Università degli Studi di
Milano-Bicocca

Sistemi Complessi: Modelli e Simulazione

Fake News Countermeasures

Raffaele Cerizza - 845512

Link al repository:
`https://github.com/raffaelecerizza/fake_news_countermeasures`

Indice

1 Introduzione

La disinformazione è un fenomeno risalente. Alcuni esempi celebri risalgono addirittura al I secolo a.C. [30][1]. Più recentemente l'avvento dei mezzi di informazione digitali ha agevolato la rapidità di diffusione della disinformazione [26]. E questo ha portato il World Economic Forum classificare la disinformazione come una delle più grandi minacce per le società umane e la democrazia [38]. Per questo motivo i ricercatori hanno analizzato questo fenomeno e ne hanno studiato le contromisure [12]. In questo quadro è stata coniata una nuova espressione per definire le false informazioni: *fake news*.
Tra i mezzi di informazione digitali più diffusi vi sono certamente i *social media*. In particolare i *social media* sono mezzi di comunicazione caratterizzati da una rete sociale (*social network*) di utenti che accedono ai servizi offerti tramite strumenti digitali [16]. I *social media* consentono agli utenti sia la fruizione che la diffusione di informazioni [26]. L'accessibilità a basso costo dei *social media* e la rapidità con cui i messaggi possono essere scambiati facilitano la diffusione delle informazioni. E offrono un terreno fertile per la disinformazione.
Un *social network* è anche un sistema complesso. Infatti prevede l'interazione di molti agenti autonomi ed esibisce comportamenti emergenti [27]. Tra questi vi è certamente la formazione di gruppi di utenti che rafforzano la polarizzazione delle opinioni. Questo fenomeno prende il nome di *echo chamber* [20]. E contribuisce anch'esso a rafforzare la diffusione delle *fake news*.
In questo lavoro verrà analizzato il fenomeno della diffusione di *fake news* in *social network* con l'obiettivo di individuarne possibili contromisure. Questa analisi verrà condotta studiando il fenomeno come sistema complesso di agenti interagenti.
In particolare l'esposizione seguirà questo ordine:

- nella Sezione 2 verrà illustrato lo stato dell'arte in tema di modellazione di reti sociali e verrà descritto il modello proposto per questo lavoro;
- nella Sezione 3 verrà illustrato lo stato dell'arte in tema di modellazione della diffusione di *fake news* e verrà descritto il modello utilizzato per questo lavoro;

[1] In particolare ci si riferisce alla campagna di disinformazione con cui Ottaviano dipinse il rivale Marco Antonio come burattino di Cleopatra.

- nella Sezione 4 verranno inquadrati i modelli proposti all'interno dei tipici modelli basati su agenti;
- nella Sezione 5 verranno descritte e analizzate le operazioni di calibrazione, di validazione e di analisi di sensitività dei parametri dei modelli proposti;
- nella Sezione 6 verranno descritte e analizzate alcune simulazioni svolte utilizzando i modelli proposti al fine di individuare contromisure efficaci alla diffusione di *fake news*;
- infine verranno rassegnate le conclusioni del presente lavoro e proposti alcuni possibili sviluppi futuri.

2 Modello di rete sociale

Per simulare la diffusione di *fake news* in una rete sociale occorre anzitutto modellare quest'ultima. In letteratura esistono diversi modelli di reti sociali. In questa Sezione verranno anzitutto descritti alcuni modelli rappresentativi dello stato dell'arte. Dopodiché verrà definito il modello di rete sociale utilizzato per il presente lavoro.

2.1 Definizioni

In questa Sezione si farà largo uso di termini appartenenti alla teoria dei grafi. Per agevolare la comprensione dei Paragrafi successivi occorre dunque fornire alcune brevi definizioni di questi termini. In particolare:

- **Grafo**: un grafo è definito come una struttura composta da nodi e archi. Formalmente un grafo è definito come $G = (V, E)$, dove V è l'insieme dei nodi ed E è l'insieme degli archi.
- **Nodo**: un nodo è un'entità del grafo che può presentare attributi e può essere in relazione con altri nodi.
- **Arco**: un arco rappresenta una connessione tra nodi. Anche gli archi possono presentare attributi.
- **Multi-archi**: si parla di multi-archi quando più archi insistono sulla stessa coppia di nodi con la stessa direzionalità.
- **Cappio**: un cappio è un arco che parte da un nodo e termina sullo stesso nodo.
- **Grafo orientato**: un grafo orientato è un grafo in cui gli archi hanno una direzione. La direzione dell'arco consente di determinare l'origine e la destinazione della relazione.
- **Grafo non orientato**: un grafo non orientato è un grafo in cui gli archi non hanno una direzione. In questo caso la presenza di un arco tra due nodi determina l'esistenza di una relazione bidirezionale. Pertanto entrambi i nodi dell'arco sono allo stesso tempo sorgenti e destinazioni della relazione.

- **Grado** o ***degree***: il grado di un nodo specifica il numero di archi che incidono su di esso.
- ***In-degree***: l'*in-degree* di un nodo rappresenta il numero di archi entranti in esso.
- ***Out-degree***: l'*out-degree* di un nodo rappresenta il numero di archi uscenti da esso.

In questo lavoro si userà spesso anche il termine "rete" come sinonimo di "grafo". L'uso di questi termini come sinonimi è frequente nella letteratura scientifica [4].

2.2 Stato dell'arte

2.2.1 Modello di Barabási-Albert

Il primo modello di rete sociale che viene qui presentato è quello proposto da Barabási e Albert [5]. Secondo questo modello le reti sociali presentano quantomeno due caratteristiche principali denominate *growth* e *preferential attachment*.
La prima caratteristica specifica che il numero di nodi di una rete sociale cresce progressivamente. La seconda caratteristica invece specifica che i nodi tendono a connettersi preferibilmente ai nodi con più connessioni nella rete [4].
La creazione di una rete sociale con queste caratteristiche avviene come segue:

- Si parte da una rete avente un piccolo numero di nodi pari a m_0. A ogni istante di tempo viene aggiunto un nuovo nodo con un numero di archi pari a $m \leq m_0$. Questi archi connettono il nuovo nodo ai nodi già presenti nella rete. In questo modo viene soddisfatta la caratteristica denominata *growth*.
- La probabilità P che il nuovo nodo si connetta al nodo i della rete dipende dal grado del nodo i. In particolare questa probabilità è definita come:
$$P(k_i) = \frac{k_i}{\sum_j k_j},$$
dove k_i rappresenta il grado del nodo i. In questo modo viene soddisfatta anche la caratteristica di *preferential attachment*. Infatti i nodi

con grado più alto nella rete hanno una probabilità maggiore di ricevere connessioni dai nuovi nodi [19].

Una rete sociale che presenta queste due caratteristiche viene detta *scale-free*. In questo caso i gradi dei nodi sono distribuiti secondo una legge di potenza (*power-law*). In particolare una distribuzione a legge di potenza può essere espressa come:

$$p_k \sim k^{-\lambda},$$

dove k rappresenta il grado, p_k rappresenta la probabilità che un nodo presenti il grado k e λ rappresenta l'esponente del grado [4]. Questa distribuzione comporta che nella rete vi saranno pochi nodi con grado alto e molti nodi con grado basso. E questa proprietà è invariante per scala e nel tempo. In Figura 1 viene mostrata una distribuzione *power-law*.
Il modello di rete sociale proposto da Barabási e Albert presuppone che il grafo sottostante sia non orientato [24]. Una rappresentazione di una rete *scale-free* è mostrata in Figura 2.

2.2.2 Modello di Bollobás

Il secondo modello di rete sociale che viene qui presentato è stato proposto da Bollobás *et al.* nel 2003 [6]. Per semplicità si farà riferimento a questo modello come modello di Bollobás.
Il modello di Bollobás segue le orme del modello di Barabási e Albert. Infatti anche questo modello presenta le caratteristiche di *growth* e *preferential*

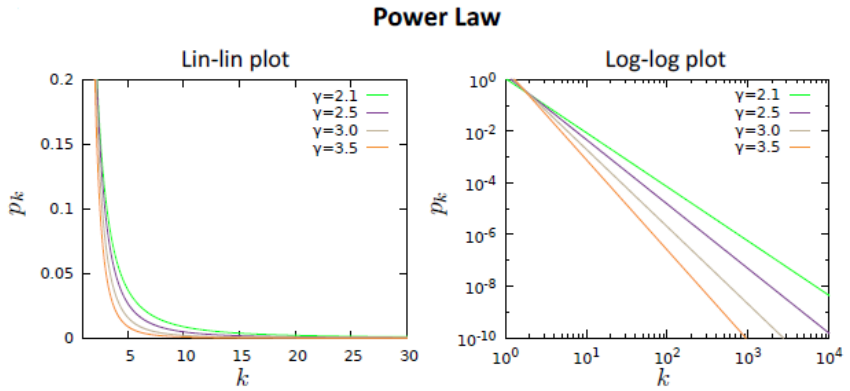

Figura 1: Rappresentazione di una distribuzione *power-law*. Questa rappresentazione è tratta da [4].

attachment. Pertanto la rete generata è *scale-free*. Tuttavia il modello di Bollobás differisce da quello di Barabási e Albert in quanto presuppone che il grafo sottostante sia orientato.
Il modello di Bollobás presenta diversi parametri:

- Una prima serie di parametri comprende i parametri α, β e γ. Questi parametri hanno un valore reale compreso fra 0 e 1. La loro somma deve essere pari a 1.

- Una seconda serie di parametri comprende i parametri δ_{in} e δ_{out}. Anche questi parametri hanno un valore reale compreso fra 0 e 1.

La costruzione della rete secondo il modello di Bollobás segue questa procedura. Sia G_0 il grafo iniziale al tempo t_0. Questo grafo presenta un solo vertice senza archi. Pertanto il numero iniziale di nodi $n(t_0)$ è pari a 1. Per ogni istante di tempo $t > t_0$ viene aggiunto un arco al grafo secondo queste regole:

- Con probabilità α viene aggiunto un nuovo nodo v e un arco da v a un nodo già esistente w. Il nodo w viene scelto con una probabilità proporzionale a $d_{\text{in}} + \delta_{\text{in}}$, dove d_{in} è l'*in-degree* del nodo w. Più

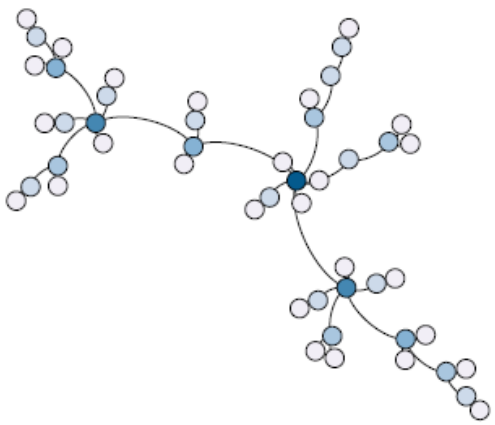

Figura 2: Rappresentazione di una rete *scale-free*. Il colore dei nodi rappresenta il grado. I nodi più scuri hanno un grado maggiore. Questa rappresentazione è tratta da [34].

precisamente:

$$P(w) = \frac{(d_{\text{in}}(w) + \delta_{\text{in}})}{(t + \delta_{\text{in}} n(t))}.$$

- Con probabilità β viene aggiunto un arco da un nodo già esistente v a un altro nodo già esistente w. In questo caso il nodo w viene scelto con una probabilità proporzionale a $d_{\text{in}}+\delta_{\text{in}}$, come appena illustrato. Invece il nodo v viene scelto con una probabilità proporzionale a $d_{\text{out}} + \delta_{\text{out}}$, dove d_{out} è l'*out-degree* del nodo v. Più precisamente:

$$P(v) = \frac{(d_{\text{out}}(v) + \delta_{\text{out}})}{(t + \delta_{\text{out}} n(t))}.$$

- Con probabilità γ viene aggiunto un nuovo nodo w e un arco da un nodo già esistente v al nodo w. In questo caso il nodo v viene scelto con una probabilità proporzionale a $d_{\text{out}} + \delta_{\text{out}}$, come appena illustrato.

Il modello di Bollobás ammette cappi e multi-archi.

2.2.3 Modello di Anwar

Il terzo modello di rete sociale che viene qui presentato è stato proposto da Anwar *et al.* nel 2021 [2]. Per semplicità si farà riferimento a questo modello come modello di Anwar.
Questo modello rappresenta un'evoluzione del modello di Bollobás. Infatti il modello di Anwar aggiunge l'omofilia al modello di Bollobás. In particolare in una rete sociale l'omofilia esprime la propensione dei nodi a legarsi ad altri nodi con caratteristiche simili. In questo caso la similitudine riguarda le opinioni dei nodi che rappresentano gli utenti della rete sociale. Nel modello di Anwar le opinioni hanno valore binario: 0 o 1. Pertanto esistono solo due categorie di utenti.
Poiché il modello di Anwar rappresenta un'evoluzione del modello di Bollobás, genera anch'esso una rete *scale-free*.
Anche il modello di Anwar utilizza i parametri α, β, γ, δ_{in} e δ_{out} già descritti per il modello di Bollobás. A questi parametri si aggiunge il parametro h che regola l'omofilia della rete. Questo parametro ha un valore reale compreso fra 0 e 1. E viene utilizzato per calcolare il valore $h(v, w)$ per ogni coppia di nodi come segue:

$$h(v, w) = \begin{cases} h & \text{se } v \text{ e } w \text{ appartengono alla stessa categoria} \\ 1 - h & \text{altrimenti} \end{cases}$$

Inoltre il modello di Anwar utilizza il parametro p_M che rappresenta la frazione di nodi appartenenti alla categoria di maggioranza che si vuole ottenere con la generazione della rete.
La costruzione della rete secondo il modello di Anwar segue questa procedura. Sia G_0 il grafo iniziale al tempo t_0. Questo grafo presenta un solo vertice senza archi. Pertanto il numero iniziale di nodi $n(t_0)$ è pari a 1. Per ogni istante di tempo $t > t_0$ viene aggiunto un arco al grafo secondo queste regole:

- Con probabilità α viene aggiunto un nuovo nodo v e un arco da v a un nodo già esistente w. Il nodo v viene assegnato alla categoria di maggioranza con probabilità p_M. Altrimenti viene assegnato alla categoria di minoranza. Il nodo w viene scelto con una probabilità proporzionale a $h(v,w)d_{\text{in}} + \delta_{\text{in}}$. Più precisamente:
$$P(w) = \frac{(h(v,w)d_{\text{in}}(w) + \delta_{\text{in}})}{(\sum_z h(v,z)d_{\text{in}}(z) + \delta_{\text{in}})}.$$
- Con probabilità β viene aggiunto un arco da un nodo già esistente v a un altro nodo già esistente w. In questo caso il nodo w viene scelto con una probabilità proporzionale a $h(v,w)d_{\text{in}} + \delta_{\text{in}}$. Invece il nodo v viene scelto con una probabilità proporzionale a $d_{\text{out}} + \delta_{\text{out}}$. Più precisamente:
$$P(v) = \frac{(d_{\text{out}}(v) + \delta_{\text{out}})}{(\sum_z d_{\text{out}}(z) + \delta_{\text{out}})}.$$
- Con probabilità γ viene aggiunto un nuovo nodo w e un arco da un nodo già esistente v al nodo w. Il nodo w viene assegnato alla categoria di maggioranza con probabilità p_M. Altrimenti viene assegnato alla categoria di minoranza. In questo caso il nodo v viene scelto con una probabilità proporzionale a $h(v,w)d_{\text{out}} + \delta_{\text{out}}$. Più precisamente:
$$P(v) = \frac{(h(v,w)d_{\text{out}}(v) + \delta_{\text{out}})}{(\sum_z h(z,w)d_{\text{out}}(z) + \delta_{\text{out}})}.$$

Anche il modello di Anwar ammette cappi e multi-archi.

2.3 Modello proposto

In questo Paragrafo verrà descritto il modello innovativo proposto con questo lavoro. Prima di descrivere il modello occorre però svolgere alcune considerazioni preliminari in modo da rendere più chiare le scelte modellistiche effettuate.

Obiettivo. La scelta del modello è stata guidata dall'obiettivo prefissato. In particolare l'obiettivo è ottenere una rete sociale su cui simulare la diffusione di *fake news*. Le reti sociali presentano tipicamente le caratteristiche di *growth* e *preferential attachment* descritte al Paragrafo 2.2.1. Inoltre gli utenti delle reti sociali presentano tipicamente opinioni variegate e difficilmente catalogabili in due sole categorie. E le opinioni degli utenti influiscono sulla topologia della rete attraverso la formazione di *echo chamber*. Infine la rete sociale che si vuole modellare deve essere basata su un grafo non orientato. Il motivo di questa scelta risiede nei *dataset* utilizzati per la validazione, come verrà chiarito meglio nella Sezione 5.

Modello. Il modello innovativo proposto si basa sul modello di Anwar. Infatti quest'ultimo modello: (*i*) soddisfa le caratteristiche di *growth* e *preferential attachment*; (*ii*) considera l'omofilia nella formazione della rete; e (*iii*) presuppone che il grafo della rete sociale sia orientato.
Tuttavia il modello proposto differisce dal modello di Anwar per alcune caratteristiche significative. In particolare:

- Anzitutto le opinioni dei nodi hanno un valore (non più binario, ma) reale compreso fra 0 e 1. In questo modo si ritiene di cogliere meglio la varietà di opinioni degli utenti di un *social network*. In particolare le opinioni sono campionate da una distribuzione uniforme analogamente a quanto proposto in [7].

- Inoltre il fattore $h(v, w)$ utilizzato nel modello di Anwar viene ridefinito come:
$$h(v, w) = |h - |op_v - op_w||,$$
dove op_v è l'opinione del nodo v. In questo modo il fattore $h(v, w)$ viene adattato al fatto che l'opinione dei nodi ha valore (non più binario, ma) reale. L'utilizzo dei valori assoluti produce due effetti[2]. Il primo è che il fattore $h(v, w)$ non può assumere valori negativi o superiori a 1. Il secondo è che se il parametro h è minore di 0.5, allora la probabilità di collegare nodi con opinioni distanti cresce. Viceversa se il parametro h è maggiore di 0.5, allora la probabilità di collegare nodi con opinioni distanti decresce. In questo modo viene modellata correttamente l'omofilia nella rete.

[2]Si noti che il calcolo di $|op_v - op_w|$ equivale al calcolo della distanza euclidea tra i valori reali delle opinioni dei nodi v e w.

- Infine il modello proposto non prevede né cappi né multi-archi. Infatti per questo lavoro l'orientamento degli archi della rete rappresenta la relazione unidirezionale di *follow* tipica di alcuni *social media* come Twitter [37]. E tipicamente nei *social media* un utente non può (*i*) né instaurare una relazione di *follow* con se stesso (*ii*) né instaurare due diverse relazioni di *follow* con lo stesso utente.

