ISSN 1826-3534

federalismi.it
RIVISTA DI DIRITTO PUBBLICO ITALIANO, COMPARATO, EUROPEO

LA DISINFORMAZIONE ONLINE
24 APRILE 2020

Capire la diffusione della disinformazione e come contrastarla

di Giancarlo Ruffo
Professore associato di Informatica
Università di Torino

e Marcella Tambuscio
Postdoctoral researcher
Austrian Academy of Science in Vienna

Capire la diffusione della disinformazione e come contrastarla *

di Giancarlo Ruffo
Professore associato di Informatica
Università di Torino

e Marcella Tambuscio
Postdoctoral researcher
Austrian Academy of Science in Vienna

Abstract [It]: La proliferazione di fake news è uno dei temi più discussi degli ultimi anni: un problema non nuovo, ma amplificato dalle nuove tecnologie digitali che permettono scambi di informazioni sempre più rapidi e disintermediati. Proponiamo qui un modello per studiare la diffusione della disinformazione come un processo virale in cui bufale e relative smentite (debunking) competono tra loro, focalizzando la nostra attenzione soprattutto sul ruolo della struttura (topologia) della rete sociale sottostante. La segregazione strutturale aiuta o limita la propagazione? Le simulazioni del nostro modello mostrano che la risposta non è univoca. Infine consideriamo possibili strategie per suggerire policies efficaci e potenziare piattaforme di fact-checking.

Abstract [En]: Fake news proliferation has been largely discussed in the last years: a problem that is not new, but recently amplified by new digital technologies that provide decentralised tools to exchange information in real time. We propose here a model to study misinformation spreading as a viral process in which fake news and their debunking compete, focusing specially on the role of the underlying social network's structure (topology). Does structural segregation help or limit the propagation? Empirical simulations of our model show that there is not a unique answer. Finally we consider possible strategies to suggest effective policies and to empower fact-checking platforms.

1. Premessa

Nel 2013 il World Economic Forum ha inserito la disinformazione nella lista dei rischi tecnologici e geopolitici dei nostri tempi[1], mentre nel 2016 gli *Oxford dictionaries* hanno scelto la parola "post-verità" come parola dell'anno[2], indicando che viviamo in una società dove "i fatti effettivi hanno meno influenza nel formare l'opinione pubblica degli appelli emotivi e legati alle credenze personali". Chiaramente non si tratta di un fenomeno nuovo, ma in un mondo iperconnesso dove il modo di produrre e ricevere informazioni è cambiato radicalmente, Internet e nuovi media hanno un ruolo cruciale nella diffusione delle *fake news*. Negli ultimi anni è quindi emersa la necessità di studiare il fenomeno anche da un punto di vista quantitativo, perché ci si è resi conto della pericolosità delle conseguenze: un esempio su tutti, la bufala secondo la quale i vaccini causerebbero l'autismo che ha contribuito a determinare un calo delle

* Articolo sottoposto a referaggio.
[1] World Economic Forum (a cura di), *Digital wildfires in a hyperconnected world*, 2013, disponibile su: <http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/>.
[2] Oxford Dictionaries (a cura di), *Word of the year 2016*, disponibile su: <https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016>.

vaccinazioni volontarie in molti paesi. In Italia ad esempio si è passati da valori di copertura vaccinale anti-polio di 96,1% nel 2011 a 93,4% nel 2015.[3]

La domanda che ci poniamo è se la verifica dei fatti (*fact-checking*) possa avere un ruolo nel contrasto alla diffusione della disinformazione.

