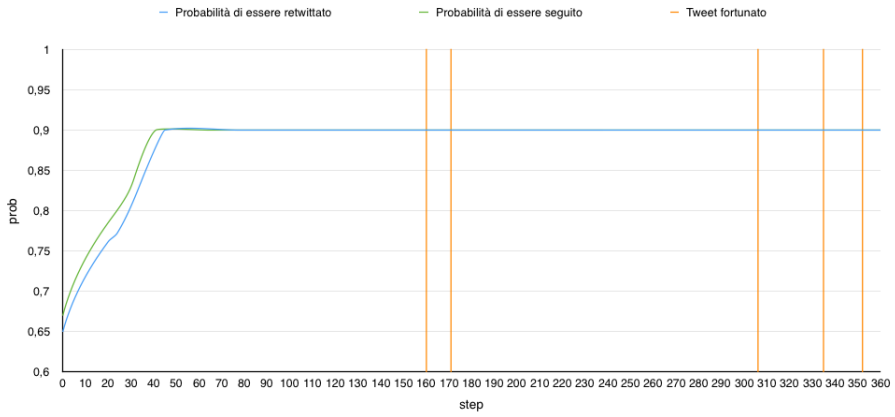


Figura 6: Probabilità di essere seguito e di essere retwittato di un nodo vip

del modello; i nodi vip avranno quindi più archi entranti (e dunque più follower) rispetto ai nodi non vip. È opportuno segnalare come, nel caso di un nodo che ha già raggiunto il massimo valore di probabilità (impostata a 0.9 per lasciare un minimo di "incertezza"), il tweet fortunato non sembri sortire alcun effetto sul valore di probabilità. In realtà il valore di probabilità viene incrementato anche per i nodi che hanno già raggiunto il valore massimo (con un opportuno controllo per evitare che tale valore superi l'unità). L'incremento resta valido per un singolo turno, poi la probabilità viene reimpostata al suo valore originario. Tale variazione viene però effettuata prima del salvataggio della rete e, pertanto, non risulta nelle statistiche. In Figura 7 sono rappresentate le curve di crescita del numero di follower e di followee dell'agente che stiamo analizzando. Dal grafico

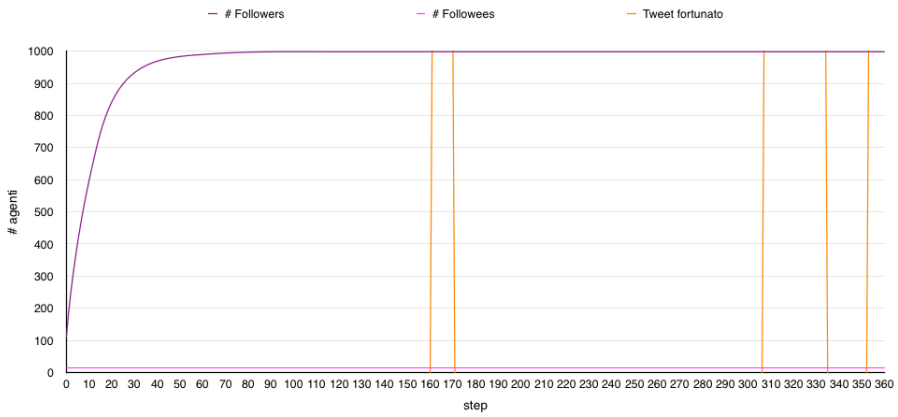


Figura 7: Numero di follower e di followee di un nodo vip

è evidente come il numero di follower del nodo cresca rapidamente in un breve periodo di tempo, mentre il numero di followee si mantenga costante per tutta la simulazione. Se il primo risultato è parzialmente accettabile, il secondo ha senz'altro un margine di miglioramento.
Abbiamo ritenuto interessante analizzare questi stessi valori per due nodi non vip che avessero, al termine della simulazione, caratteristiche tra loro differenti. Il grafico di Figura 8 riporta le probabilità di un nodo che al termine della simulazione presenta un numero di follower e un valore delle probabilità più modesto rispetto alla media. Dato il valore di probabilità così basso, ci aspettiamo un

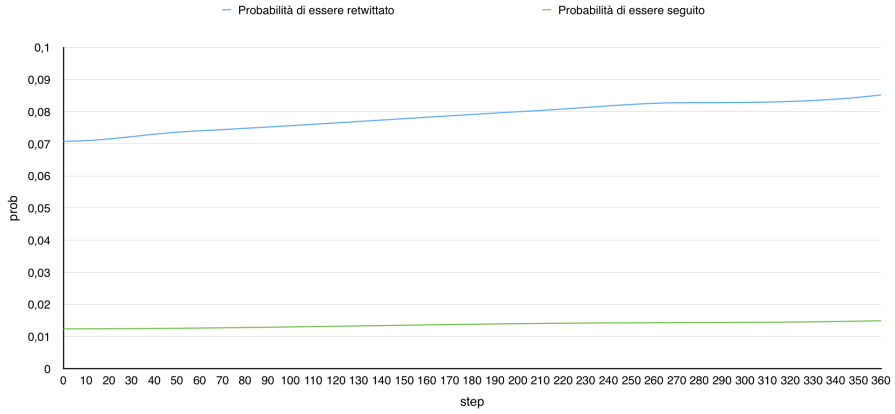


Figura 8: Probabilità di seguire e di essere seguito di un nodo non vip

numero di follower e di followee inferiore a quello visto precedentemente. Questa ipotesi trova conferma nel grafico dei follower/followee sottostante del nodo in questione (Figura 9).

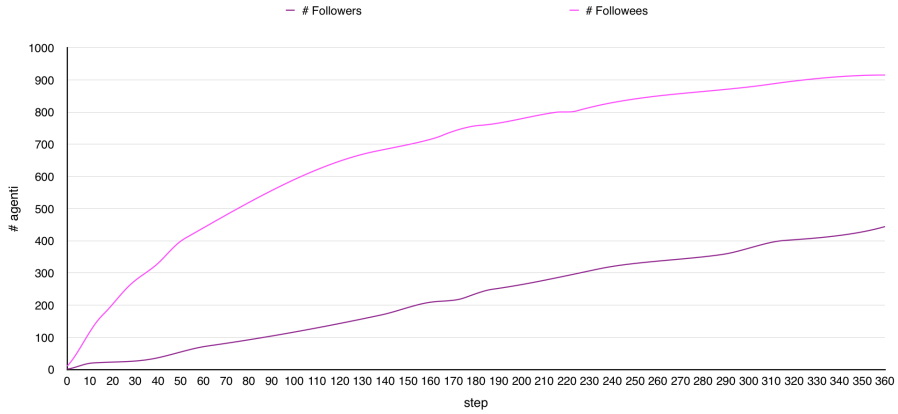


Figura 9: Numero di follower e di followee di un nodo non vip

È interessante notare come la curva di crescita del numero di followee segua una tendenza logaritmica; questo fenomeno è strettamente dipendente dal fatto che, nello stabilire un arco tra due nodi, uno degli elementi più rilevanti è l'omofila di gruppo.
Tale rilevanza scaturisce dal numero molto elevato di agenti dello stesso tipo (in questo caso non vip) e dal valore della loro omofilia di gruppo. Infine, presentiamo i grafici di un nodo con una "storia" più varia rispetto a quella precedentemente vista.

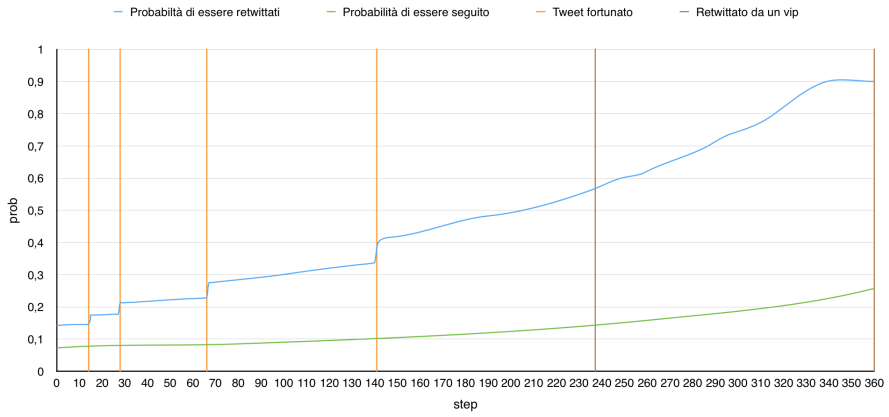


Figura 10: Probabilità di seguire e di essere seguito di un nodo non vip

Notiamo in Figura 10 come vi sia, in corrispondenza di quelli che abbiamo definito "tweet fortunati", un consistente incremento delle probabilità di essere retwittati.
Nel caso del nodo in questione abbiamo anche un considerevole aumento della probabilità di essere seguito; ciò è più evidente in questo nodo rispetto a quello vip analizzato in precedenza in quanto questa probabilità viene aumentata in percentuale e, dunque, risulta essere maggiore per i nodi con una probabilità iniziale più elevata. È opportuno far notare come, invece, il retweet da parte di un vip non sembri avere effetti rilevanti sulle probabilità o sul numero di follower del nodo. A questo proposito riportiamo anche il grafico dei follower/followee anche per questo terzo nodo (Figura 11). Passiamo ora all'analisi dei risultati del modello analogo a quello appena visto ma meno stringente.