La creazione della rete termina quando viene raggiunto il numero di nodi desiderato.
Una volta creata la rete non è prevista alcuna possibilità di modificarne la struttura. Quindi non possono essere aggiunti né rimossi nodi o archi. Questo non invaliderà le conclusioni raggiunte sulla validazione del modello esposte nella Sezione 5. Infatti si ritiene che i cambiamenti della struttura della rete siano trascurabili quando la validazione avviene su un *dataset* circoscritto in un arco temporale limitato [8]. E per questo motivo per la validazione sono stati usati *dataset* che coprono solo pochi mesi di tempo.
I parametri del modello proposto sono riassunti nella tabella mostrata come Figura 3.
In sintesi è stato implementato un modello di rete sociale che si basa su un grafo (*i*) diretto; (*ii*) *scale-free*; (*iii*) con omofilia; e (*iv*) con opinioni con valore continuo. Per l'implementazione del modello è stato utilizzato il linguaggio di programmazione Python e la libreria `NetworkX` [14].

Parametri	Significato
n	Numero di nodi della rete
α	Probabilità di aggiungere un nuovo arco che ha come sorgente un nodo nuovo
β	Probabilità di aggiungere un nuovo arco tra due nodi già esistenti
γ	Probabilità di aggiungere un nuovo arco che ha come destinazione un nodo nuovo
h	Parametro che regola l'omofilia della rete
δ_{in}	Parametro utilizzato per scegliere un nodo già esistente come destinazione di un arco che parte da un altro nodo
δ_{out}	Parametro utilizzato per scegliere un nodo già esistente come sorgente di un arco che ha come destinazione un altro nodo

Figura 3: Parametri del modello di rete sociale proposto.

3 Modello di diffusione di *fake news*

Sin qui è stato descritto lo stato dell'arte in tema di modellazione di reti sociali ed è stato presentato il modello innovativo proposto per questo lavoro. Ora si procederà a presentare il modello di diffusione di *fake news*. Per fare questo si seguirà questo ordine. Anzitutto verranno descritti alcuni modelli rappresentanti lo stato dell'arte che sono strettamente correlati al modello qui proposto. Dopodiché verrà presentato il modello innovativo proposto per la diffusione di *fake news*.

3.1 Stato dell'arte

3.1.1 Modello di Serrano

Il primo modello di diffusione di *fake news* che viene qui presentato è stato proposto da Serrano *et al.* [35, 34]. Per semplicità si farà riferimento a questo modello come modello di Serrano.

Modello SIR. Il modello di Serrano trae ispirazione dal modello epidemiologico proposto da Kermack e McKendrick nel 1927 [17]. Quest'ultimo modello suddivide la popolazione in tre categorie: persone suscettibili all'infezione (S); persone infette (I); e persone guarite dall'infezione o morte (R). La transizione degli individui tra le categorie è regolata da equazioni differenziali. Il nome delle categorie del modello di Kermack e McKendrick ha portato a identificare comunemente questo modello come modello SIR.
Ora, la diffusione di *fake news* presenta analogie significative con la diffusione epidemica. Infatti anche in questo caso vi sono: (*i*) individui che promuovono la divulgazione di *fake news*; (*ii*) individui che credono all'informazione falsa; e (*iii*) individui che smentiscono (o non credono più) all'informazione falsa. Per questo motivo il modello di Kermack e McKendrick offre una base anche per la modellazione della diffusione di *fake news*.

Categorie di popolazione. In particolare il modello di Serrano innova il modello SIR suddividendo la popolazione in quattro categorie:

- ***neutral***: individui suscettibili di credere o no alla *fake news*;
- ***infected***: individui che credono alla *fake news*;

- ***vaccinated***: individui che non credono alla *fake news* senza mai essere stati infettati;
- ***cured***: individui che non credono più alla *fake news* dopo essere stati infettati.

Parametri. Il modello di Serrano prevede l'utilizzo di alcuni parametri. In particolare prevede:

- un numero iniziale di persone infette;
- un parametro rappresentante la probabilità di un individuo infetto di infettare i vicini non infetti;
- un parametro rappresentante la probabilità che un individuo neutrale diventi vaccinato quando un vicino infetto prova a infettarlo;
- un parametro rappresentante la probabilità che un individuo vaccinato provi a curare un vicino infetto o vaccinare un vicino neutrale.

A ogni istante di tempo gli individui infetti provano a infettare i vicini innescando un meccanismo di infezione, vaccinazione e cura. In questo meccanismo gli individui curati assumono un ruolo puramente passivo in quanto non possono provocare né la vaccinazione né la cura dei vicini.
In Figura 4 viene mostrato uno schema delle possibili transizioni di stato epidemiologico secondo il modello di Serrano.

Rete sociale. Infine si precisa che il modello di Serrano presuppone l'utilizzo di una rete generata secondo il modello di Barabási-Albert illustrato al paragrafo 2.2.1. Pertanto presuppone che il grafo di nodi e archi sia non orientato.

3.1.2 Modello di Gausen

Il secondo modello di diffusione di *fake news* che viene qui presentato è stato proposto da Gausen *et al.* nel 2021 [12]. Per semplicità si farà riferimento a questo modello come modello di Gausen.

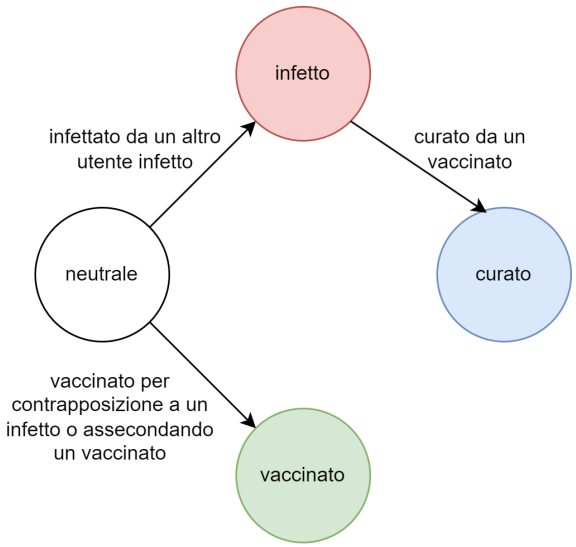

Figura 4: Schema delle transizioni di stato epidemiologico secondo il modello di Serrano.

Categorie di popolazione. Il modello di Gausen si basa sul modello di Serrano. In particolare suddivide la popolazione nelle stesse categorie previste dal modello di Serrano, sebbene con nomi diversi. Più precisamente ogni individuo può assumere questi stati:

- ***susceptible*** che corrisponde allo stato *neutral* del modello di Serrano;
- ***believe*** che corrisponde allo stato *infected* del modello di Serrano;
- ***deny*** che corrisponde allo stato *vaccinated* del modello di Serrano;
- ***cured*** che corrisponde all'omonimo stato del modello di Serrano.

Parametri. Il modello di Gausen eredita dal modello di Serrano anche i relativi parametri. In particolare tra questi si ricordano: (*i*) la probabilità che gli individui infetti contagino i vicini (P_{inf}); (*ii*) la probabilità degli individui vaccinati di vaccinare i vicini neutrali (P_{vacc}); e (*iii*) la probabilità degli individui vaccinati di curare i vicini infetti e degli individui neutrali di diventare vaccinati a contatto con la *fake news* (P_{deny}).

Contromisure. D'altro canto il modello di Gausen si differenzia dal modello di Serrano per alcune peculiarità significative. Queste peculiarità riguardano principalmente le contromisure alla diffusione di *fake news*. Più precisamente:

- Anzitutto il modello di Gausen introduce la figura dell'*influencer*. Questa figura è ortogonale alle categorie di popolazione appena ricordate. In particolare l'*influencer* possiede una maggiore probabilità di infettare o vaccinare. Questa probabilità viene modellata tramite il parametro P_{infl}.

- Inoltre il modello di Gausen introduce la possibilità di bloccare gli individui infetti. In particolare un individuo vaccinato può presentare un reclamo contro un individuo infetto con una probabilità P_{block} che viene stimata essere pari a 0.1. Un utente che riceva almeno 3 reclami viene bloccato. L'utente bloccato non può più interagire con gli altri utenti.

- Ancora, il modello di Gausen introduce il parametro P_{inoc}. Questo parametro riduce la probabilità di infezione, che diventa:

  $$P_{\text{inf}} = P_{\text{inf}} - P_{\text{inoc}}.$$

  Il parametro P_{inf} viene stimato essere pari a 0.065.

- Infine il modello di Gausen introduce il parametro P_{acc}. Questo parametro incrementa la probabilità di vaccinazione, che diventa:

  $$P_{\text{vacc}} = P_{\text{vacc}} + P_{\text{acc}}.$$

  Il parametro P_{acc} viene stimato essere pari a 0.048.

3.1.3 Modello di Lotito

Il terzo modello di diffusione di *fake news* che viene qui presentato è stato proposto da Lotito *et al.* nel 2021 [21]. Per semplicità si farà riferimento a questo modello come modello di Lotito.

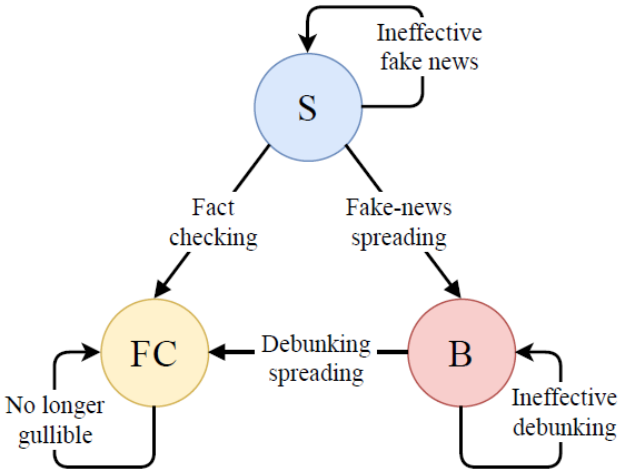

Figura 5: Schema delle transizioni di stato epidemiologico secondo il modello di Lotito. Lo stato "S" rappresenta lo stato *susceptible*. Lo stato "B" rappresenta lo stato *believer*. Lo stato "FC" rappresenta lo stato *fact-checker*. Questo schema è tratto da [21].

Tassonomie della popolazione. Il modello di Lotito prevede due tassonomie della popolazione. Queste tassonomie sono sovrapposte e ortogonali tra loro.
La prima tassonomia riprende le categorie del modello SIR. Più precisamente gli utenti vengono suddivisi in:

- ***susceptible***: utenti suscettibili di credere alla *fake news*;
- ***believer***: utenti che credono alla *fake news*;
- ***fact-checker***: utenti che non credono alla *fake news* perché l'hanno verificata o perché sono entrati in contatto con un altro *fact-checker*.

Le transizioni di stato epidemiologico sono riassunte nella Figura 5.
La seconda tassonomia attribuisce invece un ruolo agli utenti. In particolare gli utenti possono essere:

- ***common***: normali utenti della rete;
- ***influencer***: utenti che hanno una maggiore capacità diffusiva rispetto ai *common*;

- ***bot***: *account* automatizzati che diffondono la *fake news* e la mantengono viva;
- ***eternal fact-checker***: utenti che combattono continuamente la diffusione di *fake news*.

La sovrapposizione tra le due tassonomie opera come segue:

- i *common* e gli *influencer* possono essere *susceptible*, *believer* o *fact-checker*;
- i *bot* possono essere solo *believer*;
- gli *eternal fact-checker* possono essere solo *fact-checker*.

Inoltre ogni utente possiede attributi specifici. In particolare ogni utente presenta:

- una probabilità di seguire opinioni altrui. Questa probabilità viene campionata da una distribuzione normale con media 0.5 e deviazione standard 0.2;
- una probabilità di condividere una notizia. Questa probabilità viene campionata da una distribuzione normale con media 0.5 e deviazione standard 0.2;
- una probabilità di verificare se la notizia è vera (operazione di *fact-checking*). Questa probabilità viene campionata da una distribuzione normale con media 0.2 e deviazione standard 0.2;
- cinque attributi di interesse campionati da una distribuzione normale con media 0 e deviazione standard 0.4. Questi attributi di interesse hanno valore compreso fra -1 e 1;
- una coppia di coordinate geografiche campionate da una distribuzione uniforme con valore compreso fra 0 e 1.

Rete sociale. Dopo aver dettagliato gli utenti, il modello di Lotito procede a definire la costruzione di una rete sociale con determinate caratteristiche. In questa rete i nodi rappresentano gli utenti e gli archi rappresentano le relazioni tra gli utenti. Più precisamente:

- se la distanza euclidea tra le coordinate geografiche di due nodi è minore di 0.03, allora viene creato un arco tra i nodi;
- se la distanza euclidea tra gli attributi di interesse di due nodi è minore di 0.03, allora viene creato un arco tra i nodi;
- gli archi creati secondo le precedenti due regole vengono rimossi con probabilità pari a 0.5 in modo da introdurre maggiore casualità nella rete;
- a ogni arco viene attribuito un peso pari alla media della distanza geografica e della distanza degli attributi di interesse. Questo peso rappresenta una costante che influisce sulla probabilità di infezione;
- i *bot* sono collegati ad altri nodi in modo casuale con una probabilità pari a 0.02;
- a ogni arco che collega un *bot* a un altro nodo viene attribuito un peso pari a 0.1.

In questo quadro la topologia della rete è influenzata direttamente dall'omofilia tra i nodi.

Dinamica temporale. Infine il modello di Lotito disciplina la dinamica temporale sotto due profili:

- Anzitutto si prevede che gli utenti accedano al *social network* (non a ogni istante di tempo, ma) saltuariamente. In particolare il tempo di accesso degli utenti è modellato secondo una distribuzione esponenziale con media pari a $\frac{1}{\lambda_{\text{common}}}$, dove $\lambda_{\text{common}} = 16(\text{minuti}^{-1})$. I *bot* accedono al *social network* con una frequenza quattro volte maggiore.
- Inoltre l'interesse degli utenti a diffondere le notizie (vere o false) decresce col tempo. Per modellare questa decrescita viene introdotto un coefficiente di coinvolgimento E la cui variazione nel tempo segue questa equazione differenziale:

$$\frac{\partial E}{\partial t} = -\lambda E,$$

dove t è l'istante di tempo e λ è la costante di decrescita. Quest'ultima viene posta pari a $\frac{2}{T_{\text{sim}}}$, dove T_{sim} è l'istante di tempo della simulazione. La soluzione della precedente equazione differenziale è pari a:

$$E(t) = E_0 e^{-\lambda t}.$$

3.1.4 Modello di Törnberg

Il quarto modello di diffusione di *fake news* che viene qui presentato è stato proposto da Törnberg nel 2018 [38]. Questo modello è stato ideato per studiare l'impatto delle *echo chamber* sulla disinformazione.

Rete sociale. Il modello di Törnberg utilizza un grafo non orientato creato secondo il modello di Erdős–Rényi per simulare la diffusione di *fake news* [10]. In particolare il modello di Erdős–Rényi prevede che i nodi vengano connessi tra loro in modo random con uguale probabilità. Se il numero di nodi della rete è molto grande, la distribuzione dei gradi dei nodi segue una distribuzione (non a legge di potenza, ma) di Poisson [25].

Echo chamber. Il modello di Törnberg individua due proprietà delle *echo chamber*:

- La prima proprietà è la polarizzazione delle opinioni. Questa proprietà determina quanto gli utenti condividano la stessa opinione su un argomento ed è quindi strettamente legata all'omofilia. Questa proprietà viene modellata tramite la probabilità P_o.
- La seconda proprietà è la polarizzazione della rete. Questa proprietà determina quanto gli utenti appartenenti all'*echo chamber* siano densamente connessi rispetto al resto della rete. Questa proprietà viene modellata tramite la probabilità P_n.

A questi parametri il modello di Törnberg aggiunge anche: il parametro k rappresentante il grado medio dei nodi della rete; il parametro θ come soglia di attivazione/infezione dei nodi; e il parametro c rappresentante la frazione di nodi della rete che appartiene all'*echo chamber*.

Modello di diffusione. La diffusione di *fake news* nel modello di Törnberg avviene come segue:

1. Anzitutto viene creata una rete sociale secondo il modello di Erdős–Rényi. Questa rete presenta N nodi dei quali cN nodi appartengono all'*echo chamber*.

2. Dopodiché viene calcolato il numero di archi $E = N\frac{k}{2}$ ottenuto dividendo il grado medio k per 2, dato che il grafo è non orientato.

3. Successivamente viene selezionato un numero di archi pari a $P_n kE$ tali per cui esattamente un estremo di ogni arco rappresenta un nodo dell'*echo chamber*. Questi archi sono rimossi e vengono sostituiti con altrettanti archi tra nodi interni all'*echo chamber*.

4. A questo punto vengono fissate le soglie di attivazione. In particolare i nodi esterni all'*echo chamber* hanno una soglia di attivazione pari a θ; i nodi interni all'*echo chamber* hanno una soglia di attivazione pari a $\theta - P_o$. Questo comporta che la soglia di attivazione per i nodi interni all'*echo chamber* sia più bassa rispetto al resto della rete. In questo modo i nodi dell'*echo chamber* sono più propensi a credere alla *fake news*.

5. Poi viene attivato un nodo random dell'*echo chamber*. E parimenti vengono attivati i nodi collegati a esso.

6. Infine viene disciplinata l'attivazione degli altri nodi. In particolare un nodo viene attivato se i nodi collegati a esso sono attivati per una frazione superiore alla specifica soglia di attivazione. L'attivazione del nodo avviene nell'istante di tempo successivo al superamento della soglia.

3.2 Modello proposto

In questo paragrafo verrà descritto il modello innovativo proposto per la diffusione di *fake news*. Le caratteristiche principali di questo modello possono essere riassunte come segue: (*i*) la popolazione viene divisa secondo le categorie epidemiologiche del modello di Serrano; (*ii*) agli utenti viene attribuito un ruolo in conformità al modello di Lotito; (*iii*) gli utenti possono essere bloccati analogamente al modello di Gausen; *(iv)* il coinvolgimento degli utenti decresce nel tempo in modo simile al modello di Lotito; e (*v*) la diffusione

delle *fake news* viene influenzata dalla presenza di *echo chamber* attraverso alcuni meccanismi illustrati nel modello di Törnberg. Ora verranno descritte più approfonditamente le caratteristiche del modello proposto.

Le categorie epidemiologiche. Le categorie epidemiologiche in cui è divisa la popolazione sono riprese dal modello di Serrano (v. Paragrafo 3.1.1). Più precisamente gli utenti possono essere:

- **neutrali**: utenti suscettibili di infezione o vaccinazione in quanto non ancora entrati in contatto con persone infette o vaccinate;
- **infetti**: utenti che credono alla *fake news* e cercano di diffonderla (salvo il blocco di cui *infra*);
- **vaccinati**: utenti che non credono alla *fake news* e che non sono mai stati infettati;
- **curati**: utenti che non credono più alla *fake news* in quanto curati da utenti vaccinati.