2. Il ruolo dei nuovi media

Diversi studi psicologici condotti negli ultimi decenni hanno evidenziato alcuni elementi che possono favorire la propagazione dei *rumor*, come la ripetizione[4], la popolarità dell'argomento, l'ambiguità[5], la fiducia nella fonte e la concordanza con le opinioni personali[6]. Le nuove tecnologie digitali amplificano alcune di queste caratteristiche: infatti una completa de-centralizzazione dell'informazione su vasta scala ha portato ad una democratizzazione della conoscenza rendendo molti contenuti accessibili a tutti, ma d'altronde ha anche reso le reti estremamente vulnerabili, ad esempio, alla manipolazione delle notizie. In particolare, *social media* come Facebook o Twitter, dove ogni utente può potenzialmente raggiungere milioni di persone in pochi minuti, vengono utilizzati sempre più spesso per diffondere informazioni[7], fornendo un terreno fertile per bufale, contenuti di scarsa qualità e teorie del complotto, per almeno tre ragioni. Primo, caratteristiche come la presenza di lingue diverse o la diversità dei contenuti (ad esempio satira e informazione) possono creare ambiguità. Una ricerca del 2017 ha infatti evidenziato che fattori rilevanti nel determinare la viralità di un'informazione sono la limitata capacità di attenzione di ogni individuo e la sovraesposizione informativa (*information overload)* data dal bombardamento mediatico, più che la qualità[8]. In secondo luogo, l'informazione, una volta divenuta virale, genera delle vere e proprie cascate di condivisioni[9], esponendo ripetutamente gli utenti agli stessi contenuti. In terzo luogo, l'esposizione continua ad un certo tipo di informazione può rafforzare alcuni pregiudizi, alimentando processi cognitivi come il *confirmation bias*[10], ovvero la scelta mirata di alcune argomentazioni per supportare e consolidare una propria opinione. Questo avviene quotidianamente sulle bacheche dei *social network*, che sono costruiti appositamente per connettere persone con le stesse idee o per ricostruire *online*

[3] Istituto superiore di sanità (a cura di), *Epicentro. Le vaccinazioni in Italia.* (<https://www.epicentro.iss.it/vaccini/dati_Ita>).
[4] R.H. KNAPP, *A psychology of rumor*, in *Public opinion quarterly*, 8, n. 1/1944. pp. 22-37.
[5] G.W. ALLPORT - L. POSTMAN, *The psychology of rumor*, Oxford, 1947.
[6] N. DI FONSO - P. BORDIA, *Rumor psychology: Social and organizational approaches*, Washington, 2007.
[7] A. MITCHELL - J. GOTTFRIED - M. BARTHEL - E. SHEARER, *The modern news consumer*, in *Pew Research Center*, disponibile su: <http://www.journalism.org/2016/07/07/the-modern-news-consumer/>.
[8] X. QIU - D. OLIVERA - A. S. SHIRANZI - A. FLAMMINI - F. MENCZER, *Limited individual attention and online virality of low-quality information*, in *Nature Human Behavior*, n. 1/2017, pp. 1-32.
[9] P.A. DOW - L.A. ADAMIC - A. FRIGGERI, *The anatomy of large facebook cascades*, in *Proc. of the 7th Intern. AAAI Conf. on Weblogs and Social Media (ICWSM)*, Ann Harbor (MI), 2013.
[10] D. CENTOLA, *The spread of behavior in an online social network experiment*, in *Science*, 329, n. 5996/2010. pp. 1194-1197.

comunità già esistenti, secondo il principio dell'omofilia[11], ovvero il fenomeno che vede aumentare la probabilità alla connessione reciproca tra persone che hanno interessi o caratteristiche simili. Inoltre, queste piattaforme spesso utilizzano algoritmi che filtrano i contenuti in base alle preferenze dell'utente o alla sua attività passata: la conseguenza più immediata di questo tipo di architettura è un'esposizione molto selettiva che porta, in taluni casi, alla formazione delle ormai famose bolle, o *echo-chambers*. Recenti ricerche hanno mostrato come la disinformazione possa apparire molto polarizzata nelle conversazioni *online*[12], tuttavia è bene ricordare che il ruolo della struttura della rete sociale sottostante nei processi di diffusione delle informazioni non è ancora del tutto chiaro[13].

3. Il Fact-Checking e altre possibili soluzioni

Quali soluzioni adottare dunque? Gli autori del già citato rapporto del *World Economic Forum* auspicavano lo sviluppo futuro di norme per un'etica digitale che rendano il comportamento degli utenti più responsabile, sottolineando il problema di fornire delle regole senza limitare i diritti di libertà di parola.
Ad oggi, l'unica contromisura che il singolo utente può adottare per la diffusione di notizie false è il *fact-checking*, ovvero un processo di verifica e accertamento dei fatti, seguito dalla diffusione del *debunking*, la smentita di una "bufala". Nell'ultimo decennio sono apparsi molti siti Internet con questo scopo (e.g., Snopes, Politifact, Factcheck - oppure Attivissimo e Butac in Italia), mentre famose testate giornalistiche hanno costruito intere piattaforme appositamente dedicate alla verifica di certe affermazioni in occasione per esempio delle elezioni politiche, o di temi particolarmente dibattuti (ad esempio: vaccini, migranti, terrorismo). In alcuni casi questi servizi si sono rivelati effettivamente utili ad arginare la diffusione delle notizie false[14], mentre in altri casi è stato osservato come il *debunking* possa addirittura peggiorare la situazione[15]. Questo fenomeno, studiato anche nelle scienze politiche ed in ambito socio-psicologico[16], è chiamato *backfire effect* e si presenta quando la bufala che si vuole smentire è fortemente in linea con le