Caso OR

Contrariamente a quanto descritto per il modello precedente, la situazione finale della rete in questo modello è quella mostrata in Figura 12. Questa immagine evidenzia come funziona la condizione meno stringente, infatti tutti gli utenti sono diventati vip e solo alcuni hanno pochi archi in ingresso. È possibile rilevare questo dato anche nella Figura 13, la quale rappresenta l'andamento del numero di nodi vip e dei nodi non vip nel corso della simulazione.
Dai grafici precedenti risulta evidente quanto differente sia questo modello da

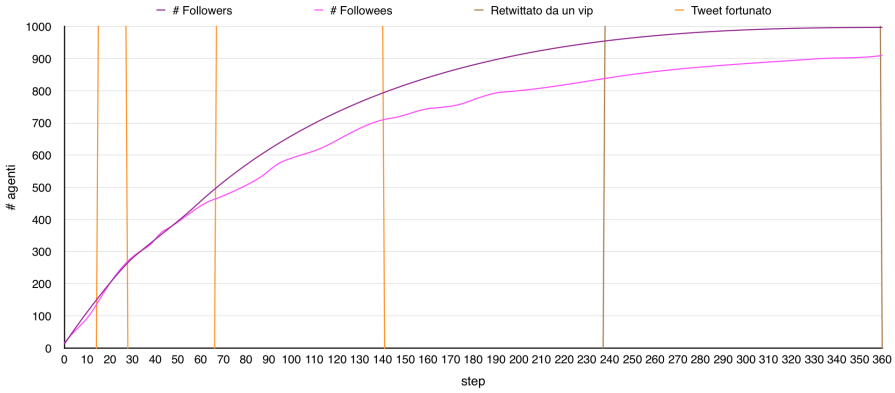


Figura 11: Numero di follower e di followee di un nodo non vip

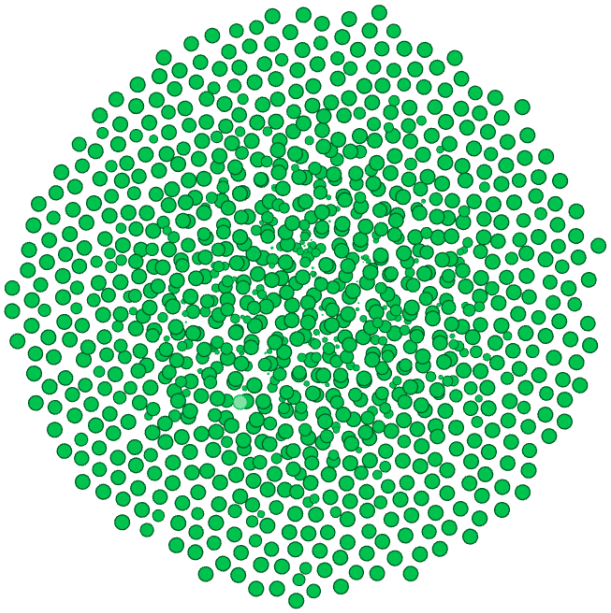


Figura 12: Stato della rete al termine della simulazione 1

quello in AND; in pochi step infatti ogni nodo della rete diventa vip.
Se escludiamo i primi venti step, che per il nostro modello sono "di assestamento" e per i quali non viene considerata la condizione in OR, i nodi della rete diventano vip appena viene valutato il numero di follower.
Analizzando uno dei nodi impostati come vip sin dall'inizializzazione della rete, abbiamo trovato un'ulteriore differenza tra i due modelli: se per il numero di follower, infatti, la tendenza è la stessa, non possiamo dire la stessa cosa per

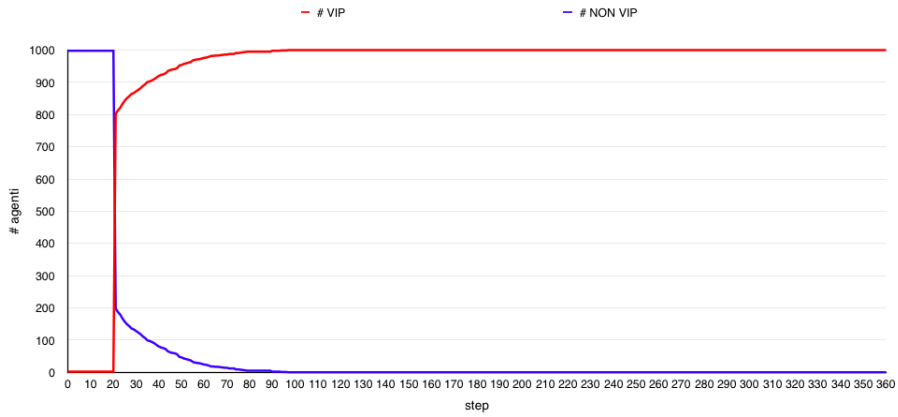


Figura 13: Numero di utenti vip e non vip

quanto riguarda il numero di followee.

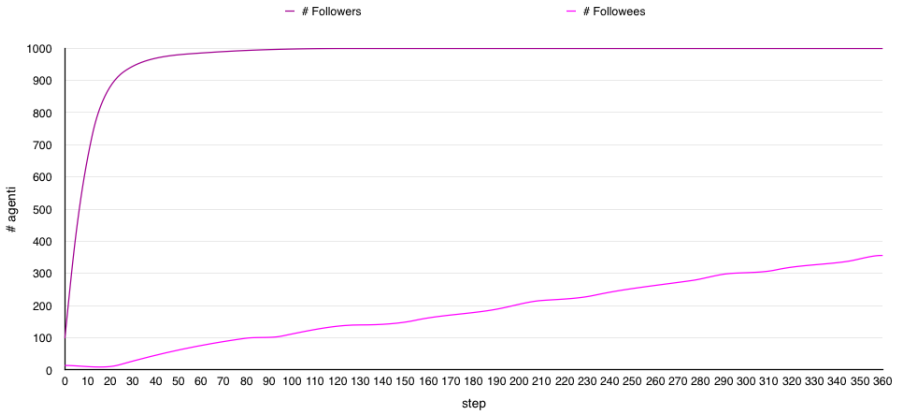


Figura 14: Follower e followee del nodo vip

Questa differenza è data dal fatto che, in questo modello, il numero di agenti vip aumenta molto velocemente, facendo crescere così il valore dell'omofilia di gruppo che influisce sulla dinamica di creazione degli archi.
Verranno analizzati i dati di agenti che all'inizio della simulazione non erano vip, ma che lo sono diventati successivamente.
Il seguente nodo riguarda un utente che è diventato vip subito dopo gli step che, in precedenza, abbiamo chiamato "di assestamento".
In questo caso, l'agente diventa vip allo step venti a causa del numero di follower; l'andamento logaritmico di questi due dati deriva dal fatto che il numero di nodi che diventano vip aumenta rapidamente e quindi gli agenti hanno una tendenza a seguirsi e di essere seguiti maggiore. Inoltre, questo utente ha effet-

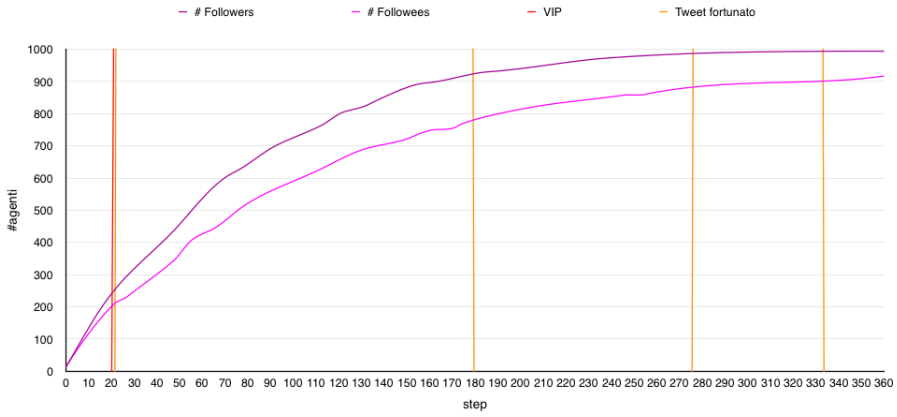


Figura 15: Follower e followee di un nodo nato non vip

tuato anche dei tweet interessanti che hanno aumentato la probabilità di essere retwittato, come mostrato in Figura 16.

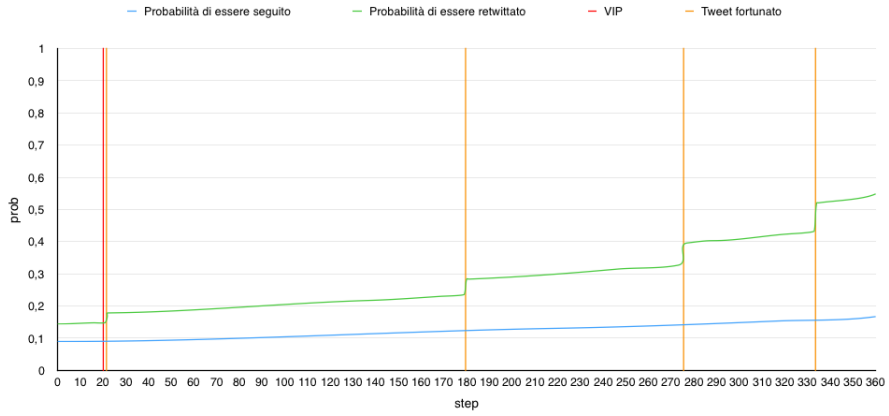


Figura 16: Probabilità di un nodo nato non vip

Per quanto riguarda la sua probabilità di essere seguito, questa non cresce in maniera rilevante dato che l'incremento è in percentuale e che il valore è molto basso sin dall'inizio.

6.1.1 Modifiche per simulazione successiva

Alla luce dei dati analizzati per entrambe le simulazioni, abbiamo deciso che una prima modifica da apportare dovesse essere quella di modificare i valori dei filtri in modo tale che fossero generati meno archi tra i nodi. Abbiamo inoltre ristretto l'intervallo di probabilità di seguire e la probabilità di un vip di essere

seguito.
Abbiamo infine modificato il valore delle penalità sui retweet in modo da renderli meno probabili.

6.2 Simulazione 2

In seguito alle modifiche descritte sopra, ci aspettiamo di rilevare, nell'analisi dei dati, una sostanziale diminuzione nel numero di archi (quasi 900.000 nella prima simulazione) creati al termine dell'esecuzione del modello; ci aspettiamo, inoltre, che il numero dei follower dei vip cresca meno rapidamente rispetto alla simulazione 1.
Per maggiore chiarezza, cercheremo di mantenere uno schema comune nell'analisi dei dati; ciò significa che, laddove possibile (e utile), analizzeremo sempre il grafico dell'andamento del numero di vip e del numero di non vip, il grafico di un nodo "nato vip", quello di un nodo diventato vip nel corso dell'esecuzione del modello e quello di un nodo nato non vip e rimasto tale durante la simulazione. Nel caso in cui, dall'analisi dei file .gephi della rete, dovessero emergere altri aspetti degni di analisi, non ci tratterremo dall'analizzarli.
Caso AND

Dall'analisi del numero di vip e non vip al termine della simulazione emerge come, anche nel corso di questa esecuzione, la situazione non sia cambiata rispetto al caso precedente; dopo 360 step, infatti, abbiamo ancora due soli nodi vip. Riscontriamo tuttavia una distinzione maggiore, rispetto alla simulazione precedente, tra la dimensione (e dunque il numero di follower) dei nodi non vip e quella dei nodi vip. Per quanto riguarda la diminuzione del numero di archi, è evidente come la modifica dei parametri abbia influito notevolmente sulla generazione di nuove relazioni tra i nodi: al termine della simulazione 2 abbiamo 300 000 archi in meno rispetto alla precedente per un valore di circa 600 000.