Il modello proposto presenta alcuni parametri che regolano la transizione tra le categorie. In particolare:

- un primo parametro P_{inf} rappresenta la probabilità di un utente neutrale di essere infettato quando legge un messaggio di un utente infetto;
- un secondo parametro P_{vacc} rappresenta la probabilità di un utente neutrale di essere vaccinato quando legge un messaggio di un utente infetto o di un utente già vaccinato;
- un terzo parametro P_{cure} rappresenta la probabilità di un utente infetto di essere curato quando legge un messaggio di un utente vaccinato.

I ruoli. Il modello proposto associa agli utenti un ruolo. Questo ruolo è ortogonale allo stato epidemiologico dell'utente. In particolare i ruoli considerati sono ripresi dal modello di Lotito (v. Paragrafo 3.1.3). Più precisamente i ruoli sono i seguenti:

- ***common***: normali utenti della rete;

- ***influencer***: utenti che hanno un elevato numero di *follower*. A questi utenti viene attribuita una probabilità $P_{\text{influencer}}$. Questa probabilità si somma alle probabilità degli utenti di essere infettati, vaccinati o curati. Quindi:

$$P_{\text{inf}} = P_{\text{inf}} + P_{\text{influencer}},$$
$$P_{\text{vacc}} = P_{\text{vacc}} + P_{\text{influencer}},$$
$$P_{\text{cure}} = P_{\text{cure}} + P_{\text{influencer}}.$$

  Pertanto un messaggio di un *influencer* ha una maggiore efficacia diffusiva. Naturalmente rimane invariata la probabilità di essere vaccinati come risposta a un messaggio di un utente infetto;

- ***bot***: account automatizzati che diffondono continuamente la *fake news* e non possono essere curati;

- ***eternal fact-checker***: utenti che combattono continuamente la diffusione della *fake news* e non possono essere infettati.

Come nel modello di Lotito, anche qui i *common* e gli *influencer* possono essere neutrali, infetti, vaccinati o guariti. I *bot* possono essere solo infetti. Gli *eternal fact-checker* possono essere solo vaccinati. In questo caso si assume che gli *influencer* non siano né *bot* né *eternal fact-checker*.

Il blocco degli utenti. Ogni utente pubblica un messaggio negli istanti di tempo in cui è attivo. A questa regola fanno eccezione gli utenti curati. Questi utenti non pubblicano alcun messaggio dopo la cura. Si vuole in questo modo replicare il modello di Serrano, per il quale gli utenti che hanno riconosciuto l'errore nel credere alla *fake news* non sono comunque psicologicamente propensi a combatterne la diffusione [35].
Un'altra eccezione alla regola predetta è rappresentata dal blocco degli utenti. In particolare il modello proposto prevede che un utente vaccinato presenti un reclamo nei confronti di un utente infetto con una probabilità $P_{\text{complaint}}$ quando vede un messaggio di quest'ultimo. Quando un utente riceve 3 reclami viene bloccato. Dopo essere stato bloccato l'utente non può più pubblicare messaggi. Questo meccanismo di blocco replica in parte il meccanismo proposto nel modello di Gausen (v. Paragrafo 3.1.2).

La dinamica temporale. La diffusione della *fake news* è scandita dagli istanti di tempo della simulazione. In particolare gli utenti controllano i messaggi degli utenti "seguiti" negli istanti di tempo in cui sono attivi. E il controllo consiste nella lettura dei messaggi che gli utenti "seguiti" hanno pubblicato dall'ultimo istante di tempo in cui l'utente è stato attivo. Per ogni messaggio letto si attiva il meccanismo di infezione, vaccinazione e cura illustrato *supra*.
Il modello proposto prevede che l'interazione degli utenti sul tema della *fake news* decresca nel tempo. Per modellare questa decrescita è stato utilizzato un fattore di decrescita esponenziale analogamente al modello di Lotito (v. Paragrafo 3.1.3). In particolare il prossimo istante di attività dell'utente è determinato come:

$$t_{\text{next}} = t_{\text{current}} + \max\left\{1, \left\lfloor \frac{1}{e^{-\lambda t_{\text{current}}}} \right\rceil\right\},$$

dove λ è campionato da una distribuzione uniforme con valore compreso fra 0.01 e 0.05 individuato come appropriato attraverso valutazioni empiriche. Questa equazione somma all'istante corrente il massimo fra 1 e $\left\lfloor \frac{1}{e^{-\lambda t_{\text{current}}}} \right\rceil$ in modo tale che il prossimo istante di tempo non sia mai uguale a quello corrente. Il termine frazionario viene arrotondato all'intero più vicino in modo da ottenere un nuovo istante di tempo intero. In questo quadro il prossimo istante di tempo in cui l'utente sarà attivo tenderà ad essere sempre più distante con l'avanzare degli istanti di tempo delle simulazioni.
Nel modello proposto i *bot* sono invece attivi in ogni istante di tempo in quanto si suppone che siano (non persone fisiche, ma) *account* automatizzati.

Omofilia ed *echo chamber*. L'ultima caratteristica del modello di diffusione proposto riguarda l'omofilia e le *echo chamber*. In particolare al Paragrafo 2.3 è stata precisata l'influenza dell'omofilia sulla topologia della rete. Questa influenza si traduce nella creazione di *echo chamber*. E l'efficacia diffusiva nelle *echo chamber* è particolarmente forte, come specificato da Törnberg [38].
In questo quadro si è deciso di arricchire il modello proposto al fine di tenere in considerazione questa maggiore efficacia diffusiva. Più precisamente:

- Ogni utente presenta un'opinione rappresentata da un valore reale compreso fra 0 e 1, come già chiarito al Paragrafo 2.3.

- La probabilità di infezione, vaccinazione e cura è influenzata dalla differenza fra le opinioni degli utenti. In particolare viene definito un fattore op_{diff} calcolato come segue:

$$op_{\text{diff}} = 1 - |op_{\text{x}} - op_{\text{x}}|,$$

  dove op_{x} e op_{y} rappresentano rispettivamente le opinioni dei nodi x e y. Il valore assoluto consente di evitare valori negativi della differenza fra le opinioni. Il valore assoluto viene sottratto a 1 in modo tale che il fattore ottenuto sia più grande quando la differenza fra le opinioni è piccola. Infatti questo fattore rappresenta la fiducia che un utente ripone nell'utente di cui legge i messaggi.

- Il fattore op_{diff} viene moltiplicato alle probabilità di infezione, vaccinazione e cura. Pertanto queste probabilità vengono ridefinite come segue:

$$P_{\text{inf}} = P_{\text{inf}} * op_{\text{diff}},$$

$$P_{\text{vacc}} = P_{\text{vacc}} * op_{\text{diff}},$$

$$P_{\text{cure}} = P_{\text{cure}} * op_{\text{diff}}.$$

  Inoltre se il messaggio proviene da un *influencer*, viene comunque sommata la probabilità $P_{\text{influencer}}$ come illustrato in precedenza.

- Infine viene modellata una soglia P_{echo}. In particolare:

  - un utente neutrale diventa infetto se la frazione di utenti infetti da lui "seguiti" supera questa soglia;
  - un utente neutrale diventa vaccinato se la frazione di utenti vaccinati da lui "seguiti" supera questa soglia;
  - un utente infetto diventa curato se la frazione di utenti vaccinati da lui "seguiti" supera questa soglia.

  Pertanto se la soglia è superiore a 0.5, l'utente seguirà l'orientamento della maggioranza degli altri utenti appartenenti all'*echo chamber*. In particolare nel modello proposto si è previsto che il meccanismo di *echo chamber* appena descritto si attivi solo quando la soglia è superiore a 0.5. In questo modo si evita che un utente possa essere indotto a seguire un'opinione minoritaria all'interno di un'*echo chamber*.

Parametri	Significato
P_{inf}	Probabilità di essere infettato
P_{vacc}	Probabilità di essere vaccinato
P_{cure}	Probabilità di essere curato
$P_{influencer}$	Maggiore probabilità degli *influencer* di infettare o vaccinare o curare
P_{echo}	Soglia per il meccanismo delle *echo chamber*
$P_{complaint}$	Probabilità di presentare un reclamo contro un utente infetto

Figura 6: Parametri del modello di diffusione delle *fake news* proposto.

Sintesi. In questo quadro le transizioni di stato epidemiologico possono essere riassunte come mostrato in Figura 7. I parametri del modello proposto sono riassunti nella tabella riportata come Figura 6.
Il *workflow* delle attività di un utente in ogni istante di tempo è invece riassunto in Figura 8. In particolare l'utente verifica subito se il proprio stato debba essere aggiornato in base al meccanismo delle *echo chamber*[3]. Se lo stato dell'utente non deve essere mutato, procede alla lettura dei messaggi degli utenti "seguiti" che sono stati pubblicati dopo il suo ultimo istante di attività. L'ordine degli utenti "seguiti" di cui vengono letti i messaggi è reso casuale in modo da ridurre eventuali *bias*. Infine l'utente pubblica un proprio messaggio e definisce il prossimo istante di attività. L'aggiornamento dello stato degli utenti avviene solo dopo che tutti hanno terminato la propria attività. L'ordine di aggiornamento dello stato degli utenti è ancora casuale in modo da ridurre eventuali *bias*. L'aggiornamento del blocco degli utenti viene compiuto dalla rete al termine dell'attività di tutti gli utenti.

[3]Si è preferito verificare subito il mutamento di stato per effetto dell'*echo chamber* in modo da ottenere *echo chamber* coerenti di utenti che condividano (non solo l'opinione, ma) anche lo stato.

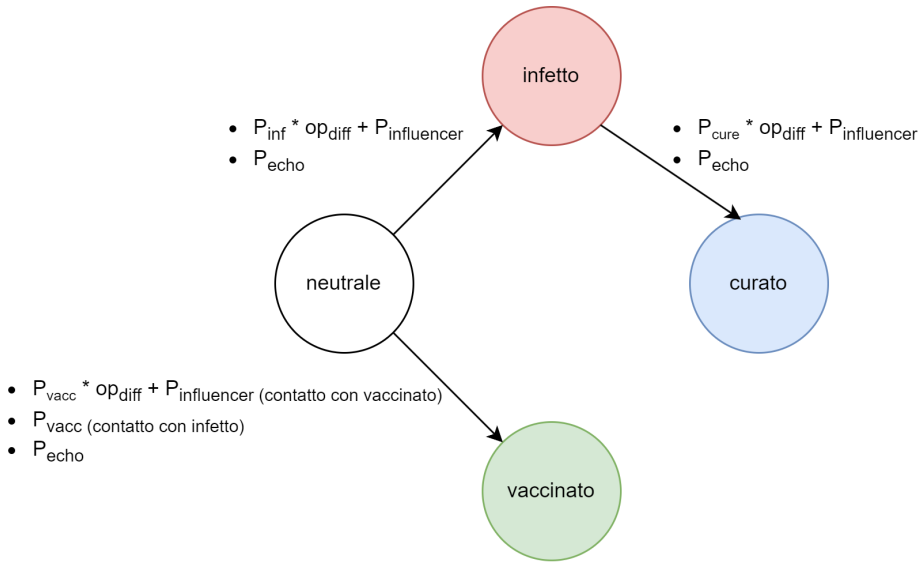

Figura 7: Schema delle transizioni di stato epidemiologico secondo il modello proposto.

4 Modellazione basata su agenti

Sin qui sono stati descritti i modelli proposti per simulare la diffusione di *fake news* in *social network*. In particolare questi modelli sono stati implementati come modelli basati su agenti. Infatti un modello basato su agenti è un sistema artificiale che consente di simulare le interazioni tra agenti autonomi al fine di individuare comportamenti emergenti [9]. E questo risulta appropriato nel caso di specie in quanto occorre simulare le interazioni tra gli utenti di un *social network* nella diffusione di *fake news*.
In questa Sezione verranno inquadrati i modelli proposti all'interno dei modelli basati su agenti. Anzitutto verranno richiamate alcune nozioni riguardanti i sistemi complessi. Dopodiché verranno dettagliate le caratteristiche degli agenti. Successivamente verranno descritte le particolarità dell'ambiente proposto. Infine verranno specificate le tipologie di interazioni impiegate nei modelli proposti.

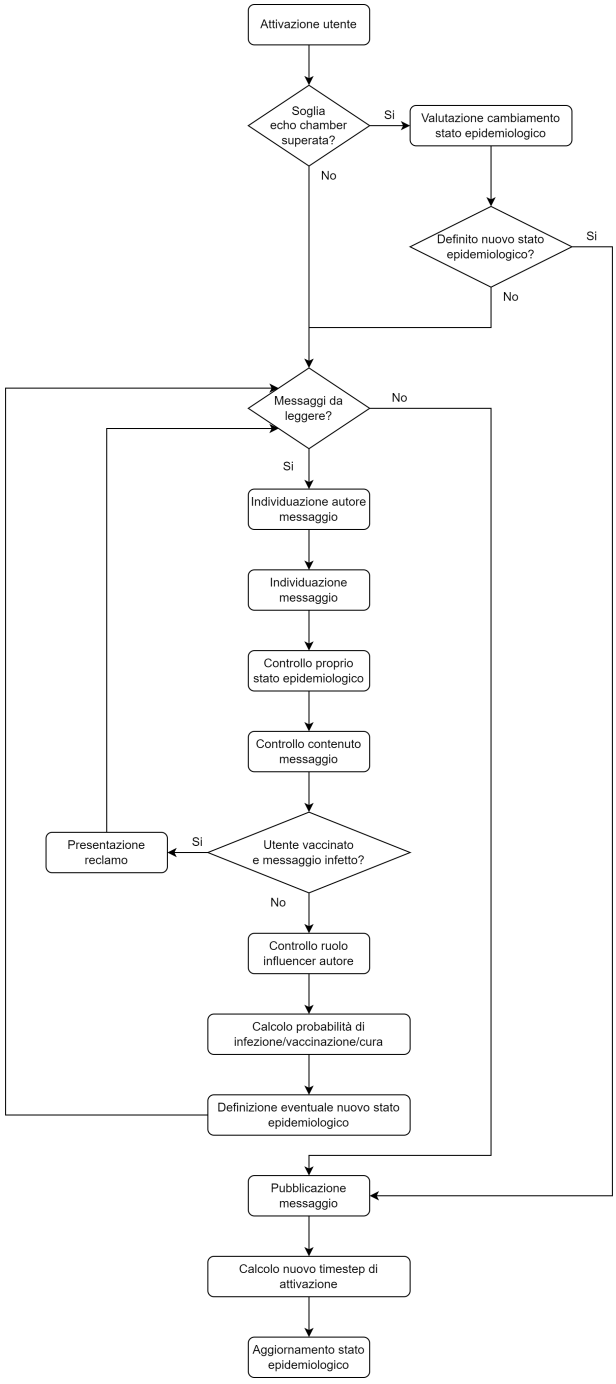

Figura 8: *Workflow* delle attività di un utente in ogni istante di tempo.

4.1 Sistema complesso

La modellazione basata su agenti è una tipica tecnica di modellazione di sistemi complessi. In particolare un sistema complesso è caratterizzato da una moltitudine di agenti che interagiscono tra loro e che esibiscono proprietà emergenti descrivibili da modelli matematici non lineari [32, 11]. Le proprietà emergenti rappresentano comportamenti derivanti dall'interazione delle componenti del sistema sia tra loro che con l'ambiente. Pertanto un sistema complesso risulta più espressivo della somma delle sue componenti [15].
Ora, un *social network* è una rete che (*i*) prevede l'interazione di molti agenti autonomi e che (*ii*) esibisce comportamenti emergenti [27, 12]. In questo caso gli agenti sono gli utenti della rete. E i comportamenti emergenti riguardano gli effetti della diffusione delle informazioni e la formazione di *echo chamber* [22]. In questo quadro il problema affrontato riguarda certamente un sistema complesso.
Inoltre il problema analizzato in questo lavoro è un certamente un problema distribuito. Infatti la diffusione di *fake news* in un *social network* coinvolge una pluralità di utenti che operano autonomamente. E i modelli proposti perseguono l'obiettivo di analizzare questa diffusione al fine di comprenderla ed eventualmente limitarla. In questo quadro il presente lavoro rappresenta anche una possibile soluzione a un problema distribuito.

4.2 Agente

Un modello basato su agenti richiede naturalmente la presenza di agenti. In questo lavoro si riprende la nozione di agente proposta da Wooldridge e Jennings [41][4]. In particolare un agente è un *software* che presenta le seguenti caratteristiche:

- **autonomia**: l'agente può operare senza l'intervento umano e ha il controllo delle proprie azioni e del proprio stato;
- **abilità sociale**: l'agente può interagire con altri agenti tramite un linguaggio;

[4]Non esiste una nozione univoca di agente. La nozione proposta da Wooldridge e Jennings soddisfa le esigenze definitorie di questo lavoro. Tuttavia anche questa nozione presenta alcune limitazioni in quanto esclude gli agenti semplici che interagiscono senza un vero e proprio linguaggio.

- **reattività**: l'agente può agire in risposta a uno stimolo esterno;
- **proattività**: l'agente può anche agire di propria iniziativa.

Nei modelli proposti gli agenti sono rappresentati dagli utenti del *social network*.
L'architettura interna di questi agenti può essere definita seguendo la tassonomia di agenti proposta da Genesereth e Nilsson [13]. In particolare secondo questa tassonomia gli agenti proposti risultano essere agenti isteretici. Infatti gli agenti isteretici sono definiti attraverso la tupla:

$$\langle I, E, P, A, i_0, see, internal, do, action \rangle.$$

E gli elementi di questa tupla trovano conferma negli agenti proposti. Più precisamente:

- $\boldsymbol{I}$ rappresenta l'insieme degli stati interni dell'agente. In questo caso lo stato interno dell'agente è composto quantomeno da: (*i*) stato epidemiologico; (*ii*) ruolo; (*iii*) probabilità di infezione, vaccinazione, cura; (*iv*) numero di reclami ricevuti; (*v*) blocco dall'invio di messaggi; (*vi*) elenco dei messaggi pubblicati; e (*vii*) opinione.
- $\boldsymbol{E}$ rappresenta l'insieme degli stati dell'ambiente. In questo caso si tratta dell'insieme degli stati degli altri utenti del *social network*.
- $\boldsymbol{P}$ rappresenta la partizione di E percepibile dall'agente. In questo caso si tratta dello stato degli utenti "seguiti" dall'agente.
- $\boldsymbol{A}$ rappresenta l'insieme delle azioni dell'agente. In questo caso vi sono due tipologie di azioni. Una prima azione consiste nella pubblicazione di un messaggio che corrisponde al proprio stato epidemiologico. Una seconda azione consiste nell'eventuale presentazione di un reclamo nei confronti di un altro utente.
- $\boldsymbol{i_0}$ rappresenta lo stato iniziale dell'agente.
- $\boldsymbol{see} : E \to P$ è la funzione sensoriale dell'agente. In questo caso è la funzione che permette tipicamente all'agente di leggere i messaggi pubblicati dagli utenti "seguiti".