[11] M. MCPHERSON - L. SMITH-LOVIN - J. M. COOK, *Birds of a feather: Homophily in social networks*, in *Annual Review of Sociology*, 27, n. 1/2001, pp. 415-444.
[12] M. DEL VICARIO - A. BESSI - F. ZOLLO - F. PETRONI - A. SCALA - G. CALDARELLI, H. E. STANLEY - W. QUATTROCIOCCHI, *The spreading of misinformation online*, in *Proceedings of the National Academy of Sciences (PNAS)*, 113, n. 3/2016, pp. 554-559.
[13] E. BAKSHY - I. ROSENN - C. MARLOW - L. ADAMIC, *The role of social networks in information diffusion,* in *Proc. of the 21st intern. conf. on World Wide Web (WWW)*, New York, 2012, pp. 519-528.
[14] A. FRIGGERI - L.A. ADAMIC - D. ECKLES - J. CHENG, *Rumor cascades*, in *Proc. of 8th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, Palo Alto, 2014, pp.101-110.
[15] A. BESSI - G. CALDARELLI - M. DEL VICARIO - A. SCALA - W. QUATTROCIOCCHI, *Social determinants of content selection in the age of (mis) information*, in *Proc. of the Intern. Conf. on Social Informatics*, Londra, 2014, pp. 259-268.
[16] B. NYHAN - J. REIFLER, *When corrections fail: The persistence of political misperceptions*, in *Political Behavior*, vol. 32, n. 2/2010, pp. 303-330.

convinzioni personali di chi ascolta, o, alternativamente, quando un *fact-checking* non convincente crea, sul lungo termine, un livello di confusione riguardo la veridicità del contenuto stesso.

I contributi dell'informatica ed in generale dalle cosiddette "scienze dure" allo studio della diffusione delle *fake-news* si possono raggruppare in due grandi categorie: da una parte lo sviluppo di tecniche di identificazione automatica che possano aiutare a classificare velocemente le notizie false o l'attività sospetta di utenti che vogliono intenzionalmente diffondere disinformazione; dall'altra, l'ideazione di modelli che possano descrivere e riprodurre i processi di propagazione, per proporre nuove contromisure e cercare di limitare i danni. Per quanto riguarda il primo gruppo, si tratta principalmente di tecniche basate su apprendimento automatico e linguistica computazionale, che mirano a misurare la credibilità di una singola notizia attraverso l'analisi di *pattern* o caratteristiche ricorrenti (ad esempio il numero di *follower*, *retweet* o la presenza di particolari menzioni, *hashtag* o informazioni legate alla posizione geografica dell'utente). Non ci dilungheremo su questa categoria di soluzioni perché fuori dallo scopo del presente contributo, ma ne indicheremo velocemente alcune caratteristiche oltre che alcuni limiti di cui i sistemi attuali risentono: sostanzialmente, la macchina viene addestrata a riconoscere la notizia potenzialmente falsa o studiandone lo stile con il quale è stata scritta, o riferendosi alla "reputazione sociale" dell'utente (o testata giornalistica, o blog, etc.) che l'ha prodotta inizialmente o diffusa. Ad esempio, esistono strumenti abbastanza efficaci di riconoscimento di *bot*[17] e *troll*, che sono rispettivamente programmi o utenti veri e propri che possono produrre notizie fasulle per un qualche tipo di interesse, tendenzialmente perseguendo un obiettivo di tipo manipolatorio. Il problema con questo tipo di approccio è soprattutto legato alla sua durata: il modello che viene appreso dalla macchina in modo automatico è, per quanto complesso, composto da regole che possono essere usate anche da chi fabbrica notizie false per capire, a sua volta, un comportamento da *non* emulare per evitare di essere riconosciuti. Pertanto, prima o poi, i *bot* ed i *troll* impareranno come passare dalle forche caudine di uno specifico sistema di riconoscimento automatico. Esiste un'intera branca dell'Intelligenza Artificiale che si chiama *Adversarial Learning* che serve proprio ad apprendere modelli di classificazione automatica che possano resistere contro attacchi di questo tipo. Nonostante questo, i sistemi di riconoscimento allo stato dell'arte sono destinati a commettere ancora molti errori e, nel tempo, ad essere superati da comportamenti sempre più sofisticati da parte dei creatori di notizie fasulle[18]. Insieme a questo problema sostanzialmente tecnico, in realtà esiste un altro aspetto di natura etica di cui la comunità scientifica è consapevole: pur supponendo che sia possibile creare il sistema di classificazione perfetto in grado di etichettare come "vere" o "false" le