Una simile considerazione può essere fatta per l'andamento della curva delle probabilità di seguire ed essere seguito del nodo vip, le cui differenze possono essere imputate semplicemente ai diversi valori di inizializzazione dei parametri osservati. È invece interessante osservare il grafico di Figura 18: notiamo infatti come, sebbene alla fine della simulazione il nodo sia comunque seguito da tutti gli agenti della rete, le modifiche effettuate dopo la simulazione 1 abbiano portato a "ritardare" questo fenomeno. Resta invece invariato l'andamento costante del numero di agenti seguiti dal nodo vip, risultato che non ci sorprende dal momento che non abbiamo fatto alcuna modifica volta a cambiare la situazione.
Non avendo la possibilità di analizzare nodi che sono diventati vip durante la simulazione, analizziamo i grafici di un nodo non vip. Notiamo come l'andamento della curva dei nodi seguiti abbia, in questa simulazione, un andamento più simile a quello lineare che a quello logaritmico riscontrato nella simulazione precedente. Questo può essere imputato al ridimensionamento delle probabilità di seguire un altro nodo ed alla modifica dei filtri che regolano la creazione degli archi. Sebbene il numero dei follower di un nodo non vip sia ancora elevato, ne riscontriamo un leggero calo (si consideri che il nodo analizzato è uno dei nodi con le dimensioni più prossime a quelle di un nodo vip e che, pertanto, le con-

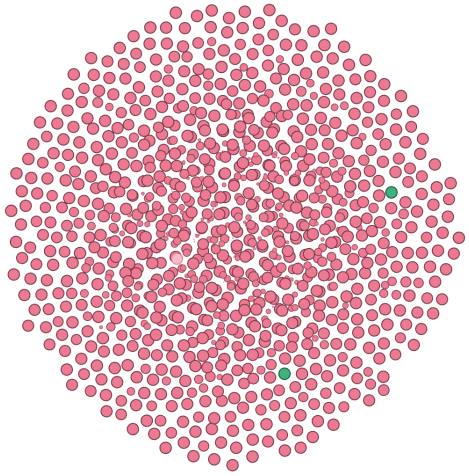


Figura 17: Stato della rete al termine della simulazione 2

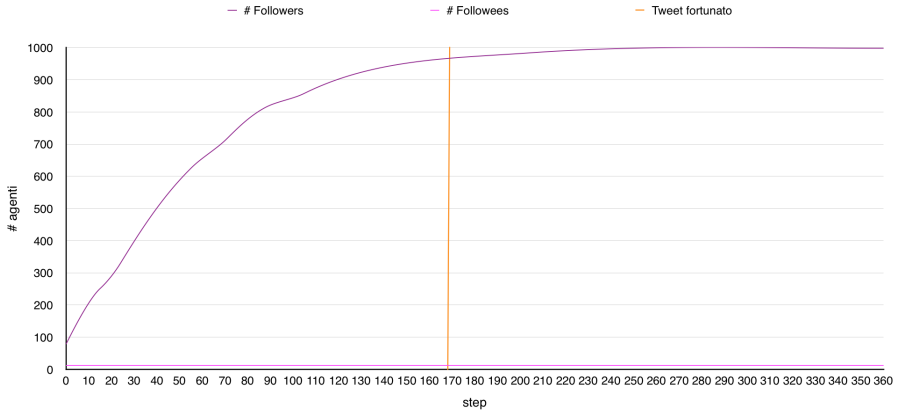


Figura 18: Probabilità di un nodo nato vip

siderazioni fatte circa il numero dei follower possono essere estese anche a gran parte degli altri). Il comportamento della curva della probabilità (Figura 20) di essere seguito non presenta cambiamenti degni di nota.

Caso OR

Come detto per il ramo AND, anche in questo caso c'è stata una significativa diminuzione degli archi creati durante la simulazione; infatti, alla fine della simulazione sono presenti meno di 600 000 archi. Nonostante la situazione finale non sia cambiata, ovvero tutti i nodi sono diventati vip, i dati evidenziano quanto ci aspettavamo con le modifiche apportate.

In Figura 21 viene mostrato il grafico del numero di vip e non vip durante

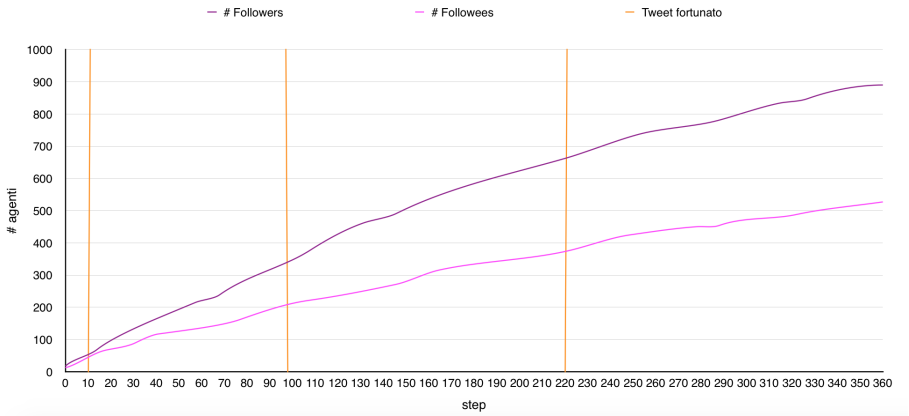


Figura 19: Follower e followee di un nodo non vip

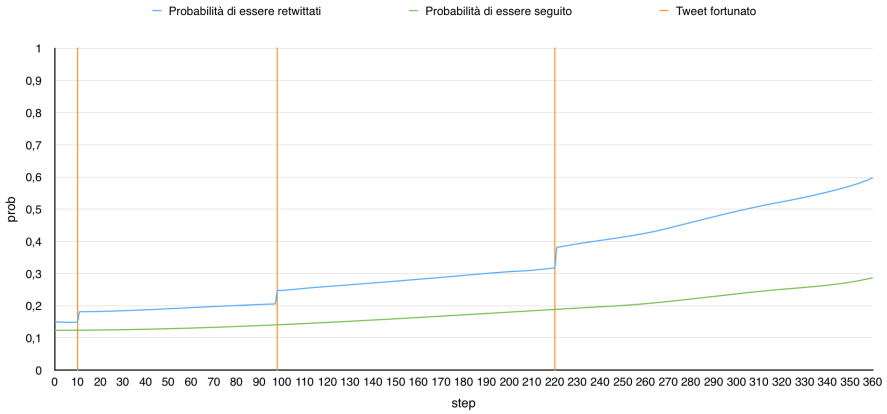


Figura 20: Probabilità di un nodo non vip

l'esecuzione del modello; si vede chiaramente come, anche questa volta, questo andamento sia asintotico verso i limiti superiore ed inferiore, ovvero i vip tendono ad aumentare fino ad arrivare a 1000, mentre i non vip decrescono fino ad arrivare a 0, ma questo andamento è meno rapido rispetto alla simulazione precedente.
Anche per quanto riguarda l'andamento dei follower e dei followee del nodo nato vip si vede come la crescita della funzione abbia rallentato; questo fatto è strettamente collegato al rallentamento del cambio di stato dei nodi da non vip a vip.

Per questa simulazione non viene mostrato lo stato della rete visualizzabile tramite la piattaforma Gephi in quanto non si evidenziano notevoli modifiche rispetto alla simulazione precedente, ma andremo comunque ad analizzare i dati

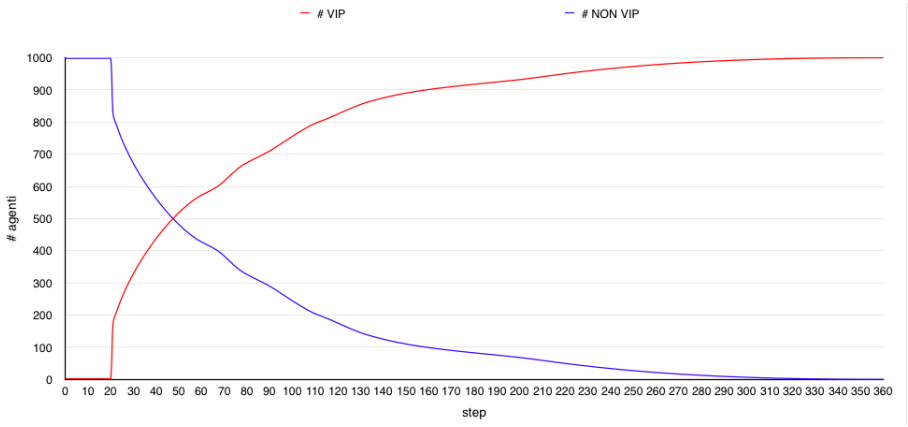


Figura 21: Numero di vip e non vip

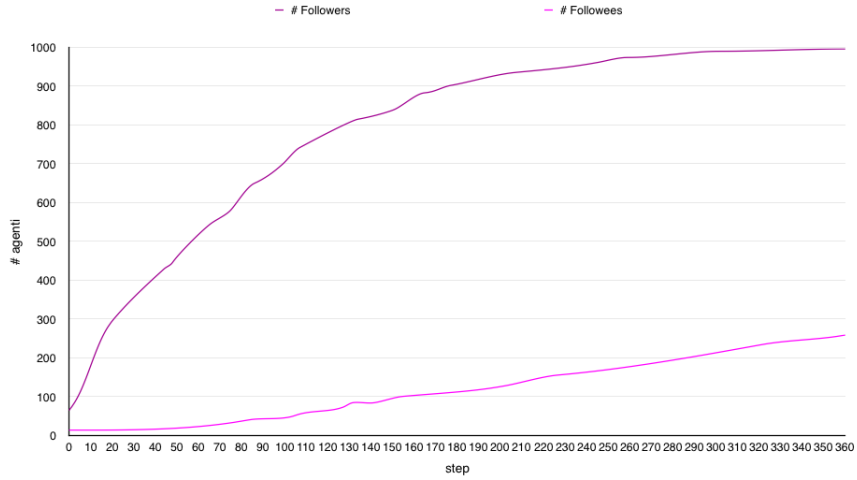


Figura 22: Follower e followee di un nodo nato vip

di alcuni nodi.