- ***internal*** : $I \times P \rightarrow I$ è la funzione che mappa lo stato interno e quanto osservato in un nuovo stato interno. In questo caso l'agente che legge i messaggi altrui aggiorna il proprio stato epidemiologico.

- ***action*** : $I \times P \rightarrow A$ è la funzione che individua l'azione da compiere sulla base dello stato interno e dell'osservazione dell'agente. In questo caso vi è (*i*) sia l'azione riguardante la scelta del messaggio da pubblicare (*ii*) sia l'azione riguardante la presentazione di un reclamo.

- ***do*** : $A \times E \rightarrow E$ è la funzione che aggiorna l'ambiente in base all'azione intrapresa dall'agente. In questo caso i reclami possono portare al blocco degli utenti.

Secondo la tassonomia proposta da Genesereth e Nilsson, questi agenti non sono né tropistici né a livello di conoscenza. In particolare non sono agenti tropistici perché possiedono quantomeno uno stato interno che influisce sulle azioni degli agenti. Inoltre non sono agenti a livello di conoscenza in quanto non presentano un *database* di fatti noti del mondo su cui fare inferenza: le uniche informazioni possedute riguardano lo stato interno dell'agente. Infine e in ogni caso si precisa che gli agenti proposti non perseguono alcun obiettivo o utilità.
Per tuziorismo gli agenti proposti possono anche essere qualificati come agenti riflessivi con stato interno secondo la tassonomia proposta da Russell e Norvig [33]. Questa qualifica è equivalente alla nozione di agenti isteretici coniata da Genesereth e Nilsson. L'architettura di questi agenti è mostrata in Figura 9.

4.3 Ambiente

Gli agenti operano all'interno di un ambiente. In questo caso l'ambiente è dato dalla rete sociale degli utenti. In particolare in questa rete i nodi sono rappresentati dagli utenti e gli archi dalla relazione unidirezionale di *follow* tipica dei *social network*.
Le caratteristiche dell'ambiente possono essere descritte seguendo la tassonomia proposta da Russell e Norvig in tema di ambienti [33]. Più precisamente l'ambiente proposto è:

- **parzialmente accessibile** in quanto gli utenti possono percepire solo il sottoinsieme di utenti che "seguono". E questo sottoinsieme ti-

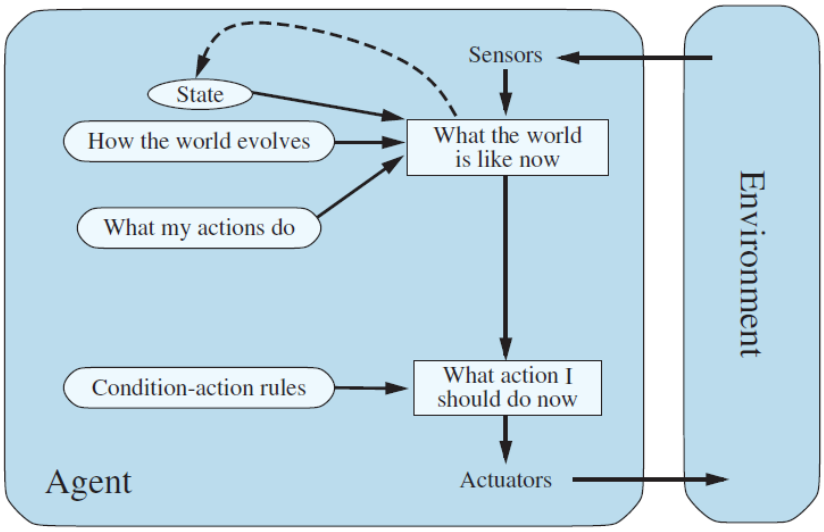

Figura 9: Architettura degli agenti riflessivi con stato interno. Questa rappresentazione è tratta da [33].

picamente non comprende tutti gli utenti della rete in quanto il caso dell'utente che "segue" tutto il resto della rete è estremamente improbabile;

- **statico** in quanto una volta terminata la creazione del grafo non vengono più aggiunti né rimossi nodi o archi. Pertanto l'ambiente non muta mentre l'agente delibera;
- **sequenziale** in quanto l'azione e l'aggiornamento dello stato interno dell'agente in un istante di tempo influiscono sulle azioni successive dell'agente[5]. Per esempio un utente neutrale che diventa vaccinato non potrà essere infettato successivamente e potrà presentare reclami;
- **stocastico** in quanto l'effetto sulla rete della pubblicazione di un messaggio dipende dalle diverse probabilità coinvolte[6];

[5]In questo caso si ritiene l'ambiente sequenziale in quanto si considera anche lo stato interno dell'agente. Se invece si dovessero considerare solo le azioni e le percezioni, allora l'ambiente potrebbe essere considerato episodico nella misura in cui in ogni episodio un utente legge i messaggi degli utenti "seguiti" e pubblica un proprio messaggio.

[6]In questo caso si preferisce il termine "stocastico" al termine "nondeterministico" in quanto le probabilità sono note e quantificate.

- **discreto** in quanto il numero di azioni e percezioni possibili è finito.

4.4 Interazione

Anche l'interazione tra gli agenti può essere classificata in base alle sue caratteristiche. In questo caso viene utilizzata la classificazione proposta da Jacques Ferber e ripresa da altri autorevoli Autori [3].
Nel modello proposto l'interazione tra gli agenti è indiretta e fondata spazialmente. Più precisamente:

- L'interazione è **indiretta** in quanto gli agenti non comunicano direttamente tra loro utilizzando un protocollo noto, ma pubblicano messaggi che vengono letti dai *follower*. Per questo motivo gli utenti che pubblicano messaggi non hanno bisogno di conoscere gli utenti che leggeranno i loro messaggi (*name uncoupling*). E gli utenti non devono neppure coesistere nello stesso istante di tempo (*time uncoupling*). Quando l'agente è attivo può leggere i messaggi pubblicati dagli utenti "seguiti" negli istanti di tempo precedenti.

- L'interazione è **fondata spazialmente**. Infatti nel caso di specie la rete sociale modellata ha intrinsecamente una struttura spaziale che influenza l'informazione che arriva all'utenza. In particolare l'ambiente degli agenti è rappresentato spazialmente come un grafo. E gli agenti hanno una precisa collocazione nel grafo e ne percepiscono una parte con il loro modello percettivo. In questo caso il contesto degli agenti è una porzione visibile del grafo: gli utenti "seguiti".

5 Validazione

I modelli proposti per simulare la diffusione di *fake news* in *social network* perseguono l'obiettivo di (*i*) consentire l'analisi di questo fenomeno e (*ii*) individuare possibili contromisure. Tuttavia questo obiettivo può essere raggiunto solo se i modelli proposti risultano corretti e adeguati allo scopo. A questo proposito allora occorre procedere alla validazione dei modelli.
Il procedimento di validazione utilizzato in questo lavoro riprende le fasi di validazione proposte da Klügl per i modelli basati su agenti [18]. In particolare queste fasi sono: (*i*) validazione *prima facie*; (*ii*) analisi di sensitività dei parametri; (*iii*) calibrazione dei parametri; e (*iv*) validazione *stricto sensu*.
L'ordine delle fasi seguito in questo lavoro è diverso da quello proposto da Klügl. In questo caso l'analisi di sensitività dei parametri è stata posticipata alla validazione in senso stretto. Il motivo è il seguente. L'analisi di sensitività mira a verificare l'impatto della variazione dei valori di un singolo parametro sulla *performance* del modello. In questo modo si può ridurre lo spazio di ricerca dei valori dei parametri e rimuovere i parametri ininfluenti. Tuttavia per valutare l'impatto del valore di un parametro occorre assumere che i valori degli altri parametri siano corretti o quantomeno plausibili. E questa assunzione risulta problematica nel caso di specie. Infatti i parametri sono molti e spesso rappresentano probabilità, per cui l'utilizzo di valori medi risulta poco appropriato. In questo quadro si è deciso di calibrare i parametri prima di svolgere l'analisi di sensitività. In questo caso l'obiettivo dell'analisi di sensitività è stato valutare l'impatto della variazione dei valori dei parametri rispetto ai dati reali utilizzati per la validazione in senso stretto.

5.1 Validazione *prima facie*

La prima fase del processo di validazione è la validazione *prima facie*. In questa fase si vuole verificare la plausibilità del modello sotto un profilo generale e strutturale. In particolare in questo caso si vuole verificare: (i) che le reti sociali generate siano effettivamente *scale-free*; e (ii) che l'omofilia della rete cresca quando il valore del relativo parametro della rete aumenta.

5.1.1 Validazione della natura *scale-free* delle reti

La validazione della natura *scale-free* delle reti generate è stata eseguita con due modalità.

Prima validazione. Anzitutto si è proceduto a calcolare la distanza tra la funzione di ripartizione di una distribuzione *power-law* e la funzione di ripartizione empirica del campione generato. Per calcolare questa distanza è stato eseguito il test non parametrico di Kolmogorov-Smirnov. Più precisamente:

1. Siano X la popolazione esaminata, F_X la distribuzione di questa popolazione e F la distribuzione cercata.

2. L'ipotesi nulla è:

   $$H_0 : F_X(t) = F(t) \qquad \text{per ogni } t \in R.$$

   In questo caso l'ipotesi nulla è che i gradi delle reti generate siano distribuiti secondo una distribuzione *power-law*.

3. L'ipotesi alternativa è:

   $$H_1 : F_X(t) \neq F(t) \qquad \text{per qualche } t \in R.$$

   Pertanto l'ipotesi alternativa è che almeno qualche grado delle reti generate non sia distribuito secondo una distribuzione *power-law*.

4. Sia $(X_1, X_2, ..., X_n)$ un campione di numerosità n estratto dalla popolazione X. In questo caso i campioni sono i gradi dei nodi.

5. La funzione di ripartizione empirica della popolazione X calcolata sulle variabili casuali ordinate in modo crescente è definita come:

   $$\hat{F}_{X,n}(t) = \frac{1}{n} \sum_{i=1}^{n} U_{(-\infty,t]}(X_i) \qquad \text{per ogni } t \in R,$$

   dove

   $$U_{(-\infty,t]}(X_i) = \begin{cases} 1 & \text{se } X_i \in (-\infty, t] \\ 0 & \text{altrimenti} \end{cases}$$

   Quindi per ogni valore si conta il numero di realizzazioni del campione che hanno valore inferiore o uguale a t. Poi si divide tale conteggio per la numerosità del campione.

6. La distanza tra la funzione di ripartizione empirica della popolazione X e la funzione di ripartizione cercata è calcolata come:

$$D_n = sup_{t \in R} |F(t) - \hat{F}_{X,n}(t)|.$$

Pertanto si considera l'estremo superiore della differenza in valore assoluto tra le due funzioni.

7. Dopodiché si calcola la regione critica

$$C = (d_{1-\alpha}, 1],$$

dove α è il livello di significatività del test e $d_{1-\alpha}$ è il quantile ottenibile dalle tavole di Kolmogorov-Smirnov. Se la distanza D_n non rientra nella regione critica, allora non si può rifiutare l'ipotesi nulla.

8. In questo quadro non si può rifiutare l'ipotesi nulla quando questa distanza D_n è piccola.

9. Inoltre si è proceduto a calcolare il *p-value* per il test. Se il *p-value* è maggiore del livello di significatività del test, allora anche per questo motivo l'ipotesi nulla non può essere rifiutata.

Ora, per eseguire questo test sono state create 30 reti di 50 nodi ciascuna con questi parametri: $\alpha = \frac{1}{3}$, $\beta = \frac{1}{3}$, $\gamma = \frac{1}{3}$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 1$, $h = 0.5$. I primi tre parametri sono stati posti pari a $\frac{1}{3}$ in modo da dare uguale importanza alle diverse modalità di creazione degli archi. I parametri δ_{in} e δ_{out} sono stati posti pari a 1 per prevenire divisioni per 0 come specificato in [2]. Il parametro dell'omofilia h è irrilevante per questo test. Pertanto è stato posto pari a 0.5.
Si è deciso si effettuare i calcoli su 30 reti al fine di ottenere risultati statisticamente più robusti. Infatti il modello è stocastico e i risultati potrebbero variare con reti diverse.
Per il calcolo della distanza D_n e del *p-value* sono state utilizzate le librerie `powerlaw` e `plfit` [1, 29]. I risultati ottenuti sono i seguenti:

- La distanza media tra le funzioni di ripartizione empiriche e la distribuzione *power-law* è di 0.13 con un intervallo di confidenza pari a $(0.12, 0.14)$ e con un livello di significatività pari a 0.05. In questo caso la regione critica è pari a:

$$C = (0.19, 1].$$

La distanza media non rientra in questa regione critica neppure considerando l'intervallo di confidenza. Pertanto non si può rifiutare l'ipotesi nulla che le reti generate siano *scale-free*.

- Inoltre il *p-value* medio calcolato è pari a 0.72 con un intervallo di confidenza pari a $(0.64, 0.80)$ e con un livello di significatività pari a 0.05. Certamente risulta che il *p-value* è maggiore di 0.05. E questo anche considerando l'intervallo di confidenza. In questo quadro anche per questo motivo non si può rifiutare l'ipotesi nulla che le reti generate siano *scale-free*.

Questi risultati sono riassunti nella Tabella 1.

Tabella 1: Risultati della prima validazione della natura *scale-free* delle reti generate

D_n	C	*p-value*	α	**risultato test**
0.13 ± 0.01	$(0.19, 1]$	0.72 ± 0.08	0.05	H_0 non rifiutata

Seconda validazione. Inoltre si è proceduto a visualizzare graficamente la distribuzione dei gradi dei nodi per verificare anche qualitativamente se questa distribuzione fosse a legge di potenza. In particolare per questa verifica è stata creata una rete di 1000 nodi con i medesimi parametri indicati per la prima validazione. I risultati ottenuti sono i seguenti.
In Figura 10 viene mostrato un grafico che presenta i gradi dei nodi sulle ascisse e la relativa probabilità sulle ordinate. La linea rossa tratteggiata rappresenta una distribuzione *power-law*. La linea continua blu rappresenta invece la distribuzione empirica generata dal modello. Si nota chiaramente che i gradi dei nodi seguono una distribuzione *power-law*.
In Figura 11 vengono mostrati due grafici analoghi a quello appena descritto. In questo caso però al posto del grado vengono rappresentati rispettivamente l'*out-degree* e l'*in-degree*. Anche queste distribuzioni seguono qualitativamente una *power-law*. Pertanto anche questi grafici confermano la bontà del modello proposto.
Infine in Figura 12 viene mostrata una rappresentazione del grafo generato. La dimensione dei nodi riflette il relativo *in-degree*. In questo grafo si nota

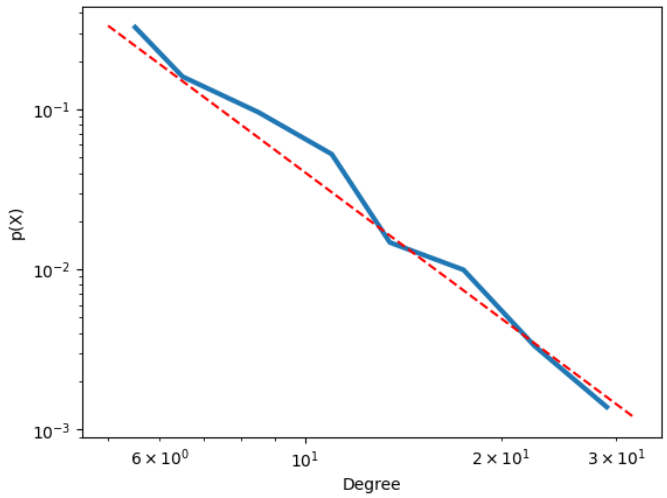

Figura 10: Grafico che rappresenta la distribuzione dei gradi dei nodi generati dal modello proposto. Il grafico è logaritmico sia sulle ascisse che sulle ordinate. La linea rossa tratteggiata rappresenta una distribuzione *power-law*. La linea continua blu rappresenta invece la distribuzione empirica generata dal modello.

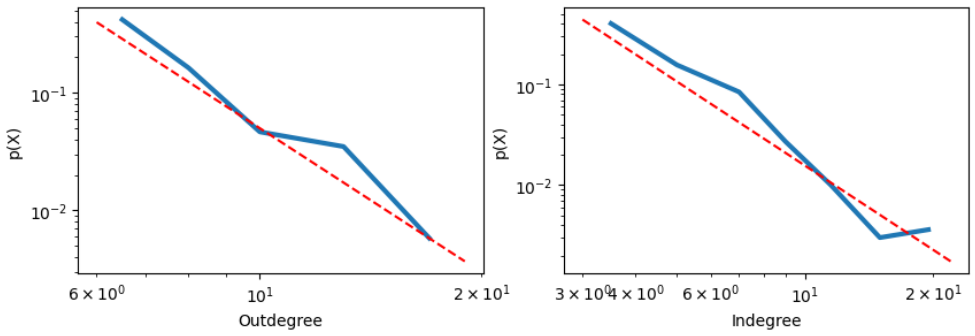

Figura 11: In questa Figura vengono mostrati due grafici. Il grafico a sinistra rappresenta la distribuzione dell'*out-degree* dei nodi generati dal modello proposto. Il grafico a destra rappresenta la distribuzione dell'*in-degree* dei nodi generati dal modello proposto. In entrambi i casi la linea tratteggiata rossa rappresenta una distribuzione *power-law* e la linea continua blu rappresenta la distribuzione empirica dei dati. Anche in questo caso i grafici sono logaritmici sia sulle ascisse che sulle ordinate.

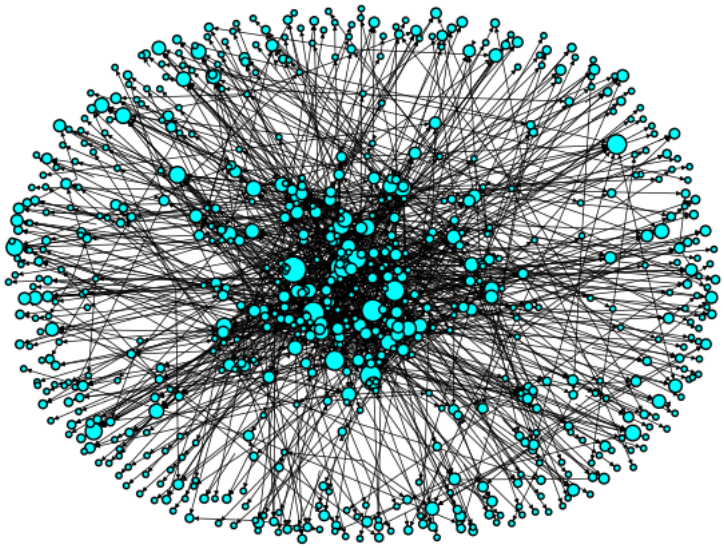

Figura 12: Grafo generato con il modello proposto. La dimensione dei nodi riflette il relativo *in-degree*.

come vi siano pochi nodi con elevato *in-degree* e molti nodi con basso *in-degree*. Questa caratteristica è tipica delle reti *scale-free*. In questo quadro anche questo grafo dimostra la correttezza del modello proposto.