[17] OSoME project (a cura di), *Botometer*, disponibile su: <https://botometer.iuni.iu.edu/#!/>.
[18] R. ZELLERS et al., *Defending against neural fake news*, in *Advances in Neural Information Processing Systems*, 2019, pp. 9051-9062.

notizie, quale autorità deve essere delegata ad eseguire il filtro dell'informazione? Se debba essere il governo, la società privata che gestisce il servizio di *social networking* o l'utente stesso, è un problema aperto. Tale problema difficilmente sarà risolto nel breve termine, a maggior ragione se pensiamo che i sistemi attuali eseguono la classificazione secondo un insieme di regole appreso automaticamente che sono racchiuse sostanzialmente in una "scatola nera" all'interno della quale è difficile trovare una spiegazione umanamente comprensibile dei processi che sono stati dedotti dall'algoritmo di intelligenza artificiale. Per approfondimenti si veda alla voce *Explainable Artificial Intelligence (XAI)*[19] e al cosiddetto *Right to Explanation*[20]. Questo senza neanche tirare in ballo la trasparenza dell'operatore, pubblico o privato, che esegue materialmente il filtraggio e che potrebbe, in un mondo non ideale, ridefinire i concetti di "vero" e di "falso" in accordo con le proprie strategie.

L'approccio che noi invece abbiamo seguito negli ultimi anni per studiare il fenomeno è invece legato all'adozione di modelli dei processi di diffusione in cui, tradizionalmente, le informazioni sono viste come virus che possono contagiare le persone. In questo ambito l'approccio più comune è quello di ricorrere ai modelli che sono stati tradizionalmente usati nell'ambito della diffusione delle epidemie. In particolare, i modelli di base sono quelli compartimentali SIR e SIS in cui la popolazione è divisa in compartimenti che indicano lo stadio della malattia (SIR = *Susceptible, Infected, Recovered* - SIS = *Susceptible, Infected, Susceptible*) e l'evoluzione del processo è regolata da probabilità di transizione da uno stadio all'altro all'interno di un sistema di equazioni differenziali. Molti modelli esistenti che simulano la diffusione di *rumor* sono basati sul modello Daley-Kendall[21], a sua volta derivante dal modello SIR, in cui un utente passa da *Susceptible* (ovvero non sa nulla) a *Infected* quando viene raggiunto dalla notizia, e successivamente da *Infected* a *Recovered* quando viene raggiunto dalla smentita. Negli anni sono state presentate altre varianti contribuendo ad arricchire una letteratura molto vasta, per cui rimandiamo il lettore interessato ad ulteriori approfondimenti che esulano dallo scopo di questo contributo.

4. Il modello SBFC

Prima di parlare del nostro specifico contributo sul problema della diffusione dell'informazione, vale probabilmente la pena ricordare brevemente al lettore cosa si intende, tecnicamente, per "teoria scientifica" e "modello". Una teoria scientifica non è semplicemente un'idea maturata da un individuo, o un'opinione valida tanto quanto un'altra: essa è una spiegazione di un determinato aspetto del mondo

[19] R. Guidotti - A. Monreale - S. Ruggieri - F. Turini - F. Giannotti - D. Pedreschi, *A Survey of Methods for Explaining Black Box Models*, in *ACM Compututing Survey,* vol. 51, n. 5/2018.
[20] L. Edwards - M. Veale. *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, in *Duke Law and Technology Review,* 2017.
[21] D.J. DALEY - D.G. KENDALL, *Stochastic rumours*, in *IMA Journal of Applied Mathematics*, vol. 1, n.1/1965, pp. 42-55.