La Figura 23 mostra il numero di follower e followee di un nodo nato non vip; anche in questo caso non vi è una crescita molto rapida.
Il grafico mostra come questi due valori crescano in maniera simile; in particolare quando il nodo diventa vip questo valore inizia a crescere più rapidamente. Ciò accade perchè la rete ha un numero di vip molto elevato.

È interessante analizzare le probabilità intrinseche di questo nodo, come visibile nella Figura 24.
Nonostante non si riscontrino evidenti cambiamenti rispetto alla precedente simulazione, risuta evidente la notevole quantità di retweet; è opportuno far notare come i due modelli si comportino in maniera diversa, infatti, se nel modello

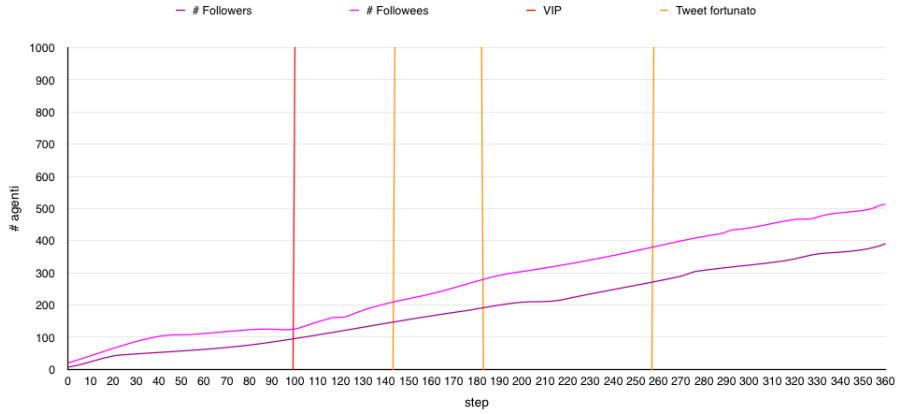


Figura 23: Follower e followee di un nodo diventato vip

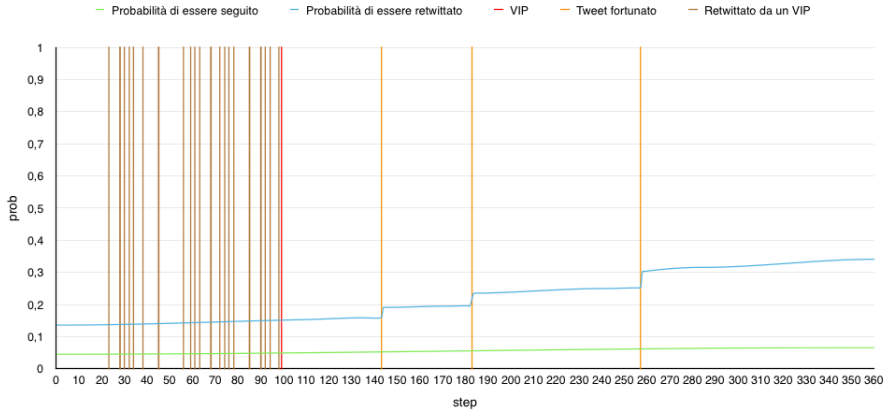


Figura 24: Probabilità di essere seguito e di essere retwittato di un nodo nato non vip

con la condizione AND si vedono raramente dei retweet, in questo se ne vedono diversi. Questo non è frutto di errori, ma semplicemente è una conseguenza del fatto che sono rappresentati solo i retweet dei nodi vip ai nodi non vip. Avendo quindi un numero di agenti vip nettamente superiore in questo modello, ci si trova in una situazione in cui tanti nodi effettuano dei retweet a pochi nodi.

6.2.1 Modifiche per simulazione successiva

Nella terza simulazione ci è sembrato interessante valutare l'impatto dell'omofilia di gruppo sul comportamento della rete; abbiamo infatti rilevato nelle simulazioni precedenti e soprattutto per il "ramo OR", che le curve rappresentanti il numero di follwer e di followee non subivano particolari cambiamenti dopo

il passaggio di stato di un nodo da non vip a vip. Abbiamo imputato questa assenza di cambiamento alla presenza dell'omofilia di gruppo che favoriva le relazioni tra nodi dello stesso tipo. Nei primi step "di assestamento", avendo un gran numero di nodi non vip, vi erano molte probabilità per un nodo non vip di stabilire relazioni con nodi "simili"; dopo gli step di assestamento, si registrava un' "esplosione" nel numero di nodi vip che annullava l'effetto dell'omofilia di gruppo per un nodo vip eguagliandola sostanzialmente a quella per i nodi non vip nello step precedente.

6.3 Simulazione 3

Eliminando l'omofilia di gruppo ci aspettiamo due principali risultati:

1. nessuna variazione nell'andamento delle curve di follower/followee dei nodi non vip al loro cambiamento di stato (questa volta l'assenza di variazione è motivata dal fatto che tutti i nodi hanno la stessa probabilità di stabilire relazioni con tutti gli altri)

2. una variazione nell'andamento della curva dei followee di un nodo vip nel modello AND (che fino ad ora è sempre rimasta costante, crediamo, per via dell'omofilia di gruppo che scoraggiava, per la natura della formula utilizzata, la relazione di following di un vip nei confronti di un non vip in misura maggiore di quanto avvenisse nella direzione opposta)

Caso AND

Un primo evidente cambiamento per questa simulazione che non avevamo considerato, ma che, a posteriori, ci sembra giustificato, è l'ulteriore distacco tra il numero di follower dei nodi vip e quello dei nodi non vip visibile in Figura 25 osservando le differenti dimensioni dei nodi verdi e di quelli rosa. Rispetto alle simulazioni precedenti sono ancora meno i nodi che hanno una dimensione prossima a quelli vip. Eliminando l'omofilia di gruppo abbiamo diminuito la probabilità dei nodi non vip di creare relazioni tra di loro: la funzione utilizzata, infatti, premia molto il gruppo "dominante" (e penalizza il gruppo "minoritario") e la sua eliminazione è dunque penalizzante per il gruppo di maggioranza rispetto al suo inserimento. Questo comporta una diminuzione nella creazione del numero di archi tra nodi non vip ed una conseguente riduzione del numero di follower dei nodi in questione. Un'altra considerazione che possiamo fare è che anche l'eliminazione dell'omofilia di gruppo non ha influito sul passaggio di stato dei nodi da non vip a vip; al termine dell'esecuzione abbiamo infatti che il rapporto nodi vip/nodi non vip è rimasto invariato. Analizziamo ora i soliti grafici alla ricerca di cambiamenti degni di nota e, in particolare, di conferme o smentite alle nostre aspettative

Come emerge dai grafici, non notiamo cambiamenti significativi nell'andamento delle curve di probabilità di essere seguito o di essere retwittato per il nodo vip che riportiamo comunque per maggiore completezza. Il grafico interessante è invece quello dei follower e dei followee del nodo; se la curva dei follower

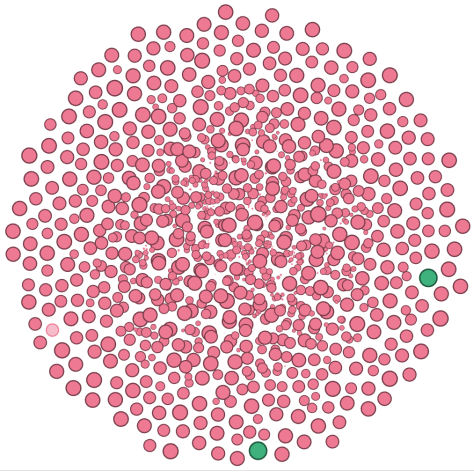


Figura 25: Stato della rete al termine della simulazione 3

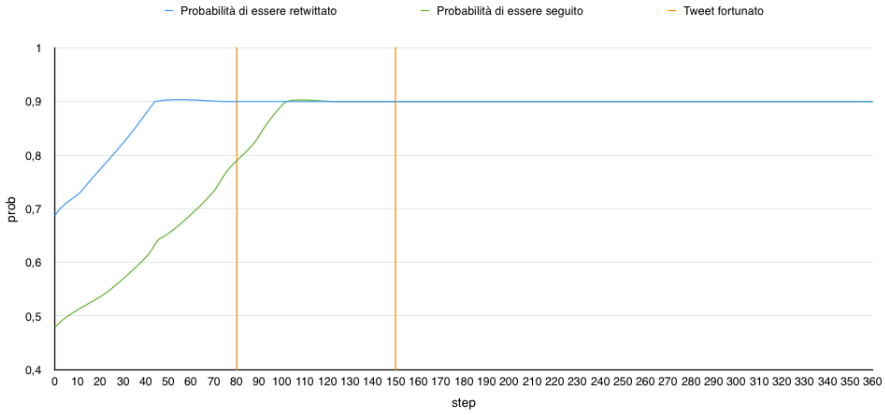


Figura 26: Probabilità di un nodo vip

non ha subito cambiamenti sostanziali, la curva dei followee si è finalmente scostata da quell'andamento costante che aveva avuto fino ad ora, mostrando una crescita lineare. L'aspettativa numero due non è quindi stata disattesa. Passiamo adesso all'analisi dei grafici dei nodi non vip per constatare la presenza di eventuali cambiamenti: saranno mostrati i grafici di un nodo non vip che, al termine della simulazione, aveva un numero di follower molto basso e di uno che, nello stesso istante, aveva un numero di follower quasi prossimo a quello del nodo vip.