5.1.2 Validazione dell'omofilia

La seconda verifica della validazione *prima facie* riguarda l'omofilia. In particolare ai Paragrafi 2.2.3 e 2.3 è stato introdotto il parametro h che regola l'omofilia della rete. Quanto più questo parametro è alto, tanto più cresce la probabilità di creare una connessione tra nodi con opinioni simili. Ora si vuole verificare questa proprietà.
La verifica è stata effettuata come segue:

- si sono considerati i seguenti valori per il parametro h: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 e 1.0;

- per ciascuno di questi valori sono state create 50 reti di 100 nodi ciascuna con i seguenti parametri: $\alpha = \frac{1}{3}$, $\beta = \frac{1}{3}$, $\gamma = \frac{1}{3}$, $\delta_{\text{in}} = 1$, $\delta_{\text{out}} = 1$;

- per ciascuna rete è stata calcolata la distanza media delle opinioni dei nodi come:

$$avg_{\text{op distance}} = \frac{1}{m} \sum_{e \in E} |op_x - op_y|,$$

  dove E è l'insieme degli archi della rete, m è il numero di archi della rete e op_x e op_y sono le opinioni dei nodi connessi;

- per ciascun valore del parametro h è stata calcolata la media delle distanze medie delle relative reti;

- si è verificato che queste ultime medie (con i relativi intervalli di confidenza) decrescono al crescere del valore h. Questo significa che le distanze tra le opinioni degli utenti decrescono al crescere di h. E ciò conferma l'ipotesi che si voleva validare.

Anche in questo caso i calcoli sono stati effettuati su più reti in modo da ottenere risultati statisticamente più robusti.
I risultati di questa verifica sono mostrati nella Figura 13. La distanza media tra le opinioni sembra crescere leggermente oltre il valore 0.9 per il parametro h. Tuttavia questa crescita è contenuta. E il limite inferiore dell'intervallo di confidenza per $h = 1.0$ è comunque più basso rispetto alla distanza media tra le opinioni per $h = 0.9$. Pertanto non risulta invalidata l'ipotesi proposta.
Inoltre si nota come la distanza media tra le opinioni sia sempre compresa fra 0.28 e 0.38. Questo è dovuto alla natura *scale-free* della rete, per cui la connessione tra i nodi è fortemente influenzata dal relativo grado. Quindi nodi con opinioni simili possono essere portati a connettersi anche se il valore di h è basso, e viceversa.

5.2 *Dataset*

Sin qui è stata descritta la validazione *prima facie*. Le successive fasi della validazione richiedono il confronto con dati reali. Per questo motivo occorrono dei *dataset*. A questo proposito sono stati utilizzati i *dataset* `palin` e `obama` descritti in [31]. Il *dataset* `palin` è stato utilizzato per la validazione anche in [35, 12].

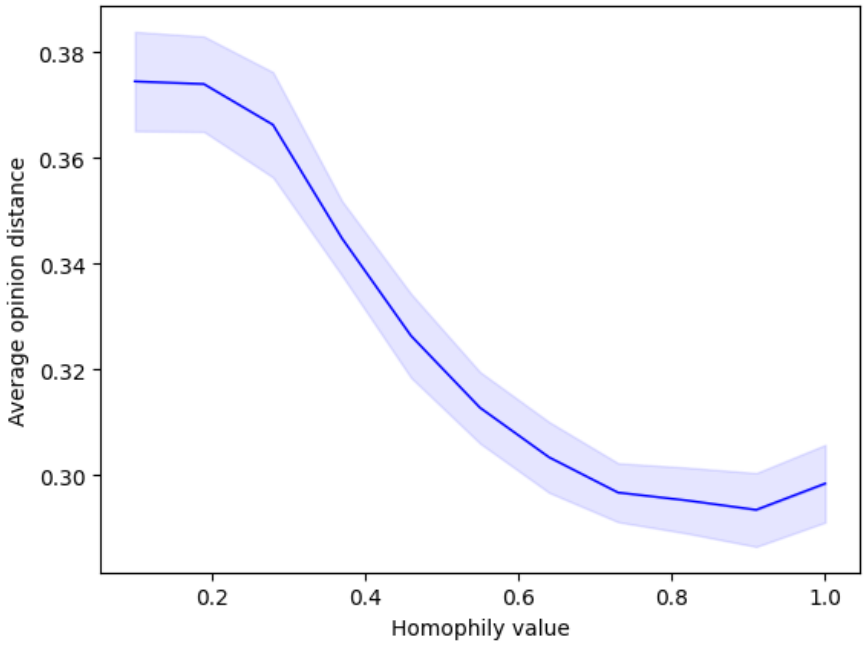

Figura 13: Grafico che mostra come al crescere del valore del parametro di omofilia h diminuiscano le distanze tra le opinioni degli utenti della rete.

Questi *dataset* sono stati resi disponibili al seguente indirizzo: `http://dit.upm.es/~eserrano/BigMarket/ESWA2015/`. Si tratta di *dataset* contenenti informazioni di *tweet* estratti da Twitter. Queste informazioni comprendono: (*i*) la data di pubblicazione; (*ii*) un identificativo anonimo dell'utente; e (*iii*) un'etichetta che descrive se il *tweet* diffonde una *fake news*, la confuta o è neutrale.
Questi *dataset* non contengono alcuna informazione sui *follower* degli utenti. Per questo motivo si è proceduto a scaricare i singoli *tweet* utilizzando l'API di Twitter [37] e la libreria `Tweepy` [36]. Per individuare con più precisione i *tweet* si è fatto ricorso ai dettagli riportati in `https://github.com/vahedq/rumors`. Tra questi dettagli sono infatti riportati anche gli identificativi dei *tweet*.
Una volta scaricati i *tweet* è stato possibile determinarne gli autori e recuperare il relativo numero di *follower*. Quest'ultimo dato è stato utilizzato per individuare gli *influencer*. A questo proposito non esiste una soglia univoca di *follower* oltre la quale si è considerati *influencer*. In questo caso si è deciso di considerare come *influencer* gli utenti con più di 50000 *follower*. Questa soglia si è rilevata idonea a individuare un numero di *influencer* ridotto, ma sufficiente a esercitare influenza sulla rete.
Infine si sottolineano i limiti dei *dataset* individuati. Anzitutto i *dataset* non contengono tutto il grafo di Twitter, ma singoli *tweet*. Inoltre i *dataset* sono circoscritti temporalmente a pochi mesi. Ancora, per questi mesi non vengono riportati tutti i *tweet* pubblicati, ma solo una selezione. Ancora, i *dataset* non contengono neppure tutti i *tweet* degli utenti menzionati, ma solo alcuni campioni. Infine non è stato possibile ricostruire le relazioni degli utenti a causa dei limiti imposti dall'API di Twitter.

5.2.1 *Dataset* `palin`

Il *dataset* `palin` è composto da 4423 *tweet* suddivisi in questo modo: (*i*) 1709 *tweet* a favore della *fake news*; (*ii*) 1895 *tweet* contrari alla *fake news*; e (*iii*) 819 *tweet* neutrali. Nella Tabella 2 viene riportato un riepilogo.
Si noti che questa suddivisione non specifica quanti siano gli utenti che sostengono o combattono la *fake news*. Infatti alcuni utenti pubblicano più messaggi. E alcuni utenti cambiano orientamento nel tempo. Il numero di utenti del *dataset* è pari a 3181.
In questo quadro si è proceduto a calcolare il numero di sostenitori della *fake*

Tabella 2: Descrizione generale dei *tweet* del *dataset* `palin`

Dataset	**favorevoli**	**contrari**	**neutrali**	**totali**
`palin`	1709	1895	819	4423

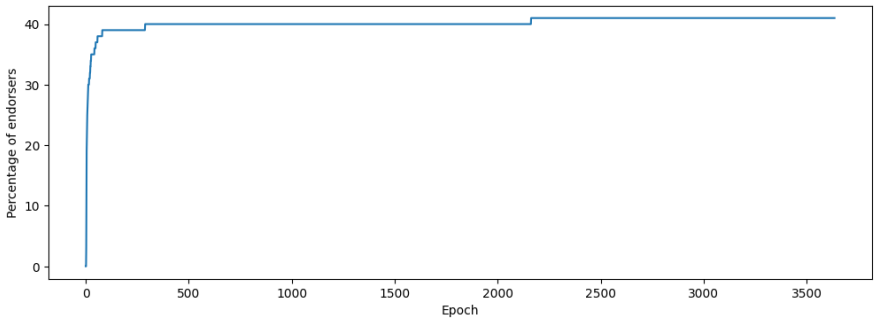

Figura 14: Percentuale di sostenitori della *fake news* per epoca relativamente al *dataset* `palin`.

news per ogni istante di tempo[7]. In questo caso ogni istante di tempo è pari a 1 ora. Questo istante di tempo è sufficientemente grande da consentire l'individuazione di *tweet* di diversi utenti. Ed è sufficientemente piccolo da limitare il mutamento di orientamento di un utente nello stesso istante di tempo. Inoltre questa unità temporale è stata adottata anche in [35].
Il calcolo dei sostenitori della *fake news* è avvenuto come segue. Anzitutto sono stati individuati il primo e l'ultimo istante di tempo del *dataset*. Poi sono stati determinati gli istanti di tempo totali del *dataset*, che sono pari a 3636. Dopodiché sono stati individuati i *tweet* per ogni istante di tempo. A questo punto è stato calcolato progressivamente il numero cumulativo di sostenitori della *fake news* per ogni istante di tempo. I sostenitori che hanno successivamente pubblicato messaggi contrari alla *fake news* sono stati rimossi dai sostenitori.
Il risultato di questo calcolo è mostrato nella Figura 14.
Nella prima epoca la percentuale di sostenitori della *fake news* rispetto al totale degli utenti è del 3%. Alla quinta epoca la percentuale sale già al 25%. All'ultima epoca la percentuale è del 41%.

[7]In questo lavoro si useranno i termini "istante di tempo" ed "epoca" come sinonimi.

5.2.2 *Dataset* `obama`

Il *dataset* `obama` è composto da 4337 *tweet* suddivisi in questo modo: (*i*) 854 *tweet* a favore della *fake news*; (*ii*) 489 *tweet* contrari alla *fake news*; e (*iii*) 2994 *tweet* neutrali. Nella Tabella 3 viene riportato un riepilogo.

Tabella 3: Descrizione generale dei *tweet* del *dataset* `obama`

Dataset	**favorevoli**	**contrari**	**neutrali**	**totali**
`obama`	854	489	2994	4337

Il numero di utenti del *dataset* è pari a 2678. La definizione degli istanti di tempo è la stessa usata per il *dataset* `palin`. In questo caso il numero di istanti di tempo è 4747. Parimenti anche il procedimento di calcolo dei sostenitori della *fake news* è il medesimo adottato per il *dataset* `palin`. Il risultato di questo calcolo è mostrato nella Figura 15.
Nella prima epoca la percentuale di sostenitori della *fake news* rispetto al totale degli utenti è dell'1%. Alla quinta epoca la percentuale sale al 3%. All'ultima epoca la percentuale è del 20%.

5.3 Calibrazione dei parametri

Per validare i modelli proposti occorre calibrarne i parametri. Il *dataset* utilizzato per la calibrazione dei parametri è il *dataset* `palin`. Il procedimento

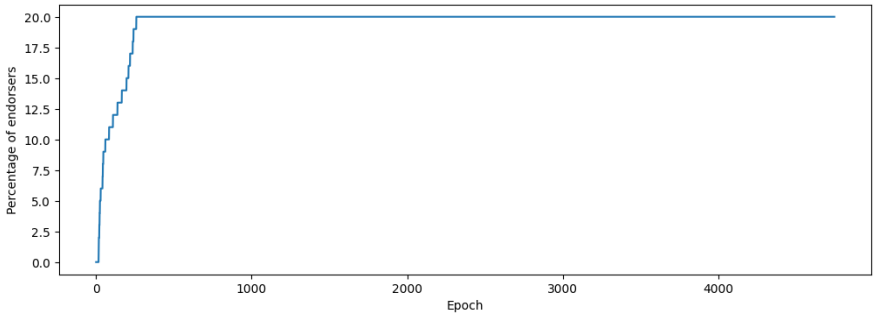

Figura 15: Percentuale di sostenitori della *fake news* per epoca relativamente al *dataset* `obama`.

di calibrazione dei parametri verrà illustrato come segue. In primo luogo verranno distinti i parametri da calibrare da quelli per cui si è stabilito da subito un valore. In secondo luogo verrà descritta la calibrazione di alcuni parametri con *grid search*. In terzo luogo verrà descritta la calibrazione degli altri parametri con l'ottimizzazione bayesiana. Infine verranno mostrati i risultati ottenuti.

5.3.1 Parametri predeterminati

I parametri predeterminati riguardano principalmente le informazioni estratte dal *dataset*. I parametri predeterminati sono riassunti nella Tabella 4.

Tabella 4: Parametri predeterminati per la calibrazione sul *dataset* `palin`

Dataset	**nodi**	$\delta_{\mathbf{in}}$	$\delta_{\mathbf{out}}$	**epoche**	***infl.***	***bot***	***e.f.c.***	**infetti iniziali**
`palin`	3181	1	1	3634	25	64	64	31

Il numero di nodi è pari al numero di utenti del *dataset*. I parametri δ_{in} e δ_{out} sono pari a 1 per le ragioni anticipate al Paragrafo 5.1 e menzionate in [2]. Il numero di epoche è pari a 3634. Dalle 3636 epoche del *dataset* sono state eliminate le prime due epoche in quanto il numero di sostenitori della *fake news* era nullo. Il numero di utenti con più di 50000 *follower* è 25: questi sono gli *influencer*. Il numero di *bot* è pari a 64, che corrisponde al 2% degli utenti del *dataset*. Questa percentuale di *bot* è ritenuta corretta da [21]. Si è ritenuto di quantificare gli *eternal fact-checker* (abbreviati *e.f.c.*) in numero equivalente ai *bot* per bilanciare infezione e vaccinazione[8]. Il numero di infetti iniziali (sostenitori della *fake news*) è pari a 31. In questo numero non figurano i *bot*. La somma degli infetti iniziali e dei *bot* è pari al 3% degli utenti del *dataset*. Questa percentuale è coerente con l'analisi del *dataset* riportata *supra*.

[8]La presenza di *bot* sui *social network* è un fatto noto. Tuttavia non si hanno dati certi sul loro numero. E non si può dire *a priori* se siano favorevoli o contrari a una notizia. Pertanto in fase di calibrazione si è ritenuto corretto bilanciare *bot* ed *eternal fact-checkers*. Inoltre i *dataset* utilizzati non menzionano la presenza di *bot* ed *eternal fact-checkers*. In questo quadro si è preferito bilanciare queste categorie in modo da limitare la presenza di *bias* nella fase di calibrazione.

5.3.2 Parametri oggetto di calibrazione

I parametri oggetto di calibrazione sono invece i seguenti: (*i*) i parametri della rete α, β e γ che regolano le modalità di connessione dei nodi; (*ii*) il parametro h che regola l'omofilia; (*iii*) i parametri del modello di diffusione che riguardano le probabilità di infezione, vaccinazione e cura; (*iv*) il parametro relativo alla maggiore probabilità di infezione, vaccinazione e cura per i messaggi degli *influencer*; e (*v*) il parametro relativo alla soglia per la propagazione dello stato epidemiologico nelle *echo chamber*.
Tra i parametri oggetto di calibrazione non è stata inserita la probabilità di presentare reclami contro gli utenti infetti. Il motivo è che il *dataset* non include informazioni su possibili blocchi degli utenti. Pertanto si è preferito non considerare il blocco degli utenti nella validazione. Questa conclusione è stata raggiunta anche in [12].

5.3.3 Calibrazione con approccio *grid search*

Una prima parte di parametri è stata calibrata con un approccio *grid search*. Questo approccio prevede di individuare un numero ridotto di possibili valori dei parametri e di valutare la *performance* del modello per tutte le combinazioni di questi valori. I valori adottati sono quelli che appartengono alla combinazione con la *performance* migliore.
I parametri calibrati in questo modo sono i parametri che riguardano la topologia della rete: α, β, γ e h. Per questi parametri si è adottato l'approccio *grid search* in quanto è stato possibile individuare *a priori* alcune combinazioni di valori plausibili e coerenti con la struttura della rete. In particolare:

- Per i parametri α, β e γ sono state valutate le combinazioni per cui la somma fosse pari a 1. In particolare i valori considerati per α e β sono stati: 0.0, 0.25, 0.33, 0.5, 0.75 e 1.0. Il valore di γ è stato inferito.

- Per il parametro h sono stati considerati i valori 0.25, 0.50 e 0.75. In questo modo è stato possibile valutare l'effetto dell'omofilia sulla *performance* del modello.

5.3.4 Calibrazione con ottimizzazione bayesiana

Gli altri parametri sono stati calibrati attraverso un procedimento di ottimizzazione bayesiana. In particolare l'ottimizzazione bayesiana utilizza una

funzione di valutazione per guidare la ricerca dei valori verso regioni più promettenti nello spazio dei valori dei parametri. É stato scelto questo approccio per i parametri del modello di diffusione in quanto il *grid search* per tutti questi parametri sarebbe stato (*i*) computazionalmente troppo oneroso e (*ii*) poco appropriato considerate le numerose combinazioni plausibili. L'ottimizzazione bayesiana invece consente di calibrare i parametri in modo più efficiente per spazi di ricerca ampi e continui.
In questo caso la funzione di valutazione utilizzata è stata il *Root Mean Squared Error* (in seguito: RMSE) tra la percentuale reale di infetti per epoca e la percentuale predetta di infetti per epoca. Pertanto:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \ ,$$

dove n è il numero di epoche (3634), Y_i è il vettore delle percentuali reali di infetti per epoca e $\hat{Y}_i$ è il vettore delle percentuali predette di infetti per epoca.
Si è scelto di calcolare l'RMSE solo sulla percentuale di infetti per due motivi. In primo luogo perché l'obiettivo del presente lavoro è individuare contromisure alla diffusione delle *fake news* e quindi alla propagazione dell'"infezione". Pertanto il numero di utenti neutrali e contrari alla *fake news* è meno rilevante. In secondo luogo perché questa misura di validazione è stata utilizzata in letteratura per lo stesso scopo [12].
Lo spazio di ricerca dei valori per la probabilità di infezione è stato circoscritto nell'intervallo $[0.25, 0.50]$, mentre quello per le probabilità di vaccinazione e cura è stato circoscritto nell'intervallo $[0.00, 0.25]$. I motivi sono due. In primo luogo i modelli calibrati nella letteratura riportano valori di questi parametri più piccoli di 0.50 [12]. In secondo luogo la letteratura scientifica specifica che le *fake news* si diffondono più rapidamente e più ampiamente rispetto alle notizie vere [39]. Pertanto si è ritenuto di individuare probabilità di vaccinazione e cura più basse rispetto alla probabilità di infezione.
Per la probabilità degli *influencer* e per la soglia delle *echo chamber* è stato invece utilizzato il seguente spazio di ricerca: $[0.0, 1.0]$. In questo modo lo spazio di ricerca comprende tutti i valori assumibili dai parametri.
In questo quadro la calibrazione dei parametri è stata operata come segue:

1. per ogni combinazione iniziale di parametri (comprensivi anche dei

parametri predeterminati) sono state create 2 reti[9];

2. per ogni rete sono state simulate le 3634 epoche;
3. in questo modo si è ottenuto un vettore di percentuali di utenti infetti per ogni epoca;
4. si è calcolata la media di questi vettori per le 2 reti create;
5. questo vettore di media è stato usato per calcolare l'RMSE rispetto alle percentuali reali di infetti per epoca;
6. a questo punto sono stati modificati i soli parametri oggetto di ottimizzazione bayesiana, sono state ripetute le simulazioni ed è stato ricalcolato l'RMSE;
7. questo processo di ottimizzazione bayesiana è stato ripetuto per 8 iterazioni;
8. infine per ogni combinazione iniziale di parametri sono stati memorizzati i relativi parametri (eventualmente modificati con l'ottimizzazione bayesiana) e l'RMSE;
9. i parametri migliori sono quelli che hanno portato a ottenere l'RMSE più basso.