naturale che è stata motivata da ripetute osservazioni e da una non trascurabile attività empirica[22]. Può essere validata o falsificata e nel secondo caso deve essere abbandonata senza rimpianti per consentire l'elaborazione di una teoria più accurata. In questo contesto, allo scopo di elaborare la nostra teoria scientifica che riguarda il modo in cui le informazioni false e la loro smentita possono diffondersi in una rete sociale, noi abbiamo costruito un modello che ci consente di partire da alcune premesse per arrivare ad alcune conclusioni tramite processi matematici deterministici, oppure stocastici. Un utile supporto sono anche le attività di simulazione, dove vengono assegnati alcuni valori specifici ai parametri iniziali del modello. L'elaborazione simulata del processo sottostante è funzionale a capire cosa può accadere a fronte di un contesto dove siano riscontrati quei valori iniziali. Infine, il modello può produrre delle "previsioni": anche in questo caso, il termine è da intendere in modo prettamente tecnico e niente ha a che fare con le sfere di cristallo o indovini. Un fenomeno viene *predetto* dal nostro modello se non è stato dato tra i parametri iniziali, ma viene restituito alla fine del processo di analisi e/o di simulazione e corrisponde ad un'osservazione già acquisita o da acquisire. Quando il fenomeno viene "predetto", diciamo quindi che esso può essere spiegato dalla nostra teoria in funzione dei parametri iniziali. Capiamo pertanto che, in quest'ottica, stiamo parlando di modelli di natura molto diversa da quelli in genere generati da un sistema di apprendimento automatico: in questo ultimo caso abbiamo predizioni sulla classe di appartenenza di certe osservazioni, ad esempio notizie che sono classificate come vere o come false, eseguite da modelli che sono difficilmente spiegabili (vedi sezione precedente); qui, al contrario, abbiamo bisogno di modelli che ci aiutino a spiegare perché un determinato fenomeno si è verificato, ad esempio perché una notizia falsa in un contesto è stata arrestata tempestivamente grazie ad un'attività di *debunking* accurata, mentre un'altra bufala, in un altro contesto, sia diventata virale.

Dopo questa premessa metodologica, passiamo a descrivere gli aspetti principali del nostro contributo in questo ambito. Infatti, nel 2014 abbiamo iniziato ad occuparci del problema di costruire una teoria scientifica che potesse spiegare perché il fact-checking a volte risultasse efficace ed altre volte inefficace nel contrasto della disinformazione online. In quel periodo pochi studiosi nell'ambito delle scienze dure stavano affrontando questo stesso problema e pertanto nessuno dei modelli esistenti sulla diffusione delle informazioni nelle reti sociali forniva una rappresentazione adeguata del fenomeno del *debunking*. Data l'enorme attenzione per le piattaforme di *fact-checking*, ci sembrò pertanto necessario cercare di acquisire una metodologia per stimare la reale efficacia di questi strumenti di contrasto analizzando il processo nella sua completezza. Nei modelli precedenti infatti, viene solo considerata la transizione da *Infected* a *Recovered*, mentre nella realtà che osserviamo ci appare chiaro come il *debunking* possa anche diffondersi

[22] T. Ghose, *"Just a Theory": 7 Misused Science Words*, in *LiveScience*, *Scientific American,* 2013.

autonomamente (ad esempio condividendo un link ad una piattaforma come Snopes o qualche *social network*), o come una notizia possa venire dimenticata nel frattempo.

Andiamo con ordine. Partendo sempre dall'epidemiologia e da un modello esistente[23] per la diffusione simultanea di due *rumor*, il nostro scopo era creare un nuovo modello[24] che rappresentasse la propagazione della disinformazione per provare a simulare una vera e propria competizione tra bufale e smentite. Innanzitutto, abbiamo bisogno di una serie di assunzioni. In primo luogo, assumiamo che la notizia falsa la cui diffusione vogliamo studiare sia effettivamente tale, ovvero esista per essa già e sia accessibile una smentita ufficiale o svolta da un *fact-checker* autorevole. Inoltre, sappiamo che i singoli individui possono essere stati esposti alla notizia attraverso vari canali ed in particolare dai propri contatti della propria rete sociale (sulla quale non facciamo assunzioni particolari per rendere il discorso quanto più generale possibile). Gli individui (o anche agenti), in un determinato istante di tempo possono essere in uno ed un solo stato tra i seguenti: Suscettibile (*Susceptible - S*), Credulone (*Believer - B*) o *Fact-Checker* (*FC*). Da queste sigle pertanto il nome del modello: SBFC.