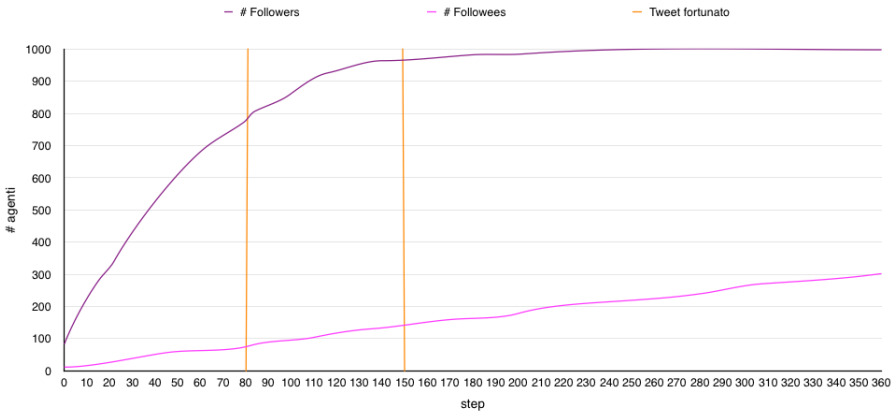


Figura 27: Follower e followee di un nodo vip

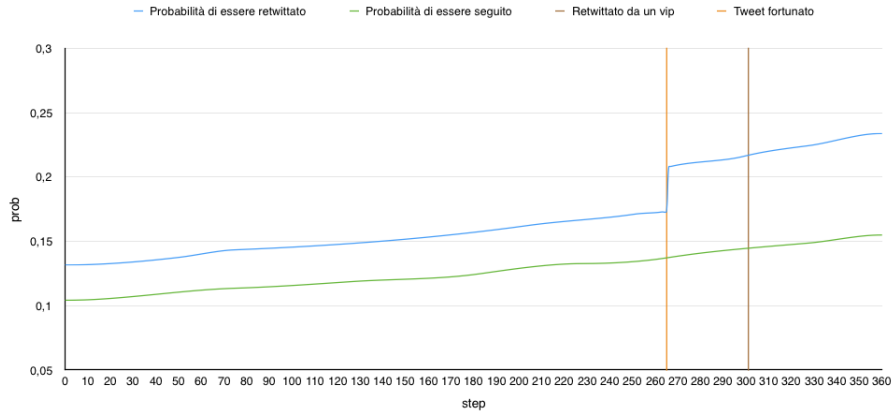


Figura 28: Probabilità di un nodo non vip con molti follower

Per quanto riguarda i nodi non vip, sia nel caso in cui questi abbiano, al termine della simulazione, molti follower sia nel caso in cui ne abbiano pochi, non riscontriamo evidenti differenze rispetto ai casi precedenti.

Caso OR

Per questo modello, l'aver rimosso l'omofilia di gruppo non ha portato cambiamenti significativi per quanto riguarda il numero di vip e non vip, infatti, la simulazione termina, ancora una volta, con 1000 utenti vip e l'andamento delle curve di questo cambiamento, ovvero il passaggio da non vip a vip, avviene in maniera molto simile a quello della simulazione precedente, mostrato in Figura 21.

Anche per quanto riguarda i nodi che sono diventati vip, il loro comportamento è molto simile a quello analizzato in precedenza. Una possibile motivazione per

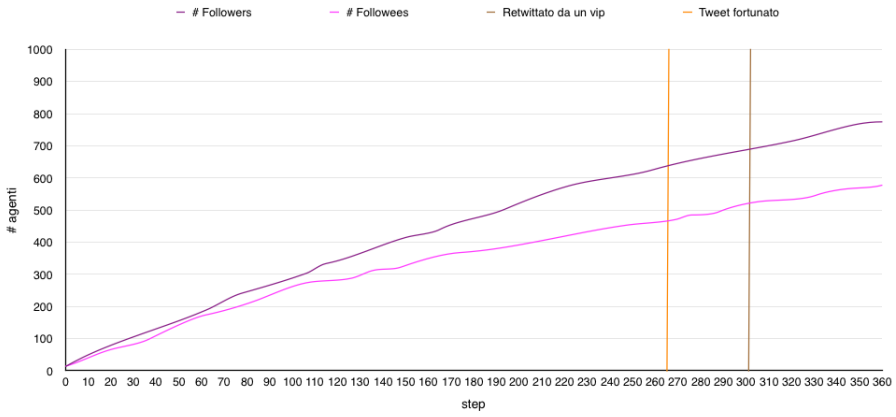


Figura 29: Follower e followee di un nodo non vip con molti follower

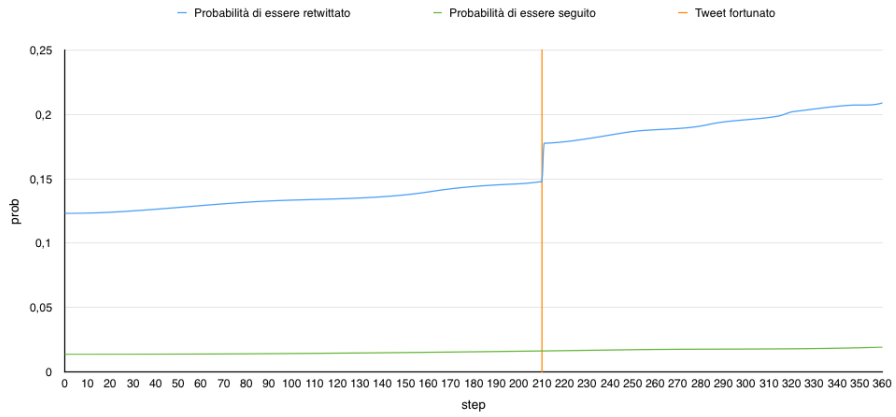


Figura 30: Probabilità di un nodo non vip con pochi follower

cui questo accade può essere che in questo modello molti agenti che nascono non vip si trasformano in vip quasi subito dopo i 20 step di assestamento.

6.3.1 Modifiche per simulazione successiva

Per l'ultima simulazione abbiamo deciso di modificare alcuni parametri che ci permettessero di analizzare aspetti ancora non considerati. I parametri modificati sono i seguenti: è stata reintrodotta l'omofilia di gruppo, è stato diminuito il "premio" derivante dal tweet fortunato, abbiamo ristretto maggiormente il filtro che permette la creazione di archi tra due nodi sia all'inizializzazione della

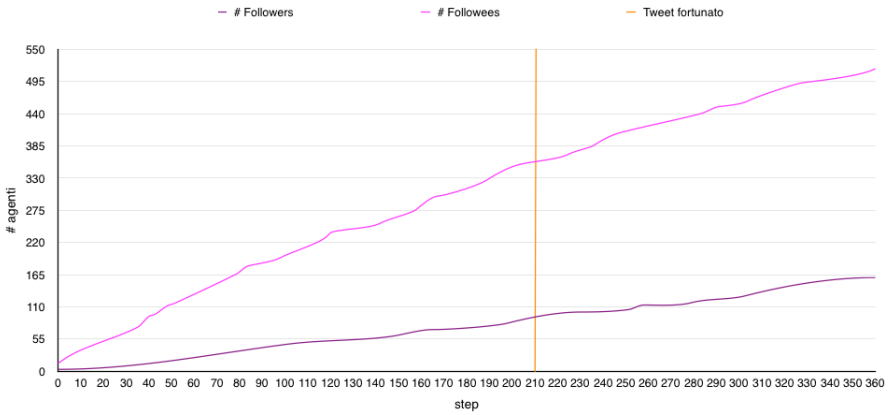


Figura 31: Probabilità di un nodo non vip con pochi follower

rete, sia durante l'esecuzione della simulazione, abbiamo aumentato i valori minimi e massimi dell'intervallo nel quale vengono selezionate le probabilità di un vip di seguire un altro utente ed abbiamo diminuito leggermente la probabilità minima di essere retwittato e di essere seguito di un vip.

6.4 Simulazione 4

In seguito alle modifiche introdotte per quest'ultima simulazione, ci aspettiamo che:

1. avendo reintrodotto l'omofilia di gruppo, il numero di followee di un nodo vip torni ad essere costante (nel modello AND)
2. il numero di archi della rete (numero di follower generali) diminuisca rispetto alla simulazione precedente sia all'inizializzazione (nella quale erano circa 13 000), sia al termine della simulazione (circa 600 000)
3. la crescita della curva del numero di follower di un nodo vip sia più lenta

Caso AND

Al termine della simulazione lo stato della rete è quello rappresentato dalla Figura 32. Notiamo che, a differenza delle altre simulaizoni, questa volta abbiamo un nuovo nodo vip. Confrontando l'immagine con quella di Figura 25 rappresentante lo stato della rete al termine della simulazione precedente, possiamo inoltre notare che, nella simulazione attuale, la dimensione dei nodi (e quindi il loro numero di follower) è ulteriormente ridotta. Questo dato si rispecchia anche nel numero totale di archi che si hanno al termine della simulazione: come ci aspettavamo, in entrambi i modelli questo numero risulta nettamente inferiore. Abbiamo infatti circa 6 500 collegamenti dopo l'inizializzazione della rete e circa 350 000 al termine della simulazione. Analizziamo ora i soliti "no-

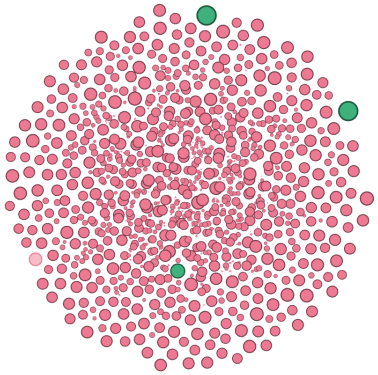


Figura 32: Stato della rete al termine della simulazione 4

di interessanti/rappresentativi". Non riportiamo il grafico dell'andamento delle probabilità del nodo vip in quanto non presenta particolari cambiamenti nell'andamento rispetto ai grafici precedenti; più interessante può essere invece l'analisi del graifico dei followers e dei followee del nodo. Dall'andamento (nuovamente) costante della curva dei followee del nodo, risulta palese la reintroduzione dell'omofilia di gruppo sulla quale non ci dilungheremo ulteriormente; dobbiamo riscontrare che l'aumento (seppur molto leggero) dell'intervallo di probabilità di seguire dei nodi vip non è servito ad aumentare i followee. Per quanto riguarda la curva di crescita dei follower si nota una leggera "distensione" rispetto al grafico precedente che indica una minore rapidità di crescita del numero dei follower del nodo dovuta sia al reinserimento dell'omofilia di gruppo, sia all'abbassamento della probabilità minima di essere seguito di un vip.

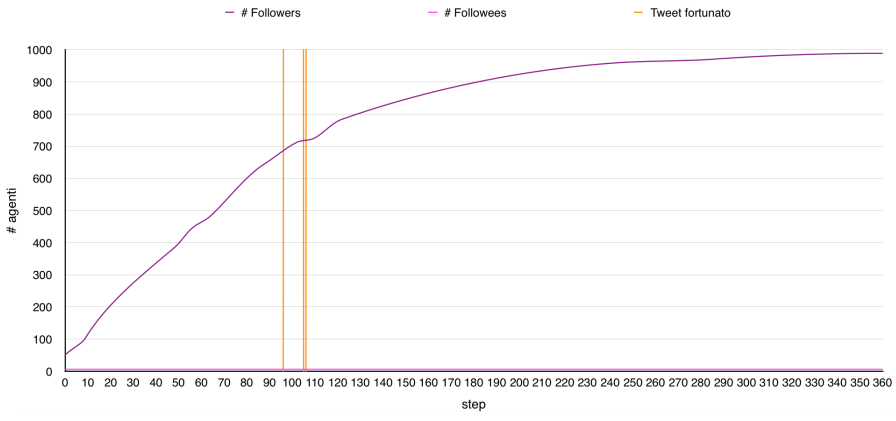


Figura 33: Follower e Followee di un nodo vip

Un nodo che non potevamo trascurare è senz'altro quello diventato vip.