Anche in questo caso sono stati effettuati calcoli con più reti per ottenere risultati statisticamente più robusti.
Per l'ottimizzazione bayesiana è stata utilizzata la libreria `scikit-learn` [28].

5.3.5 Risultati

I risultati della calibrazione dei parametri sono mostrati nella Tabella 5 riguardante i migliori parametri ottenuti con un valore di h rispettivamente pari a 0.25, 0.50 e 0.75.

[9]Per limiti computazionali si è deciso di limitare a 2 il numero di reti su cui calcolare l'RMSE per ogni combinazione di parametri.

Tabella 5: Parametri calibrati sul *dataset* `palin`

Dataset	α	β	γ	h	$P_{\mathbf{inf}}$	$P_{\mathbf{vacc}}$	$P_{\mathbf{cure}}$	$P_{\mathbf{influencer}}$	$P_{\mathbf{echo}}$	**RMSE**
`palin`	0.50	0.33	0.17	0.25	0.44	0.06	0.03	0.65	0.31	1.29
`palin`	0.50	0.33	0.17	0.50	0.36	0.06	0.22	0.10	0.62	1.41
`palin`	0.25	0.50	0.25	0.75	0.36	0.02	0.18	0.76	0.12	2.81

5.4 Validazione *stricto sensu*

Una volta calibrati i parametri è stato possibile procedere alla validazione *stricto sensu*. L'obiettivo di questa fase è verificare se i modelli proposti con i parametri calibrati sono in grado di simulare correttamente la diffusione delle *fake news* rispetto a un *dataset* diverso da quello usato per la calibrazione dei parametri.

***Dataset* e parametri.** Per questa fase della validazione è stato utilizzato il *dataset* `obama`. Anche in questo caso i parametri non calibrati sono stati estratti dal *dataset*. Questi parametri sono riassunti nella Tabella 6.

Tabella 6: Parametri predeterminati per la validazione sul *dataset* `obama`

Dataset	**nodi**	$\delta_{\mathbf{in}}$	$\delta_{\mathbf{out}}$	**epoche**	***infl.***	***bot***	***e.f.c.***	**infetti iniziali**
`obama`	2678	1	1	4730	29	27	27	27 (con *bot*)

Il numero di epoche è pari a 4730. La riduzione rispetto alle 4747 epoche menzionate al Paragrafo 5.2.2 è dovuta al fatto che per le prime epoche la percentuale di infetti era nulla. Il numero di *bot* e di *eternal fact-checker* è pari a 27. In questo caso il numero di *bot* rappresenta (non più il 2%, ma) solo l'1% degli utenti del *dataset*. Questa scelta è stata fatta in quanto la percentuale di infetti per la prima epoca è pari all'1% degli utenti. Pertanto si è ritenuto di non eccedere questo dato al fine di rispettare le caratteristiche del *dataset*.

Valutazione quantitativa. Il procedimento per la validazione è stato analogo a quello utilizzato per la calibrazione dei parametri. In questo caso però l'RMSE è stato calcolato come media (non più su 2 reti, ma) su 30 reti. La validazione *stricto sensu* è stata condotta con le migliori combinazioni di parametri ottenute per i diversi valori di omofilia. I RMSE ottenuti sono riportati nella Tabella 7.

Tabella 7: Risultati della validazione *stricto sensu* sul *dataset* `obama`

Dataset	h	**RMSE**
`obama`	0.25	14.10 ± 7.74
`obama`	0.50	$\mathbf{7.74 \pm 4.37}$
`obama`	0.75	18.18 ± 5.64

Il risultato migliore in termini di RMSE è stato ottenuto con i parametri calibrati con un valore di h pari a 0.50. Ora, questo valore non elevato del parametro di omofilia non sembra coerente con la tesi secondo cui in un *social network* le connessioni tra utenti simili sono più frequenti delle connessioni tra utenti dissimili [23]. Tuttavia questa apparente incongruenza è giustificata dai difetti del *dataset*, che non riflette in modo esatto un vero *social network*.
In Figura 16 viene riportato il grafico relativo alle percentuali di infetti per epoca ottenuto con il valore di h pari a 0.50. In questa Figura si nota come la percentuale predetta di infetti per le ultime epoche si discosti poco dalla quella reale. La differenza è di circa il 5%. E nella varianza dei valori predetti sono comunque compresi i valori reali. Tuttavia il modello proposto non replica correttamente la progressione delle percentuali di infetti. Infatti con il modello proposto la percentuale di infetti cresce repentinamente nelle prime 50 epoche per poi stabilizzarsi. Invece la percentuale reale di infetti cresce più lentamente e si arresta solo dopo circa 250 epoche. Bisogna però ricordare ancora una volta che il *dataset* utilizzato non contiene tutti i nodi e tutti i messaggi della rete sociale di Twitter. In questo quadro allora si ritiene che il modello proposto sia comunque adeguato alla simulazione della diffusione di *fake news* quantomeno perché le percentuali finali di utenti infetti risultano prossime a quelle dei *dataset* utilizzati.
Nelle successive analisi e simulazioni verranno utilizzati i parametri calibrati con il valore di h pari a 0.50 in quanto hanno portato a ottenere un basso errore tanto sul *dataset* `palin` (nella calibrazione dei parametri) quanto sul *dataset* `obama` (nella validazione *stricto sensu*)[10].

[10]Si precisa che in questa fase di validazione non si vuole verificare l'efficacia del procedimento di calibrazione dei parametri, ma individuare parametri che permettano di simulare correttamente la diffusione di *fake news* in *social network*. I parametri migliori per queste simulazioni sono stati ottenuti con una calibrazione che ha considerato un valore di h pari

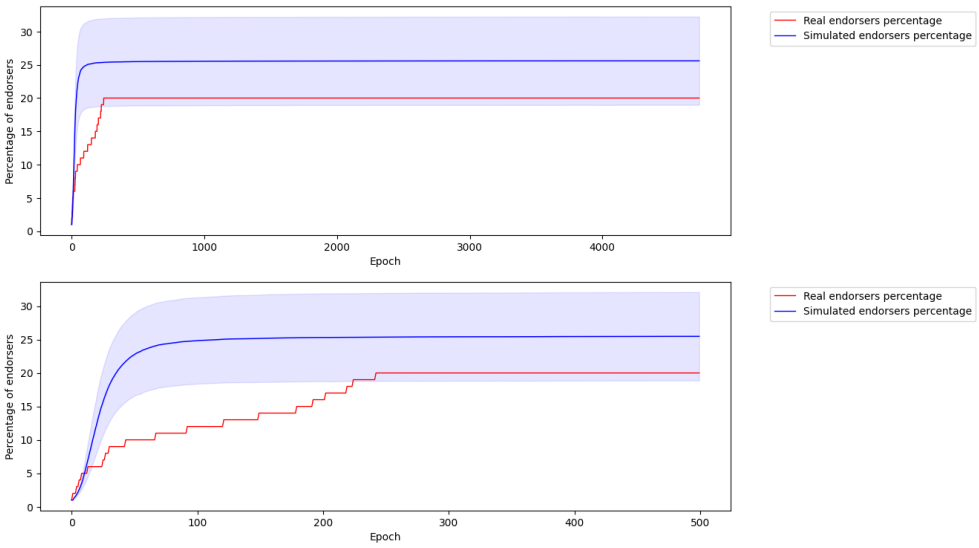

Figura 16: Grafici relativi alle percentuali medie di infetti per epoca per la validazione *strico sensu* con valore di h pari a 0.50 sul *dataset* `obama`. Il grafico superiore considera tutte le epoche. Il grafico inferiore considera solo le prime 500 epoche.

Valutazione qualitativa. In Figura 17 viene mostrata l'evoluzione degli stati epidemiologici degli utenti in uno dei grafi generati. Il grafo superiore mostra gli stati alla prima epoca. Il grafo inferiore mostra gli stati all'ultima epoca. In particolare si nota come anche all'ultima epoca molti nodi siano rimasti neutrali. Questo significa che questi nodi non sono stati raggiunti dalle informazioni diffuse.

5.5 Analisi di sensitività dei parametri

Nel Paragrafo precedente è stata descritta la validazione *stricto sensu*. In particolare si è verificato che i modelli proposti simulano correttamente la diffusione delle *fake news* in una rete sociale. In questo Paragrafo invece verrà analizzato l'impatto dei parametri calibrati con l'ottimizzazione bayesiana sulla *performance* del modello. In particolare questo impatto verrà valutato attraverso un'analisi di sensitività. Questa analisi è stata condotta solo per i parametri calibrati con ottimizzazione bayesiana (*i*) per limiti computazionali e (*ii*) in quanto più attinenti alla diffusione di *fake news* (rispetto alla topologia delle reti). Il procedimento di valutazione delle *performance* è stato lo stesso utilizzato per la validazione *stricto sensu*. Anche in questo caso il *dataset* di riferimento per valutare la bontà dei risultati è stato il *dataset* `obama`.

5.5.1 Probabilità di infezione

La prima analisi di sensitività condotta riguarda la probabilità di infezione. I valori considerati per l'analisi sono stati: 0.00, 0.25, 0.50, 0.75 e 1.00. I RMSE ottenuti sono mostrati nella Tabella 8. Nella Figura 18 invece vengono rappresentate le percentuali di infetti per epoca registrate per ogni valore analizzato.
In particolare si possono svolgere le seguenti osservazioni:

- **Confronto tra i valori del parametro**. Naturalmente la percentuale di infetti per epoca cresce all'aumentare della probabilità di infezione. Tuttavia anche quando la probabilità di infezione è massima, la percentuale di infetti non supera il 50%. Questo si verifica in quanto una parte rilevante dei nodi della rete non viene raggiunta dalle informazioni.

a 0.50. Per questo motivo si è scelto di utilizzare nel prosieguo questi parametri anche se l'RMSE più basso nella calibrazione è stato ottenuto con un valore di h pari a 0.25.

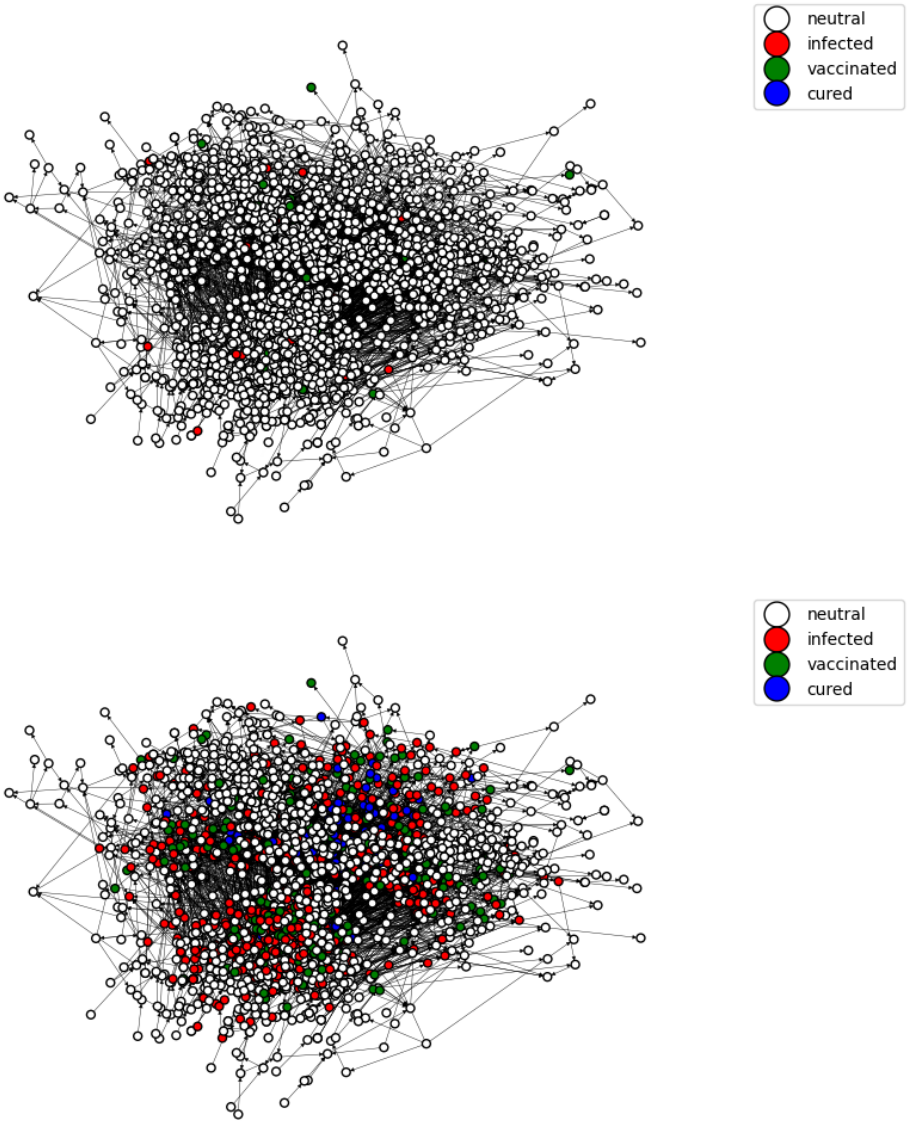

Figura 17: Grafi relativi agli stati epidemiologici degli utenti in una rete generata per la validazione *stricto sensu* con valore di h pari a 0.50 sul *dataset* `obama`. Il grafo superiore rappresenta lo stato degli utenti alla prima epoca. Il grafo inferiore rappresenta lo stato degli utenti all'ultima epoca.

Tabella 8: Analisi di sensitività per la probabilità di infezione sul *dataset* `obama`

$P_{\mathbf{inf}}$	**RMSE**
0.00	18.70 ± 3.55
0.25	5.21 ± 3.82
0.50	14.89 ± 6.00
0.75	23.64 ± 3.84
1.00	27.86 ± 4.14

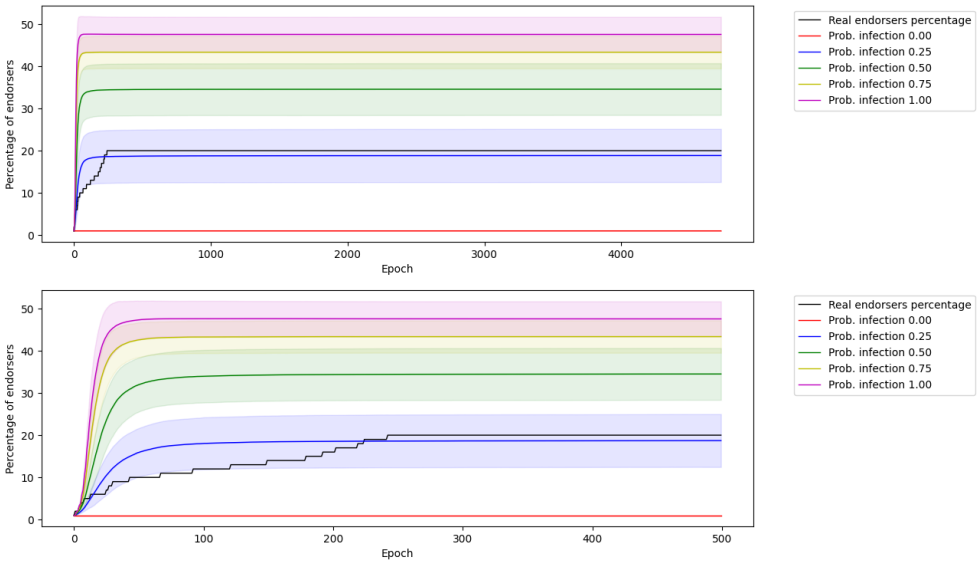

Figura 18: Grafici relativi alle percentuali medie di infetti per epoca ottenute per l'analisi di sensitività della probabilità di infezione. Il grafico superiore considera tutte le epoche. Il grafico inferiore considera solo le prime 500 epoche.

- **Confronto rispetto al *dataset***. La probabilità di infezione che consente di replicare meglio le percentuali di infetti per epoca del *dataset* è compresa fra 0.25 e 0.50. Si registra comunque un'alta variabilità dei risultati ottenuti per questi valori di probabilità di infezione.

5.5.2 Probabilità di vaccinazione

La seconda analisi di sensitività condotta riguarda la probabilità di vaccinazione. I valori considerati per l'analisi sono stati: 0.00, 0.25, 0.50, 0.75 e 1.00. I RMSE ottenuti sono mostrati nella Tabella 9. Nella Figura 19 invece vengono rappresentate le percentuali di infetti per epoca registrate per ogni valore analizzato.

Tabella 9: Analisi di sensitività per la probabilità di vaccinazione sul *dataset* `obama`

$P_{\mathbf{vacc}}$	**RMSE**
0.00	33.63 ± 3.52
0.25	16.42 ± 1.26
0.50	18.26 ± 0.22
0.75	18.55 ± 0.08
1.00	18.65 ± 0.05

In particolare si possono svolgere le seguenti osservazioni:

- **Confronto tra i valori del parametro**. La percentuale di infetti per epoca decresce all'aumentare della probabilità di vaccinazione. Tuttavia si nota come anche una bassa probabilità di vaccinazione dello 0.25 riduca la percentuale di infetti del 50% rispetto alla probabilità di vaccinazione nulla. La percentuale massima di infetti è di poco superiore 50%. Non si notano differenze significative per probabilità di vaccinazione superiori allo 0.25.