Un agente nello stato *S* può quindi essere raggiunto e potenzialmente "contagiato" da una delle due informazioni: potrà diventare rispettivamente un Credulone *(B)* o un *Fact-Checker (FC)*. Questo significa che il singolo agente non solo acquisisce un'informazione che contiene la bufala o la sua smentita, ma decide anche consapevolmente di prendere una posizione. Potrebbe rimanere indifferente e rimanere nello stato S, oppure passare ad un nuovo status, sia esso B oppure FC.

Passiamo alla definizione degli altri parametri del modello: oltre ad una "velocità di diffusione" β, che ritroviamo in quasi tutti i modelli compartimentali, abbiamo introdotto un parametro α, corrispondente ad una certa "credibilità della bufala", che fornisce alla notizia falsa un piccolo vantaggio sulla smentita: più alta è la credibilità, più probabilità ci sono che la notizia si diffonda. Inoltre, abbiamo deciso di inserire nel nostro modello i seguenti tre fenomeni: (1) probabilità di *influenza sociale,* che caratterizza il passaggio dallo stato *S* a quello *B* o *FC*: ogni agente modifica il suo stato con una probabilità che dipende dalla velocità di diffusione, dalla credibilità della bufala e dallo status dei suoi vicini, in particolare dalla maggioranza di *creduloni* e *fact-checker* tra loro; (2) probabilità di verifica, che caratterizza per ogni agente il passaggio dallo stato *B* a *FC*: ogni utente nello stato *B* può decidere di verificare la notizia con una data probabilità; (3) probabilità di *dimenticare*, che caratterizza il ritorno dallo stato *B* o *FC* ad *S*: ogni agente nello stato *B* o *FC* può effettivamente dimenticare la posizione che ha preso nei riguardi della veridicità di una notizia acquisita e tornare, con una certa probabilità, ad uno stato di suscettibilità. Per completezza,

[23] D. Trpevski - W. KS Tang - L. Kocarev, *Model for rumor spreading over networks*, in *Physical Review E*, vol. 81, n. 5/2010.
[24] M. Tambuscio - G. Ruffo - A. Flammini - F. Menczer. *Fact- checking effect on viral hoaxes: A model of misinformation spread in social networks,* in *Proc. of the 24th Inter. Conf. on World Wide Web Companion (WWW 2015)*, Firenze, 2015, pp. 977-978.

è opportuno osservare che ogni agente, ad ogni istante di tempo, può rimanere nello stesso stato in cui si trova.

Il modello ora è completo. Esso si può trasformare in un insieme di equazioni matematiche che possono calcolare, a partire da valori iniziali, delle condizioni di equilibrio. In pratica, si aspetta che il modello esegua una serie di iterazioni, fino a quando il numero di agenti in stato S, B o FC si stabilizza. Questo ci può servire, quindi, a capire se a partire da alcune determinate condizioni iniziali (valori specifici assegnati alle probabilità di dimenticare, di verifica, di influenza sociale, oltre che determinate topologie di rete sociali ed una certa percentuale di nodi che fanno partire "l'infezione", ovvero che propagano la notizia falsa e la sua smentita) possiamo stimare se e quando il sistema raggiungere l'equilibrio. Una volta che il sistema avrà raggiunto l'equilibrio, possiamo quindi stimare se avranno vinto i *creduloni* o i *fact-checker*. In termini pratici, questo significa che si calcolerà all'equilibrio del sistema (convenzionalmente definito in un momento infinito nel tempo) il numero di nodi in stato B (B_∞) e quelli in stato FC (FC_∞). Se $B_\infty > FC_\infty$ allora avrà "vinto" la *fake-news*. Se invece accadrà il contrario $FC_\infty > B_\infty$ vorrà dire che sotto quelle condizioni i *fact-checker* saranno in vantaggio. Certamente l'ideale sarebbe che $B_\infty = 0$, ma vedremo che questa eventualità è piuttosto rara.