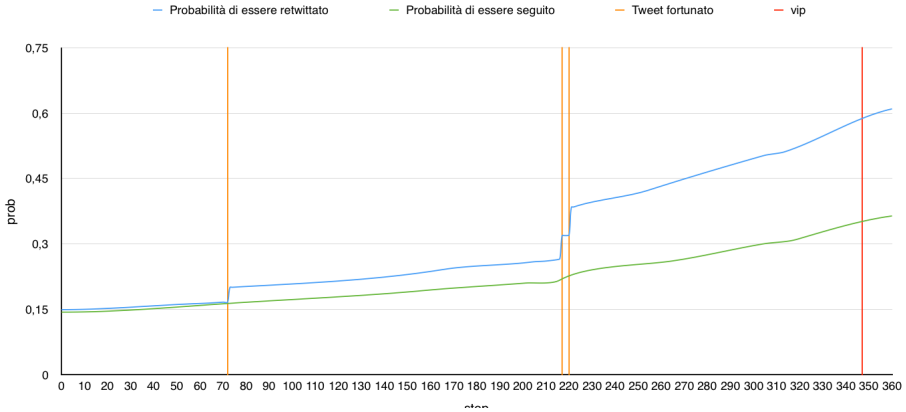


Figura 34: Probabilità di un nodo diventato nodo vip

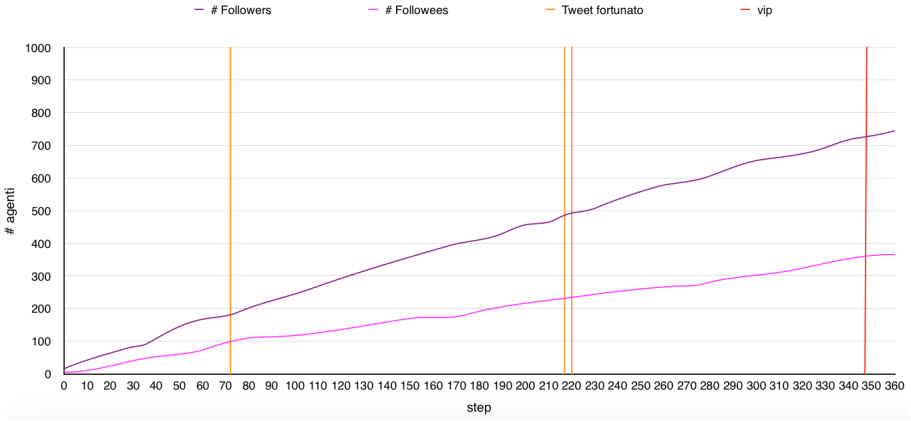


Figura 35: Follower e Followee di un nodo diventato vip

Dai grafici notiamo che il nodo è diventato vip negli ultimi step della simulazione, pertanto risulterà difficile fare delle considerazioni rilevanti sul suo comportamento in seguito al passaggio di stato. Possiamo tuttavia notare che anche dopo essere diventato vip, il nodo non modifica il proprio comportamento sebbene sia riscontrabile una leggera attenuazione nella crescita del numero di followee dovuta probabilimente all'omofilia di gruppo. Follower e followee crescono secondo un andamento lineare (fatta eccezione per la leggera flessione appena menzionata che meriterebbe di essere approfondita tramite una simulazione con più step che, purtroppo, non abbiamo la potenza di calcolo necessara per affrontare). Per quanto riguarda l'andamento delle curve di probabilità, non sembra vi sia niente di particolare da riscontrare: è giusto far notare come gli incrementi nella

probabilità di essere retwittato non siano equivalenti in corrispondenza dei tre tweet fortunati (sono infatti crescenti); la causa di tale differenza risiede nel fatto che gli incrementi sono fatti in maniera percentuale rispetto alla probabilità di essere retwittato. Vediamo infine i grafici di un nodo che, al termine della simulazione, aveva un numero di follower prossimo a quelli di un nodo vip e quelli di un nodo che ne aveva un numero esiguo.

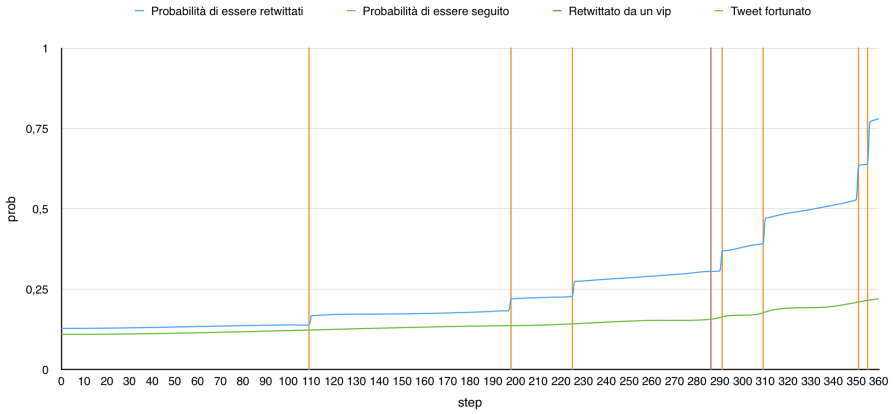


Figura 36: Probabilità nodo non vip con molti follower

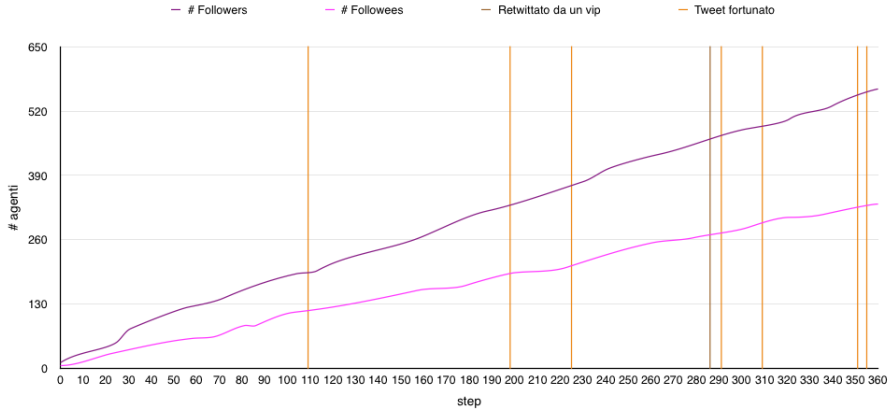


Figura 37: Follower e Followee di un nodo con molti follower

Andando in ordine, la Figura 36 e la Figura 37 mostrano i grafici del nodo con molti follower: il nodo analizzato ha incrementato notevolmente la propria pro-

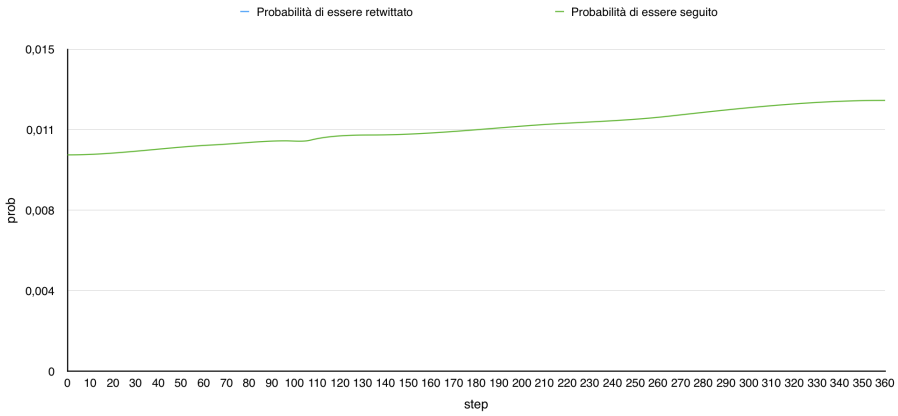


Figura 38: Probabilità di un nodo con pochi follower

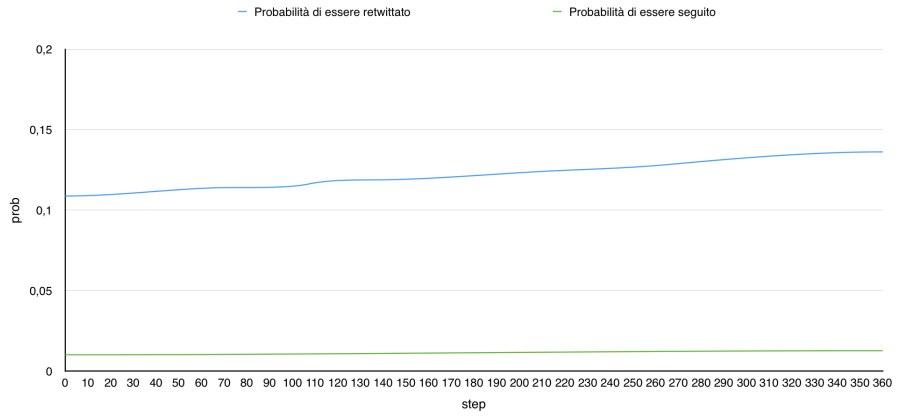


Figura 39: Probabilità di un nodo con pochi follower

babilità di essere retwittato in seguito ad una serie di tweet fortunati. Il nodo è stato inoltre retwittato da un vip, ma non si riscontrano cambiamenti evidenti. Follower e followee crescono in maniera quasi lineare; da notare il maggior tasso di crescita del numero di follower rispetto al numero di followee. Gli ultimi tre grafici rappresentano le probabilità e le interazioni del nodo con pochi follower al termine della simulazione. La motivazione della sua "sfortuna" durante l'esecuzione del modello è da identificarsi nelle basse probabilità con le quali è stato inizializzato (entrambe sotto lo 0.15) e nell'assenza di eventi rilevanti come tweet fortunati o retweet da parte di vip. I followee del nodo crescono linearmente mentre i follower crescono leggermente solo negli ultimi step della simulazione in corrispondenza di un timido incremento della probabilità di essere retwittato non riuscendo comunque a superare la cinquantina (Figura 40).