- **Confronto rispetto al *dataset***. L'analisi di sensitività conferma che un valore adeguato per la probabilità di vaccinazione è compreso fra 0.00 e 0.25. Quindi viene confermata la correttezza del valore di 0.10 ottenuto con la calibrazione del parametro. Tanto i valori inferiori

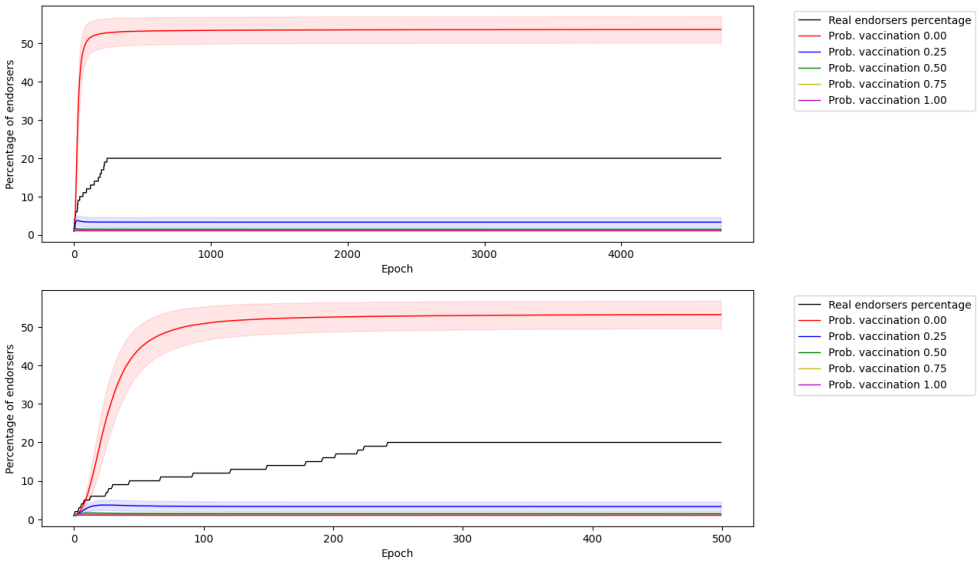

Figura 19: Grafici relativi alle percentuali medie di infetti per epoca ottenute per l'analisi di sensitività della probabilità di vaccinazione. Il grafico superiore considera tutte le epoche. Il grafico inferiore considera solo le prime 500 epoche.

quanto quelli superiori avrebbero portato a risultati molto differenti da quelli corretti.

5.5.3 Probabilità di cura

La terza analisi di sensitività condotta riguarda la probabilità di cura. I valori considerati per l'analisi sono stati: 0.00, 0.25, 0.50, 0.75 e 1.00. I RMSE ottenuti sono mostrati nella Tabella 10. Nella Figura 20 invece vengono rappresentate le percentuali di infetti per epoca registrate per ogni valore analizzato.

Tabella 10: Analisi di sensitività per la probabilità di cura sul *dataset* `obama`

$P_{\mathbf{cure}}$	**RMSE**
0.00	12.25 ± 5.70
0.25	8.00 ± 4.16
0.50	8.15 ± 4.55
0.75	7.02 ± 4.04
1.00	6.45 ± 4.50

In particolare si possono svolgere le seguenti osservazioni:

- **Confronto tra i valori del parametro**. Le percentuali di infetti più alte si registrano quando la probabilità di cura è nulla. In questo caso si arriva al 30% di infetti nelle ultime epoche. Tuttavia già con una probabilità di cura dello 0.25 si ottiene una riduzione degli infetti di circa 5 punti percentuali. Non si notano invece differenze significative con ulteriori incrementi della probabilità di cura. In ogni caso le percentuali di infetti più basse si ottengono con la massima probabilità di cura.

- **Confronto rispetto al *dataset***. Il valore massimo della probabilità di cura porta a ottenere mediamente i risultati più vicini a quelli reali. Tuttavia anche con probabilità di cura inferiori (ma non nulle) i risultati non si discostano molto da quelli reali. La distanza più elevata dai dati reali si ottiene con una probabilità di cura nulla.

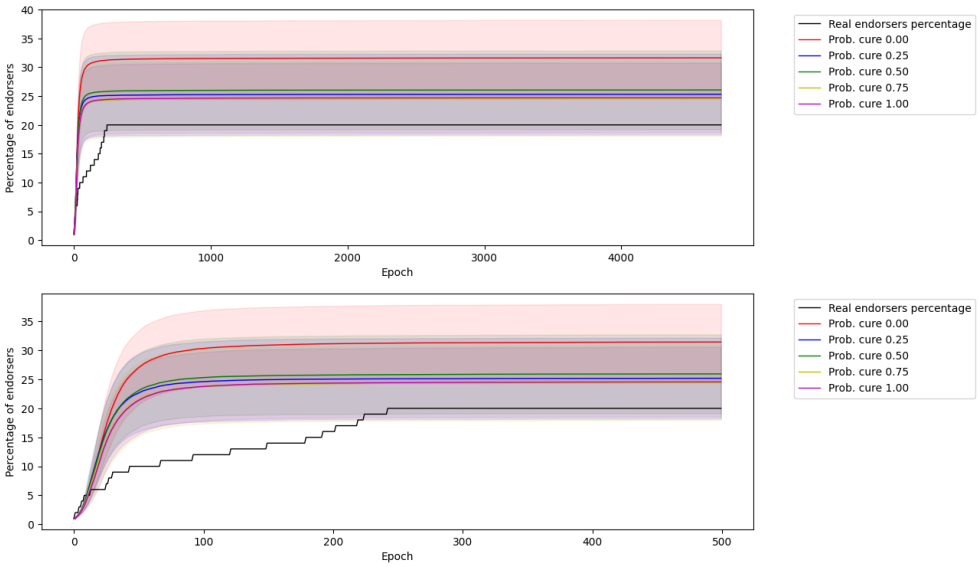

Figura 20: Grafici relativi alle percentuali medie di infetti per epoca ottenute per l'analisi di sensitività della probabilità di cura. Il grafico superiore considera tutte le epoche. Il grafico inferiore considera solo le prime 500 epoche.

5.5.4 Probabilità degli *influencer*

La quarta analisi di sensitività condotta riguarda la maggiore probabilità degli *influencer* di infettare, vaccinare e curare. I valori considerati per l'analisi sono stati: 0.00, 0.25, 0.50, 0.75 e 1.00. I RMSE ottenuti sono mostrati nella Tabella 11. Nella Figura 21 invece vengono rappresentate le percentuali di infetti per epoca registrate per ogni valore analizzato.
In particolare si possono svolgere le seguenti osservazioni:

- **Confronto tra i valori del parametro**. La variazione della probabilità di infezione, vaccinazione e cura degli *influencer* sembra non influire sulla percentuale di infetti per epoca. Infatti la variazione della probabilità non comporta differenze significative.

- **Confronto rispetto al *dataset***. Tutti i valori della probabilità considerata portano a risultati vicini a quelli reali. La differenza è di pochi punti percentuali.

Tabella 11: Analisi di sensitività per la maggiore probabilità di infettare, vaccinare e curare degli *influencer* sul *dataset* `obama`

$P_{\textbf{influencer}}$	**RMSE**
0.00	6.97 ± 4.12
0.25	9.67 ± 3.94
0.50	8.20 ± 3.56
0.75	8.58 ± 3.42
1.00	9.18 ± 3.57

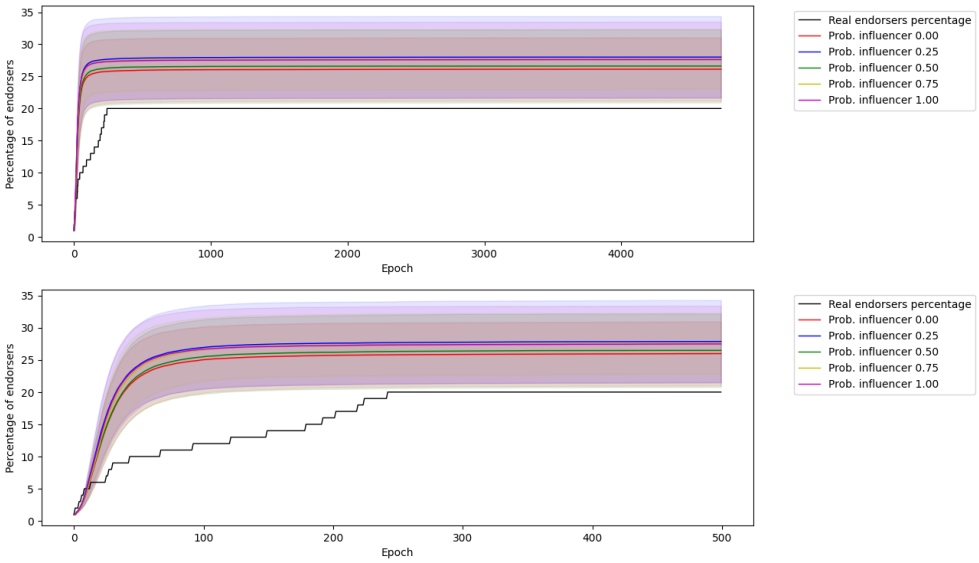

Figura 21: Grafici relativi alle percentuali medie di infetti per epoca ottenute per l'analisi di sensitività della maggiore probabilità di infezione, vaccinazione e cura degli *influencer*. Il grafico superiore considera tutte le epoche. Il grafico inferiore considera solo le prime 500 epoche.

5.5.5 Soglia delle *echo chamber*

L'ultima analisi di sensitività condotta riguarda la soglia per l'attivazione del meccanismo delle *echo chamber* descritto al Paragrafo 3.2. I valori considerati per l'analisi sono stati: 0.00, 0.25, 0.50, 0.75 e 1.00. I RMSE ottenuti sono mostrati nella Tabella 12. Nella Figura 22 invece vengono rappresentate le percentuali di infetti per epoca registrate per ogni valore analizzato.

Tabella 12: Analisi di sensitività per la soglia relativa all'effetto delle *echo chamber* sul *dataset* `obama`

$P_{\mathbf{echo}}$	**RMSE**
0.00	7.74 ± 4.37
0.25	8.05 ± 4.02
0.50	9.60 ± 4.81
0.75	17.20 ± 5.94
1.00	15.82 ± 7.17

In particolare si possono svolgere le seguenti osservazioni:

- **Confronto tra i valori del parametro**. Le percentuali di infetti più alte si ottengono con una soglia superiore a 0.75. Per soglie inferiori non si riscontrano differenze significative.
- **Confronto rispetto al *dataset***. Le soglie che portano a risultati più vicini a quelli reali sono inferiori a 0.75. Si ricorda qui che sotto la soglia di 0.50 è esclusa l'attivazione del meccanismo delle *echo chamber* descritto al Paragrafo 3.2. Il parametro calibrato in questo caso ha comunque riportato un valore di circa 0.61.

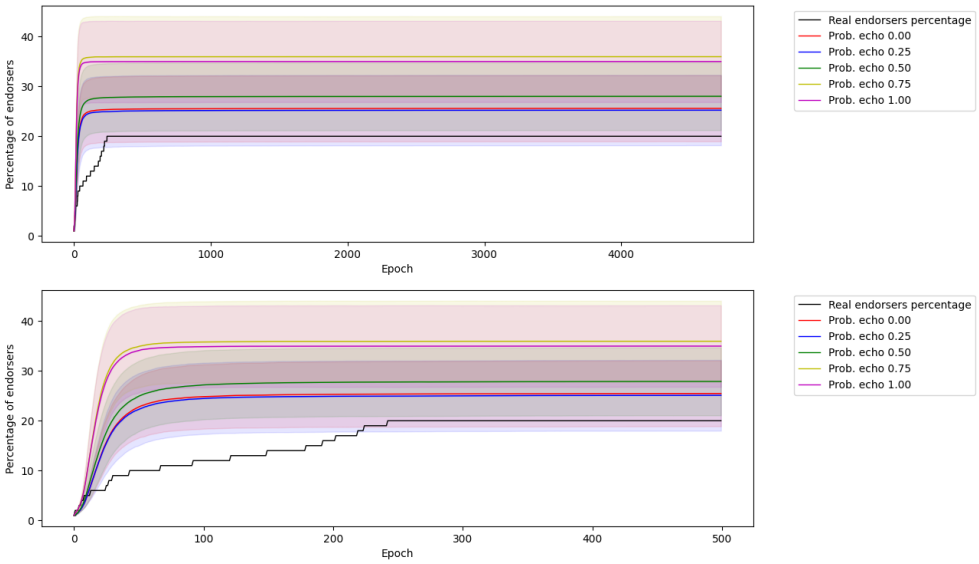

Figura 22: Grafici relativi alle percentuali medie di infetti per epoca ottenute per l'analisi di sensitività della soglia per il meccanismo delle *echo chamber*. Il grafico superiore considera tutte le epoche. Il grafico inferiore considera solo le prime 500 epoche.

6 Simulazioni

Nella Sezione precedente sono stati validati i modelli proposti. In questa Sezione verranno presentate e discusse alcune ulteriori simulazioni svolte con questi modelli. L'obiettivo di queste simulazioni è stato individuare contromisure efficaci alla diffusione di *fake news* in *social network*. A questo proposito sono state sperimentate le seguenti contromisure: (*i*) il blocco degli utenti a seguito di reclami; (*ii*) l'eliminazione dei *bot*; (*iii*) la vaccinazione degli *influencer*; e (iv) l'incremento del numero di *eternal fact-checker*.
Per le simulazioni di questa Sezione sono stati utilizzati i parametri calibrati con il valore di h pari a 0.50 indicati al Paragrafo 5.3.5 e gli ulteriori parametri indicati nella Tabella 13.

Tabella 13: Parametri predeterminati per la valutazione delle contromisure

nodi	$\delta_{\mathbf{in}}$	$\delta_{\mathbf{out}}$	**epoche**	***infl.***	***bot***	***e.f.c.***	**infetti iniziali**
2000	1	1	1000	20	40	40	100 (con *bot*)

Il numero di epoche è stato fissato a 1000 in quanto la validazione ha mostrato che ulteriori epoche non producono mutamenti significativi nei risultati. Il numero di influencer è stato fissato a 20, che è pari all'1% dei nodi. Questa percentuale è analoga a quella dei *dataset* `obama` e `palin`. Il numero di *bot* è pari al 2% dei nodi. Questa percentuale è ritenuta corretta in [21]. Il numero di *eternal fact-checker* è pari al numero di *bot* per le ragioni esposte al Paragrafo 5.3.1. Tuttavia le percentuali di *bot* ed *eternal fact-checker* sono state modificate per alcuni esperimenti. La percentuale di infetti iniziali è pari al 5% dei nodi.
Per la valutazione delle contromisure si è considerata ancora la percentuale di infetti per epoca. Questa percentuale è stata ancora calcolata come media dei risultati ottenuti su 30 reti diverse.

6.1 Simulazioni con il blocco degli utenti a seguito di reclami

Le prime simulazioni hanno riguardato il blocco degli utenti a seguito di reclami. Il numero di reclami necessario per bloccare un utente è 3. I valori considerati per la probabilità di presentare reclamo sono: 0.00, 0.25, 0.50,

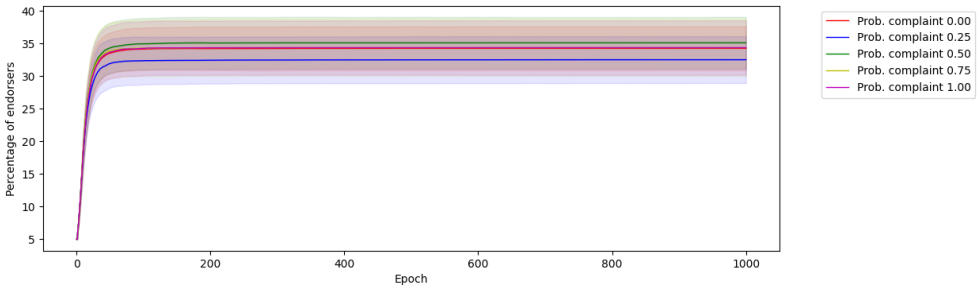

Figura 23: Grafico relativo alle percentuali medie di infetti per epoca ottenute nelle simulazioni con il blocco degli utenti.

0.75 e 1.00. La probabilità di reclamo nulla equivale all'assenza di contromisure.
Nella Figura 23 vengono rappresentate le percentuali di infetti per epoca registrate per ogni valore analizzato. La variazione della probabilità di presentare reclamo non sembra produrre risultati significativi in termini di riduzione del numero di infetti. Questo è confermato anche dal ridotto numero di utenti bloccati nelle simulazioni. Questo dato viene mostrato nella Tabella 14. Le differenze nelle percentuali di infetti sono imputabili alla stocasticità dei modelli utilizzati. In questo quadro si ritiene che il blocco degli utenti sia una

Tabella 14: Numero di utenti bloccati nelle simulazioni con i reclami

$P_{\textbf{complaint}}$	**utenti bloccati**
0.00	0.00 ± 0.00
0.25	0.67 ± 0.75
0.50	1.63 ± 1.22
0.75	1.53 ± 1.36
1.00	2.57 ± 2.36

contromisura inefficace a contenere la diffusione di *fake news*.

6.2 Simulazioni con la rimozione dei *bot*

La seconda contromisura analizzata ha riguardato la rimozione dei *bot*. In particolare si ricorda qui che nel modello proposto i *bot* sono attivi in ogni

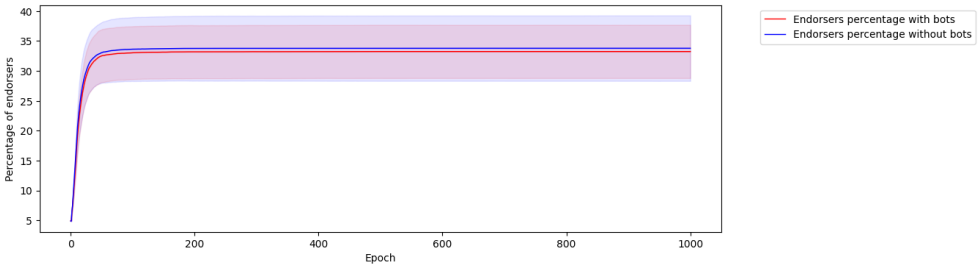

Figura 24: Grafico relativo alle percentuali medie di infetti per epoca ottenute nelle simulazioni con e senza *bot*.

istante di tempo, mentre l'attività degli altri utenti decresce progressivamente. Pertanto si vuole verificare quanto l'attività continua dei *bot* influisca sulla diffusione di *fake news*.
Nella Figura 24 vengono rappresentate le percentuali di infetti per epoca registrate con e senza *bot*. Il numero di infetti iniziali è lo stesso. Anche la rimozione dei *bot* non produce una riduzione significativa del numero di infetti per epoca. In questo quadro la rimozione dei *bot* costituisce una contromisura inefficace alla diffusione di *fake news*. E questa conclusione trova conferma anche nella letteratura scientifica, secondo cui nei *social network* reali la maggiore diffusione di *fake news* è imputabile (non tanto ai *bot*, quanto) agli umani [39].