5. Simulazioni basate su agenti

Come spiegato nella sezione precedente, modelli di questo tipo vengono trasformati in sistemi di equazioni differenziali che si risolvono per via analitica. In aggiunta alla soluzione puramente matematica, è anche possibile risolvere il sistema per via simulativa: ogni singolo individuo della rete sociale è un agente che interagisce con altri agenti della rete in un dato istante di tempo. In base a queste interazioni, allo stato dei singoli agenti e alle probabilità delle transizioni, con appositi programmi che eseguono le cosiddette 'simulazioni basate su agenti', si calcolano gli stessi valori di cui abbiamo discusso precedentemente: il numero di Susceptible, Believer e Fact-Checker ad ogni iterazione, fino a quando si osserva una convergenza di questi numeri per stimare per via maggiormente empirica i valori di e S_∞, B_∞ e FC_∞.

A questo punto è bene ricordare che il nostro è un modello basato su probabilità, quindi assegnando dei valori ai parametri iniziali ed eseguendo il modello sotto quelle condizioni potremmo avere simulazioni diverse che mostrano dei risultati diversi al raggiungimento dell'equilibrio. In pratica quindi, quando si ha a che fare con un processo stocastico di questo tipo, si eseguono moltissime simulazioni di uno stesso sistema sotto le identiche condizioni per poi individuare il comportamento medio del sistema al raggiungimento dell'equilibrio, stimando pertanto il margine di errore tra le diverse simulazioni. Questo è importante per capire che non abbiamo a che fare con un modello deterministico che restituirà sempre

gli stessi risultati a fronte delle identiche condizioni iniziali. Nei processi stocastici, invece, il caso a volte può perturbare i risultati e creare delle fluttuazioni ed è per questo che il risultato finale si valida solo a fronte dell'esecuzione di molte simulazioni.

Uno dei primi risultati che le simulazioni del nostro modello ci hanno consentito di raggiungere è la stima della soglia minima, sotto diverse condizioni iniziali, della probabilità di verifica che è necessario avere per essere ragionevolmente sicuri che la "bufala" venga rapidamente rimossa da una rete sociale[25]. Tale stima ci consente di fare delle considerazioni molto pessimistiche: infatti, la fake-news molto credibile difficilmente viene risolta da un fact-checking, anche molto capillare e pur assumendo un'alta probabilità di verifica. Le cose cambiano apparentemente se si riesce a garantire una soglia di almeno il 20% di Creduloni che viene convertita in Fact-Checker: in questi casi, la bufala scompare progressivamente dalla rete, con una vittoria netta delle strategie di *debunking*. Purtroppo è facile immaginare che tale soglia sia davvero difficile da raggiungere in pratica, dato che appare utopistico supporre che almeno il 20% di coloro che hanno creduto ad una notizia falsa si ricreda e che, anzi, diventi alfiere della posizione opposta. Si noti anche, come riprenderemo nelle conclusioni, che al momento non esiste uno studio empirico accurato dal punto di vista scientifico che fornisca un'evidenza del fatto che la probabilità di verifica sia diversa da 0, con le inevitabili conseguenze che tali osservazioni ci porterebbero a trarre.

Le simulazioni di cui abbiamo appena riferito consideravano essenzialmente due tipi di topologie della ipotetica rete sociale: una creata in modo del tutto casuale secondo il modello che in letteratura scientifica è noto con il nome dei suoi autori Erdös-Rényi[26], la seconda invece grazie ad un modello più realistico noto come *preferential attachment*[27], che tiene in considerazione l'alta eterogeneità dei nodi di una rete e del fenomeno dell'emergenza dei cosiddetti *hub*, ovvero nodi che hanno un numero notevolmente più alto rispetto alla media di contatti (ad esempio, persone estremamente popolari). In entrambi i casi i risultati delle due simulazioni erano confrontabili e per questo motivo abbiamo deciso di aggiungere alla topologia generata secondo il *preferential attachment* un ulteriore livello di realismo, considerando una società che si divide in due gruppi di persone: "scettici" ed "ingenui"[28]. Sotto questa prospettiva, non esiste una notizia falsa ugualmente credibile per tutti. Essa infatti avrà un valore di credibilità molto basso per tutti gli scettici e molto alto per gli ingenui. Inoltre, abbiamo eseguito le nostre simulazioni precedendo diversi livelli di "segregazione" tra scettici ed ingenui: gli scettici, ad esempio, possono essere molto ben collegati