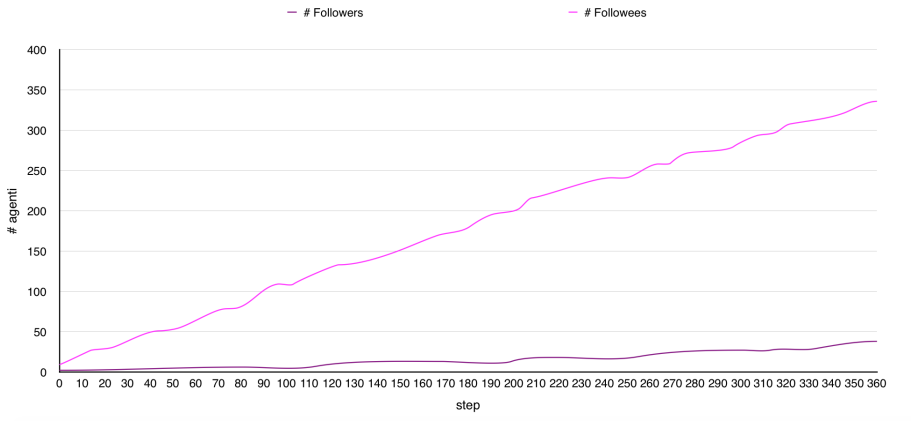


Figura 40: Follower e Followee di un nodo con pochi follower

Caso OR

Con le modifiche introdotte dopo la terza simulazione, si è verificato un netto miglioramento nel numero di vip al termine dell'esecuzione (Figura 41): ci aspettavamo questo genere di comportamento soprattutto perchè siamo andati a modificare la probabilità di creazione degli archi.

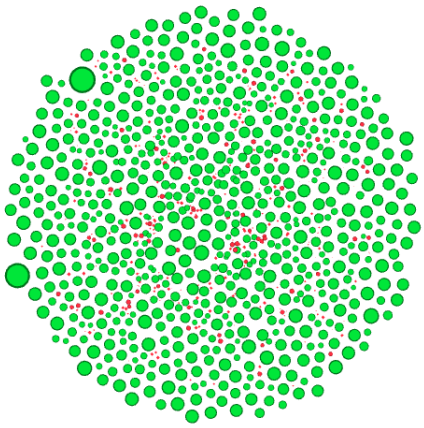


Figura 41: Stato della rete al termine della simulazione 4

Come mostrato dalla Figura 42, il numero di utenti vip inizia ad aumentare intorno allo step numero 80.
Siamo, quindi, riusciti a spostare lievemente il *punto critico*[11] più avanti, ma non riteniamo ancora questo risultato soddisfacente.

[11] In [5] viene definito come 'Quel livello oltre il quale un cambiamento diviene inarrestabile'

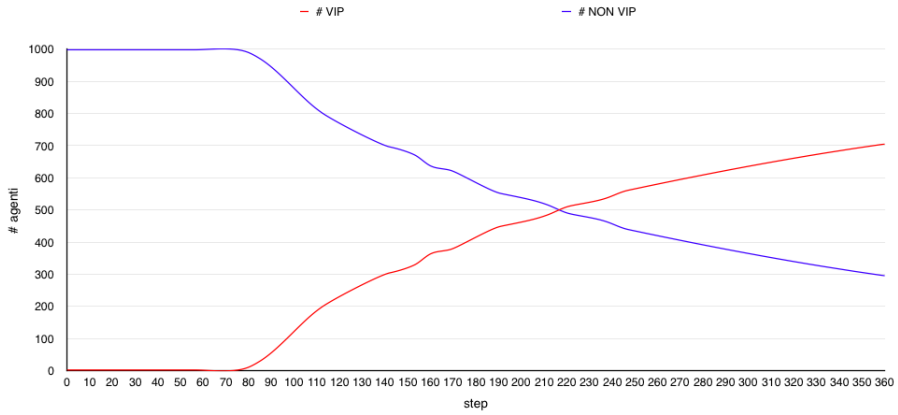
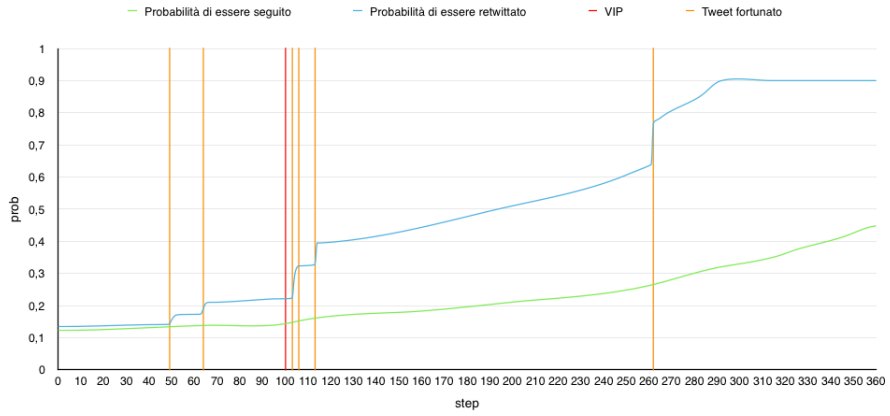


Figura 42: Andamento numero di nodi vip e non vip

L'aumento in percentuale della probabilità di essere seguito non è così evidente quanto quello della probabilità di essere retwittato, in quanto tale incremento della prima non è tanto rilevante quanto quello della seconda; è comunque possibile vedere una leggera variazione della crescita della curva in corrispondenza dei tweet fortunati (Figura 43).



g

Figura 43: Probabilità di essere seguito e di essere retwittato di un nodo diventato vip

Concentrandoci, invece, su un nodo che non è diventato vip, ma che presumibilmente lo sarebbe diventato in pochi passi, nella Figura 44 si può notare come i retweet da parte dei vip abbiamo influito in maniera molto simile sulle probabilità di essere seguito e di essere retwittato; in assenza di tweet fortunati, in questa simulazione, le probabilità di essere seguito e di essere retwittato seguono lo stesso andamento.

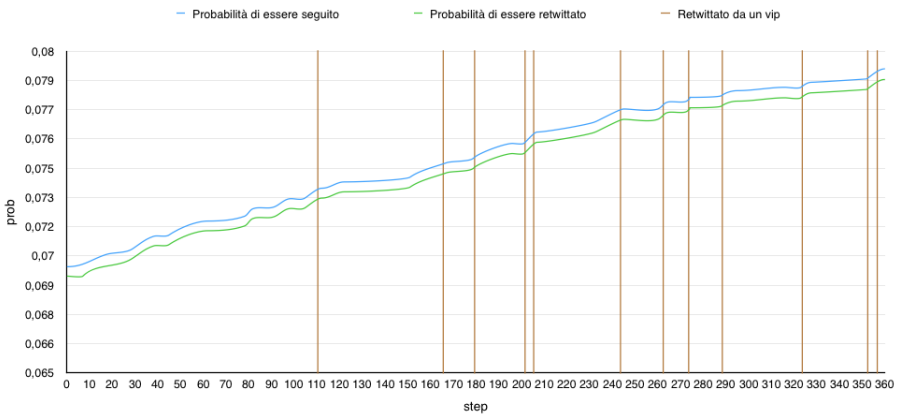


Figura 44: Probabilità di essere seguito e di essere retwittato di un nodo che è rimasto non vip

7 Problemi riscontrati

Durante lo sviluppo del progetto ci siamo imbattuti in una serie di problemi che saranno analizzati in questa sezione.

7.1 Problema 1: Clock e saturazione della rete

Uno dei problemi riscontrati nelle prime simulazioni, nelle quali ogni clock equivaleva a 24 ore, e per cui non vi era differenziazione tra le varie fasi di una giornata, è stato quello del raggiungimento di un livello di saturazione molto elevato nell'arco di pochi step. Questo problema ci ha costretti a riconsiderare il concetto di tempo nel nostro modello.

Abbiamo quindi deciso di adottare una suddivisione in quattro fasce orarie sulla base del ragionamento che ogni utente non ha lo stesso modo di interagire con un social network in qualsiasi ora del giorno e pertanto, un modello in grado di differenziare l'interazione dell'utente sulla base del periodo del giorno, sarebbe stato senz'altro più aderente alla realtà.

7.2 Problema 2: Lentezza dell'interfaccia grafica dinamica

Un secondo problema che abbiamo dovuto affrontare riguarda la creazione dell'interfaccia grafica in maniera dinamica: la libreria che abbiamo utilizzato possiede un metodo che permette di creare in automatico la GUI contenente la rete con i relativi nodi ed archi.
Questa interfaccia si è rivelata essere molto comoda nelle simulazioni iniziali, in cui creavamo una rete con un basso numero di nodi ed il cui scopo era quello di verificare la corretta esecuzione delle funzionalità.

Giunti al momento di effettuare le simulazioni con un numero di nodi superiore a quello di testing, ci siamo resi conto che l'interfaccia rallentava di molto la computazione, in quanto doveva disegnare ogni cambiamento della rete ad ogni clock eseguito. Inoltre, l'interfaccia fornita dalla libreria utilizzata non è *usabile* per reti di grandi dimensioni in quanto i nodi non possono essere spostati; per questi motivi abbiamo deciso di non mostrare l'interfaccia grafica della rete, ma di salvare i file in un formato compatibile con la piattaforma Gephi, precedentemente citata.

Analogamente al problema della GUI per la rete, anche la creazione dei grafici analitici, utili per analizzare dati come, ad esempio, il monitoraggio del numero di followers di un nodo o l'andamento di alcune sue probabilità, era molto lenta.
La parte analitica è stata perciò svolta salvando i dati della rete in formato JSON, convertendoli poi in CSV utilizzati per creare i grafici necessari a trarre le dovute conclusioni in merito ai dati forniti dalla simulazione.

7.3 Problema 3: Tempi di esecuzione degli algoritmi su grafi

Un terzo problema con cui abbiamo dovuto scontrarci è stato quello legato al tempo di esecuzione dell'algoritmo di confronto tra i nodi del grafo.