6.3 Simulazioni con la vaccinazione degli *influencer*

La terza contromisura sperimentata ha riguardato la vaccinazione iniziale degli *influencer*. I risultati sono mostrati nella Figura 25.
In questo caso si nota come la vaccinazione iniziale degli *influencer* provochi una riduzione delle percentuali di infetti superiore al 20%. Questa riduzione è significativa per due motivi. Anzitutto perché presenta una varianza piccola. Quindi la riduzione è statisticamente robusta. Inoltre perché al Paragrafo 5.5.4 si è visto che il parametro di probabilità degli *influencer* ha un basso impatto sui risultati. E questo conferma la robustezza delle conclusioni raggiunte.
Nella Figura 26 viene mostrato lo stato epidemiologico degli utenti all'ultima epoca di simulazione su una rete con gli *influencer* vaccinati. Inoltre viene evidenziata la presenza degli *influencer* nella rete. Si nota come gli *influencer*

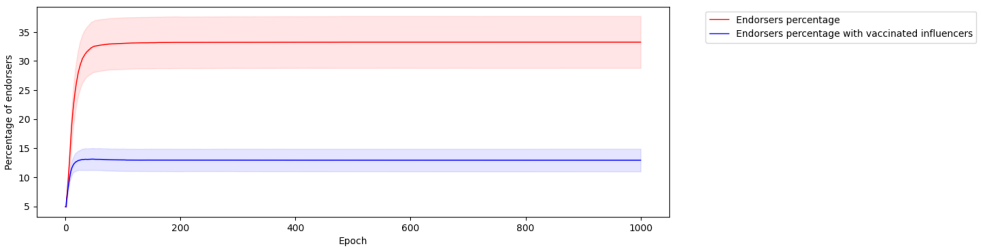

Figura 25: Grafico relativo alle percentuali medie di infetti per epoca ottenute nelle simulazioni con la vaccinazione degli *influencer*. In particolare la curva rossa rappresenta le percentuali medie di infetti senza vaccinazione degli *influencer*. La curva blu rappresenta invece le percentuali medie di infetti con la vaccinazione iniziale degli *influencer*.

vaccinati al centro della rete facilitino la diffusione della vaccinazione e della cura. E in questo modo limitano la propagazione della *fake news*.

6.4 Simulazioni con l'incremento degli *eternal fact-checker*

L'ultima contromisura sperimentata ha riguardato l'incremento del numero di *eternal fact-checker*. Si tratta dell'incremento iniziale del numero di utenti vaccinati (che non sono *influencer*). A questo proposito le percentuali di *eternal fact-checker* considerate sono state le seguenti: 2%, 5%, 7% e 10%. La percentuale del 2% è stata usata anche nelle simulazioni precedenti e corrisponde alla mancanza della presente contromisura. I risultati sono mostrati nella Figura 27.
Naturalmente all'aumentare del numero di *eternal fact-checker* decresce la percentuale di infetti per epoca. Tuttavia la percentuale di infetti rimane superiore al 20% anche con un numero di *eternal fact-checker* pari al 10% degli utenti. E la riduzione del numero di infetti presenta comunque una varianza elevata. Pertanto la contromisura in parola è efficace, ma richiede il coinvolgimento di un numero di utenti elevato per ottenere effetti significativi.

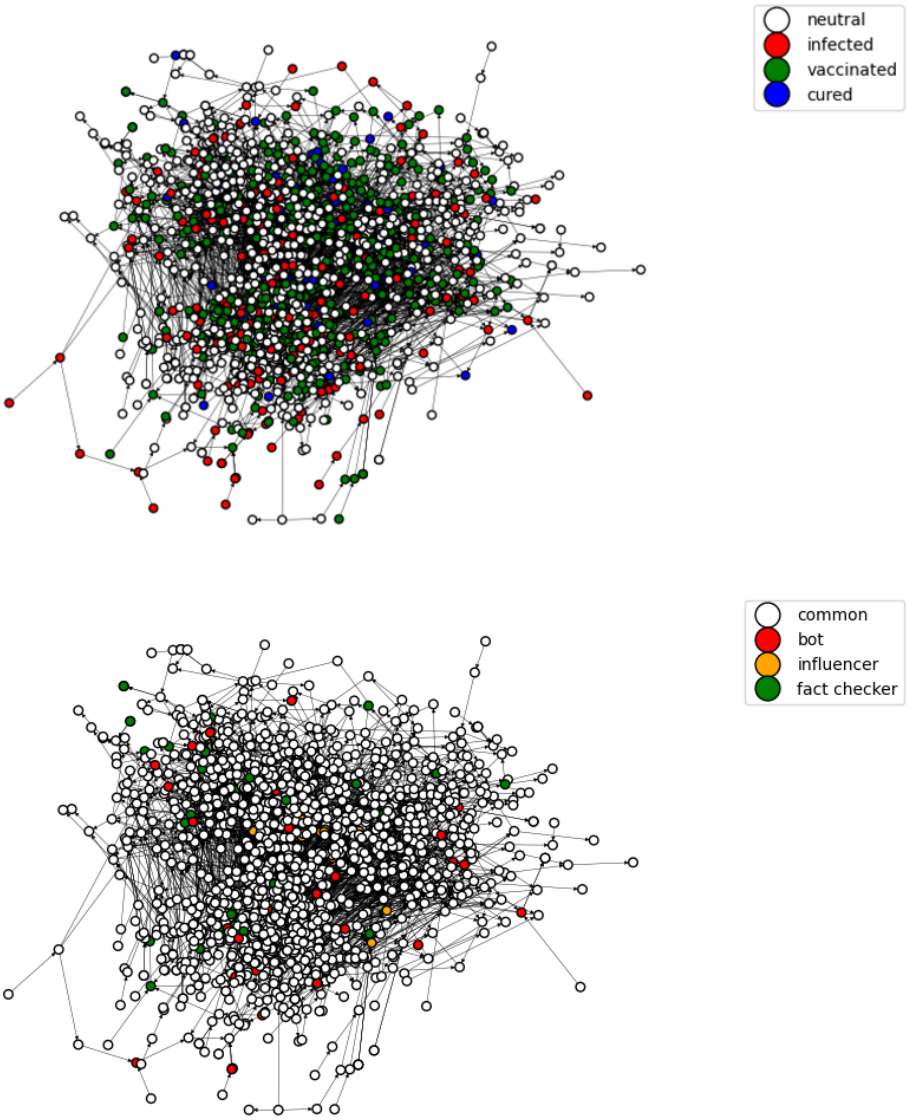

Figura 26: Il grafo superiore rappresenta lo stato epidemiologico degli utenti all'ultima epoca di simulazione su una rete generata con gli *influencer* vaccinati. Il grafo inferiore rappresenta invece il ruolo dei nodi all'ultima epoca.

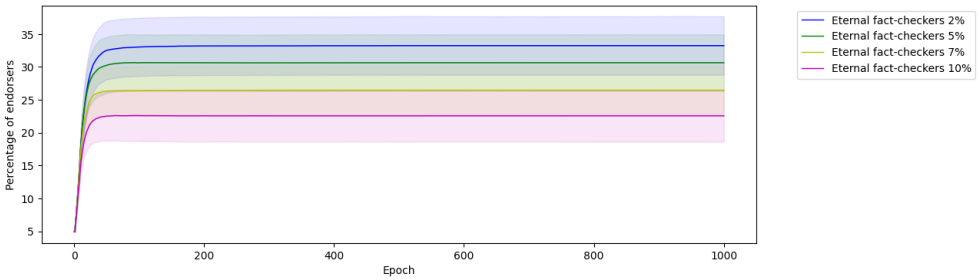

Figura 27: Grafico relativo alle percentuali medie di infetti per epoca ottenute nelle simulazioni con la variazione del numero di *eternal fact-checker*.

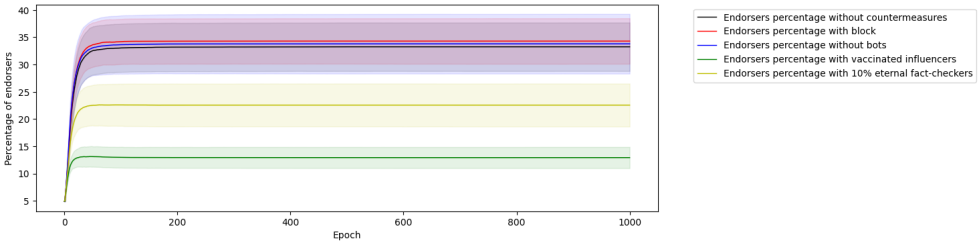

Figura 28: Grafico di confronto delle percentuali medie di infetti per epoca ottenute con diverse contromisure alla diffusione di *fake news*.

6.5 Sintesi

Nei Paragrafi precedenti sono state descritte le simulazioni svolte per individuare contromisure efficaci alla diffusione di *fake news*. In particolare le contromisure sperimentate sono state: (*i*) il blocco degli utenti a seguito di reclami; (*ii*) l'eliminazione dei *bot*; (*iii*) la vaccinazione degli *influencer*; e (iv) l'incremento del numero di *eternal fact-checker*. Nella Figura 28 vengono confrontati i risultati ottenuti con queste contromisure.
In particolare da questo confronto emerge che:

- il blocco degli utenti e la rimozione dei *bot* sono inefficaci a contrastare la diffusione di *fake news*;
- l'incremento del numero di *eternal fact-checker* riduce il numero di infetti in modo più marcato. Ma per ottenere effetti significativi occorre coinvolgere numerosi utenti. E la varianza del numero di infetti è comunque elevata;

- la vaccinazione degli *influencer* porta a ottenere i risultati migliori sia come riduzione (elevata) del numero di infetti sia come varianza (bassa) di questo numero. Inoltre questa riduzione è ottenuta più rapidamente rispetto alle altre contromisure. E questi risultati non sono influenzati in modo rilevante dal relativo parametro di probabilità. In particolare questa contromisura sfrutta la topologia della rete per raggiungere un elevato numero di utenti in quanto gli *influencer* sono i nodi con più *follower* nella rete. Pertanto il coinvolgimento di un numero anche piccolo di *influencer* permette di ottenere risultati molto significativi.

7 Conclusioni e sviluppi futuri

In questo lavoro si è analizzato il problema della diffusione di *fake news* in *social network* al fine di individuare possibili contromisure. Per svolgere questa analisi è stato necessario definire (*i*) un modello di creazione di reti sociali *scale-free* orientate e con omofilia e (*ii*) un modello di diffusione delle informazioni in un *social network*. I modelli definiti sono stati implementati come modelli basati su agenti. E questi modelli sono stati validati al fine di verificarne la correttezza.
I modelli validati sono stati poi utilizzati per svolgere alcune simulazioni. Queste simulazioni hanno avuto per oggetto diverse contromisure alla diffusione di *fake news*. In particolare le contromisure considerate sono state: (*i*) il blocco degli utenti a seguito di reclami; (*ii*) l'eliminazione dei *bot*; (*iii*) la vaccinazione degli *influencer*; e (iv) l'incremento del numero di *eternal fact-checker* (che sono utenti vaccinati diversi dagli *influencer*). Tra le contromisure considerate la vaccinazione degli *influencer* ha portato alla migliore riduzione del numero di utenti che credono alla *fake news*.
Il lavoro sin qui svolto si presta a ulteriori sviluppi. Se ne citano alcuni:

- I modelli proposti assumono che la rete creata non possa essere modificata aggiungendo o rimuovendo nodi e archi. Tuttavia i *social network* sono (non statici, ma) dinamici. Pertanto si potrebbe considerare un modello di creazione delle reti sociali che ammetta anche queste modifiche.

- I modelli proposti considerano la relazione di *follow* per la creazione di connessioni tra i nodi. Tuttavia un *social network* come Twitter presenta ulteriori forme di connessione tra gli utenti. Tra queste si cita per esempio il *retweet*, ovvero l'azione con cui un utente condivide con i propri *follower* un messaggio di un altro utente [40]. Pertanto si potrebbe definire un modello che consideri anche questa forma di connessione tra gli utenti.

- La calibrazione dei parametri svolta per questo lavoro ha incontrato diverse difficoltà. Anzitutto il *dataset* individuato per questo scopo presenta diversi difetti, tra cui l'incompletezza dei *tweet* sia rispetto agli utenti sia rispetto all'arco temporale considerato. Inoltre le limitate risorse computazionali disponibili hanno impedito una ricerca approfondita dei migliori valori dei parametri. In questo quadro il presente

lavoro potrebbe essere migliorato considerando *dataset* più completi e utilizzando macchine computazionalmente più potenti.

- Infine si potrebbe anche studiare l'impatto sugli utenti della diffusione di *fake news* attraverso fonti diverse. In particolare si potrebbe valutare l'impatto sugli utenti dell'esposizione a *fake news* sia attraverso *social media* che attraverso media tradizionali.

Riferimenti bibliografici

[1] Alstott, J., Bullmore, E., and Plenz, D. Powerlaw: a python package for analysis of heavy-tailed distributions. *arXiv preprint arXiv:1305.0215* (2013).

[2] Anwar, M. S., Saveski, M., and Roy, D. Balanced influence maximization in the presence of homophily. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (New York, USA, 2021), WSDM '21, Association for Computing Machinery, p. 175–183.

[3] Bandini, S., Manzoni, S., and Vizzari, G. Agent based modeling and simulation: An informatics perspective. *Journal of Artificial Societies and Social Simulation 12*, 4 (2009).

[4] Barabaśi, A.-L. *Network science*. Cambridge University Press, Cambridge, 2016.

[5] Barabaśi, A.-L., and Albert, R. Emergence of scaling in random networks. *Science 286*, 5439 (1999), 509–512.

[6] Bollobás, B., Borgs, C., Chayes, J., and Riordan, O. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (USA, 2003), SODA '03, Society for Industrial and Applied Mathematics, p. 132–139.

[7] Burbach, L., Halbach, P., Ziefle, M., and Calero Valdez, A. Opinion formation on the internet: The influence of personality, network structure, and content on sharing messages online. *Frontiers in Artificial Intelligence* (2020).

[8] De Domenico, M., Lima, A., Mougel, P., and Musolesi, M. The anatomy of a scientific rumor. *Scientific Reports 3* (2013).

[9] Epstein, J. M., and Axtell, R. *Growing Artificial Societies: Social science from the bottom up*. MIT Press, Cambridge, USA, 1996.

[10] Erdős, P., Rényi, A., et al. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences 5*, 1 (1960), 17–60.

[11] Foote, R. Mathematics and complex systems. *Science 318*, 5849 (2007), 410–412.

[12] Gausen, A., Luk, W., and Guo, C. Can we stop fake news? using agent-based modelling to evaluate countermeasures for misinformation on social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (2021), Association for the Advancement of Artificial Intelligence.

[13] Genesereth, M. R., and Nilsson, N. J. *Logical Foundations of Artificial Intelligence*. Elsevier Inc, Morgan Kaufmann, 1987.

[14] Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference* (Pasadena, USA, 2008), G. Varoquaux, T. Vaught, and J. Millman, Eds., pp. 11 – 15.

[15] Johnson, C. W. What are emergent properties and how do they affect the engineering of complex systems? *Reliability Engineering & System Safety* (2006), 1475–1481.

[16] Kaligotla, C., Yöcesan, E., and Chick, S. E. An agent based model of spread of competing rumors through online interactions on social media. In *2015 Winter Simulation Conference (WSC)* (2015), pp. 3088–3089.

[17] Kermack, W. O., and McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London* (1927).

[18] Klügl, F. A validation methodology for agent-based simulations. In *Proceedings of the 2008 ACM Symposium on Applied Computing* (New York, NY, USA, 2008), SAC '08, Association for Computing Machinery, p. 39–43.

[19] Liu, D., and Chen, X. Rumor propagation in online social networks like twitter – a simulation study. In *2011 Third International Conference on Multimedia Information Networking and Security* (2011), pp. 278–282.

[20] Lomonaco, F., Taibi, D., Trianni, V., Buršić, S., Donabauer, G., and Ognibene, D. Yes, echo-chambers mislead you too: A game-based educational experience to reveal the impact of social media personalization algorithms. In *Higher Education Learning Methodologies and Technologies Online* (Cham, 2023), G. Fulantelli, D. Burgos, G. Casalino, M. Cimitile, G. Lo Bosco, and D. Taibi, Eds., Springer Nature Switzerland, pp. 330–344.

[21] Lotito, Q. F., Zanella, D., and Casari, P. Realistic aspects of simulation models for fake news epidemics over social networks. *Future Internet 13* (2021).

[22] Mazzoli, M., Re, T., Bertilone, R., Maggiora, M., and Pellegrino, J. Agent based rumor spreading in a scale-free network. *arXiv preprint arXiv:1805.05999* (2018).

[23] McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27* (2001), 415–444.

[24] Newman, M. E. J. The structure and function of complex networks. *SIAM Review 45*, 2 (2003), 167–256.

[25] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Physical Review E. 64*, 2 (2001).

[26] Ognibene, D., Wilkens, R., Taibi, D., Hernández-Leo, D., Kruschwitz, U., Donabauer, G., Theophilou, E., Lomonaco, F., Bursic, S., Lobo, R. A., Sánchez-Reina, J. R., Scifo, L., Schwarze, V., Börsting, J., Hoppe, U., Aprin, F., Malzahn, N., and Eimler, S. Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence 5* (2023).

[27] Onuchowska, A., and Berndt, D. J. Using agent-based modelling to address malicious behavior on social media. In *International Conference on Interaction Sciences* (2019).

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,

Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[29] plfit. https://github.com/keflavich/plfit.

[30] Posetti, J., and Matthews, A. A short guide to the history of 'fake news' and disinformation. *International Center for Journalists* (2018).

[31] Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (Edinburgh, Scotland, United Kingdom, 2011), Association for Computational Linguistics, pp. 1589–1599.

[32] Rind, D. Complexity and climate. *Science 284*, 5411 (1999), 105–107.

[33] Russell, S. J., and Norvig, P. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020.

[34] Serrano, E., and Iglesias, C. A. Validating viral marketing strategies in twitter via agent-based social simulation. *Expert Systems with Applications 50* (2016), 140–150.

[35] Serrano, E., Iglesias, C. A., and Garijo, M. A novel agent-based rumor spreading model in twitter. In *Proceedings of the 24th International Conference on World Wide Web* (New York, NY, USA, 2015), WWW '15 Companion, Association for Computing Machinery, p. 811–814.

[36] Tweepy. https://www.tweepy.org/.

[37] Twitter. https://developer.twitter.com/en/products/twitter-api.

[38] Törnberg, P. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS ONE 13* (2018), 1–21.

[39] Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science 359*, 6380 (2018), 1146–1151.

[40] Weng, L., Menczer, F., and Ahn, Y.-Y. Virality Prediction and Community Structure in Social Networks. *Scientific Reports 3* (2013).

[41] Wooldridge, M., and Jennings, N. R. Intelligent agents: Theory and practice. *Knowledge Engineering Review* (1995), 115–152.