In una versione meno efficiente di alcune funzioni del nostro codice il numero di confronti di tutti i nodi con tutti gli altri era $O(n^2)$.
Considerata la necessità di effettuare questo confronto tra un gran numero di nodi più volte nel corso dell'esecuzione del modello e considerata la potenza di calcolo dei nostri elaboratori, l'esecuzione di un numero accettabile di step (intorno a 300), rischiava di diventare eccessivamente onerosa.
Per ovviare a questo problema abbiamo provato a ridurre, laddove possibile, il numero di confronti.
Ci siamo infatti resi conto che, in determinate funzioni, lo svolgimento di alcune operazioni tra due nodi i e j poteva essere computato una volta sola ed applicato in entrambe le direzioni (da i a j e da j a i) previa strutturazione del codice in maniera opportuna. Prendiamo come esempio la funzione di aggiornamento degli archi: questa dovrebbe calcolare, per ogni nodo, la sua omofilia con tutti gli altri.
Dati i nodi i e j, la versione ingenua di questo algoritmo clacolerebbe l'omofilia "da i a j" e, successivamente, quella da "j a i". Studiamo la formula dell'omofilia per interesse[12]: è evidente che, nel calcolo dell'omofilia, dobbiamo prendere i valori del nodo i e del nodo j due volte, una per direzione.
Ragioniamo ora su ciò che accade quando dobbiamo confrontare una lista di nodi con se stessa: è possibile evitare di ripetere due volte ogni confronto se paragoniamo la lista completa solo con la porzione della stessa con cui non è ancora stato eseguito.
Per comprendere meglio questa soluzione, supponiamo di avere una lista di sei elementi e di voler confrontare ognuno con tutti gli altri escluso se stesso (che è in sostanza quello che vogliamo fare con il nostro algoritmo). Alcuni confronti possono essere risparmiati perché già eseguiti precedentemente; in particolare possiamo eseguire il confronto dell' i-esimo elemento con i soli elementi della lista dall'i+1-esimo in poi. In questo modo la complessità dei confronti viene ridotta da $O(n^2)$ ad $O(n)$.
Per riuscire ad effettuare questa riduzione abbiamo dovuto implementare delle apposite funzioni ausiliarie che, presi i nodi i e j, calcolassero il valore richiesto in entrambe le direzioni.

Moltiplicando questo risparmio per il numero di funzioni che necessitano il confronto fra nodi e che è possibile semplificare attraverso l'uso di funzioni ausiliarie, abbiamo ottenuto un notevole risparmio in termini di computazione.
Occorre tuttavia sottolineare che, nonostante le semplificazioni descritte, l'esecuzione della simulazione risulta comunque irrimediabilmente onerosa per un computer normale.

[12] Vedere sezione "Riferimenti matematici"

8 Sviluppi Futuri

In seguito a quanto emerso dall'analisi dei risultati, abbiamo identificato alcuni sviluppi che potranno essere utili per rendere il modello ancora più fedele alle dinamiche reali.
Essendo questo uno dei primi modelli sviluppati in questo campo, vi è un largo margine di miglioramento e quindi un discreto numero di sviluppi futuri che è possibile prendere in considerazione. Ne riportiamo di seguito alcuni.

8.1 Sviluppo 1: Caratterizzazione di agente

Un primo sviluppo che si può pensare di implementare è sicuramente quello dell'aggiunta di parametri che caratterizzino un agente in maniera più specifica. Si potrebbero inserire attributi quali etnia, età, religione e genere e considerare anche questi nel calcolo dell'omofilia. In questo modo si otterrebbe un'ulteriore differenziazione tra le omofilie relative a ciascun agente. Un ulteriore sviluppo, emerso da [6], ma che non abbiamo implementato per motivi di tempo è la distinzione tra agenti che l'articolo citato distingue tra *Informers* e *Meformers*: i primi sono quegli utenti che pubblicano contenuti prevalentemente a scopo informativo (es: news, notizie, articoli), i secondi, invece, tendono a twittare informazioni riguardanti il proprio stato personale. Si potrebbe, infine, caratterizzare un agente introducento il concetto di *Influencer* inteso come persona non vip, ma con molti follower.

8.2 Sviluppo 2: Differenziazione del numero e del tipo interessi

In [6] emerge un'idea interessante che abbiamo deciso di considerare come un possibile sviluppo futuro del nostro modello: gli utenti che hanno un maggior numero di interessi hanno maggiore probabilità di essere seguiti rispetto a quelli che ne hanno meno. Lo sviluppo che potrebbe essere interessante implementare è quello della differenziazione del tipo di interessi (che al momento è uguale per tutti gli attori) e del numero di questi ultimi così da poter avere una rete composta da agenti che hanno un numero di interessi differente e che non sono necessariamente gli stessi.

8.3 Sviluppo 3: Caratterizzazione di tweet

Una terza modifica potrebbe essere quella di caratterizzare ogni tweet con uno o più argomenti (analoghi agli interessi di ciascun agente), così da poter migliorare la funzione di retweet aumentando la probabilità di retweet nel caso di corrispondenza tra gli interessi. Per avvicinarsi ulteriormente al risultato reale, si potrebbe pensare di inserire anche un valore "qualità del tweet" che differenzi i tweet autorevoli da quelli che lo sono meno, influenzandone la probabilità di essere retwittati. Unire questa caratteristica con quella descritta nello Svilup-

po 1 (*Informers, Meformers*) potrebbe portare a dei risultati molto simili alla dinamica reale.

8.4 Sviluppo 4: Dinamicità della rete

Uno sviluppo sicuramente rilevante è quello di permettere la creazione o l'eliminazione di nodi della rete: attualmente, infatti, la rete è statica; il numero di nodi con cui è inizializzata non varia durante la simulazione. In un caso reale, invece, è possibile che gli utenti creino od eliminino i loro account Twitter. Analogamente a quanto detto sopra, per raggiungere un maggiore livello di realismo, potrebbe essere sensato implementare un *meccanismo di unfollowing* che consenta agli agenti di decidere di non seguire più degli utenti precedentemente seguiti.

8.5 Sviluppo 5: Inserimento utenti inattivi

Consultando le informazioni fornite da Twitter sulla sua pagina web[13] ed effettuando ulteriori ricerche in rete[14] è possibile avere una stima degli utenti Twitter attivi (circa 313M nel 2016) e del totale degli utenti registrati al social network (circa 1.3B nel 2015). Da questi dati nasce l'idea per un ulteriore sviluppo futuro: è possbile infatti implementare un modello che tenga in considerazione la presenza di attori "inattivi", intesi come nodi che restano iscritti al social network senza però utilizzarlo.

8.6 Sviluppo 6: Assegnazione di un tempo ai tweet

Attualmente ogni tweet è considerato solo in funzione della sua presenza: con questo intendiamo dire che è sufficiente che un agente abbia eseguito un tweet per far in modo che possa essere retwittato in qualsiasi step dell'esecuzione del modello. Quest'ultima considerazione deriva dal fatto che, in Twitter, è possibile che un tweet venga retwittato anche a distanza di tempo. Si può pensare invece di associare ad ogni tweet un tempo così da poter differenziare la probabilità di retwittare un tweet molto vecchio (che dovrebbe essere inferiore) da quella di uno appena fatto (che dovrebbe essere maggiore).

8.7 Osservazioni

Un'osservazione che è doveroso esplicitare è la seguente: attualmente i nodi che diventano vip non modificano la loro probabilità di retweettare un altro nodo. Abbiamo fatto questa scelta implementativa pensando che non fosse realistico modificare la probabilità di un nodo di retwittarne un altro in seguito al suo "passaggio di stato". Ci è sembrato, infatti, che tale modifica fosse analoga all'asserire che un agente modifica il proprio atteggiamento nei confronti di un social network in virtù del fatto che ha un numero di follower superiore alla

[13] `https://about.twitter.com/company`
[14] `http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/`

media.
Potrebbe essere interessante far variare dinamicamente questo parametro e studiarne l'effetto sul comportamento generale della rete.

9 Conclusioni

In seguito alla documentazione letta, allo sviluppo dei modelli e alle simulazioni effettuate, possiamo concludere che i modelli implementati presentano comportamenti differenti, ma comunque in linea con le nostre aspettative.
Riteniamo opportuno fare una distinzione tra il modello AND e il modello OR relativamente al concetto di vip che si decide utilizzare; i nostri modelli sono stati implementati utilizzando la definizione di vip definita nel glossario e pertanto il modello AND sembra essere il più coerente con tale definizione. Uniformando, invece, il concetto di vip a quello di influencer (persona molto seguita), il modello OR avrebbe un'aderenza maggiore al caso reale.
In seguito a quanto detto, se dovesse essere implementata la distinzione tra utente vip e *influencer* nel modello AND, il modello OR non avrebbe più motivo di esistere.
Una problematica comune ad entrambi i modelli è il raggiungimento della saturazione della rete, infatti, se le simulazioni durassero un tempo molto più lungo, le reti rappresentate da entrambi i modelli avrebbero tutti i nodi etichettati come vip. Questa è una conseguenza diretta della staticità della rete, dell'assenza di un meccanismo di unfollowing e del fatto che tutti gli agenti sono attivi (utilizzano frequentemente Twitter).
Sebbene abbiamo assunto che un agente non modificasse il proprio comportamento nei confronti del social network e quindi non modificasse, ad esempio, la propria probabilità di retwittare, questo comportamento porta all'aumento esponenziale del numero di vip all'interno della rete; questo fenomeno risulta molto evidente nel modello OR, ma, a lungo andare, sarebbe visibile anche nel modello AND. Questa considerazione ci ha portato a rivalutare la nostra assunzione iniziale come descritto nella sezione *Osservazioni* del capitolo *Sviluppi futuri*. Concludiamo con una considerazione riguardante l'omofilia: abbiamo introdotto due concetti fondamentali quali l'omofilia per interesse e l'omofilia di gruppo ed entrambi contribuiscono in positivo al funzionamento realistico dei modelli. L'omofilia di gruppo ha delle realizzazioni diverse nei modelli presentati: se nel modello AND ha l'effetto di lasciare invariato il numero di followee di un nodo vip, nel modello OR ciò non avviene. Infatti, se inizialmente l'omofilia di gruppo rallenta la rapidità con cui un nodo vip crea collegamenti con i nodi di tipo differente, a lungo andare, con l'incremento degli agenti vip, il suo valore tende al valore uno rendendo il suo effetto ininfluente.

Riferimenti bibliografici

[1] Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[2] A.-L. Barabasi. *La scienza delle reti*. Einaudi, 2004.

[3] A.-L. Barabasi, A. Reka, and J. Hawoong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 1(1):173–187, 1999.

[4] S. Currarini, M. O. Jackson, and P. Pin. An economic model of friendship: Homophily, minorities and segregation. *Econometrica*, 2007.

[5] M. Gladwell. *Il punto critico - I grandi effetti dei piccoli cambiamenti*. BUR, 2006.

[6] C. Hutto, S. Yardi, and E. Gilbert. A longitudinal study of follow predictors on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 821–830, New York, NY, USA, 2013. ACM.

[7] P. J. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1996.