[25] M. Tambuscio - G. Ruffo - A. Flammini - F. Menczer. *Fact- checking effect on viral hoaxes: A model of misinformation spread in social networks,* in *Proc. of the 24th Inter. Conf. on World Wide Web Companion (WWW 2015)*, Firenze, 2015, p. 979.
[26] P. Erdős - A. Rényi, *On Random Graphs. I*, in *Publicationes Mathematicae,* n. 6/1959, pp. 290–297.
[27] A.-L. Barabási - R. Albert, *Emergence of scaling in random networks*, in *Science,* vol. 286, n. 5439/1999, pp. 509–512.
[28] M. Tambuscio - D. F. M. Oliveira - G. L. Ciampaglia - G. Ruffo, *Network segregation in a model of misinformation and fact-checking*, in *Journal of Computational Social Science*, vol. 1, n. 2/2018, pp. 261-275.

tra di loro e debolmente rispetto agli ingenui (alta segregazione), oppure molto meno assortiti nello stabilire amicizie (bassa segregazione). Fondamentalmente, cerchiamo di eseguire le nostre simulazioni in contesti diversi dove le cosiddette camere d'eco possono manifestarsi o meno. I risultati, tenendo conto di queste nuove condizioni iniziali, iniziano ad essere particolarmente interessanti. Infatti, si vede che la vittoria tra falso e vero adesso dipende da una relazione complessa (non lineare) tra livello di segregazione e probabilità di dimenticare. Molto brevemente, abbiamo due situazioni abbastanza opposte: se la probabilità di dimenticare è molto bassa (gli agenti quindi difficilmente dimenticano la posizione che hanno assunto tornando ad uno stato *Susceptible*) la bufala avrà una maggiore probabilità di diffusione in una rete molto segregata, ma rimarrà contenuta soltanto tra gli "ingenui"; se, al contrario, la probabilità di dimenticare è alta, le connessioni tra scettici ed ingenui non faranno che peggiorare la situazione, facendo diffondere la bufala anche tra chi teoricamente è maggiormente portato a verificare le notizie e le fonti. Abbiamo pertanto un risultato molto interessante che consente di spiegare un fenomeno osservabile anche empiricamente: in presenza di una bufala "innocua", ovvero di quel "gossip" senza fondamento che non incide particolarmente sui nostri convincimenti più radicati, una minore disponibilità a dialogare con i nostri contatti più ingenui aiuterebbe a contenere la notizia falsa in tempi molto più veloci e diminuire il rischio di diffusione della bufala tra gli scettici. Al contrario, quando abbiamo a che fare con una vera e propria "teoria del complotto" (laddove gli agenti che hanno preso una posizione difficilmente dimenticheranno la loro opinione sulla quale hanno modo di confrontarsi quasi quotidianamente, vedi il caso della disputa pro-vax contro no-vax), il modo migliore per contenere il danno sembrerebbe proprio cercare di essere quanto più interattivi possibile con coloro che appartengono a bolle sociali diverse dalla nostra, e utilizzare questi legami per diffondere il debunking della bufala.

Purtroppo, l'osservazione dei dati che riusciamo ad estrarre dai *social media* riguardo alle più popolari teorie del complotto ci restituisce la situazione peggiore, secondo la nostra teoria scientifica e le nostre simulazioni: in questi casi, la comunità sono ideologicamente molto polarizzate. Si tratta quindi, nella maggiore parte dei casi, di capire quale possa essere la strategia migliore per contenere il più possibile la diffusione delle *fake-news*. Pertanto nel nostro terzo e più recente studio di carattere simulativo[29] abbiamo effettuato una cosiddetta *what-if analysis*, che consiste nell'iniettare il seme del *fact-checking* a partire da agenti con caratteristiche sociali diverse tra di loro e rendere questi *debunker* sempre attivi. Ad esempio, abbiamo supposto che i primi *debunker* della rete sociale possano essere selezionati (i) a caso tra i vari agenti della rete, (ii) tra coloro che hanno il maggior numero di contatti (i.e., gli *hub*), (iii) tra gli agenti "scettici" che

[29] M. Tambuscio - G. Ruffo, *Fact-checking strategies to limit urban legends spreading in a segregated society*, in *Applied Network Science*, 4, n. 116/2019.

hanno più contatti tra gli "ingenui" rispetto agli altri (i.e., i cosiddetti *bridge*, ponti, delle reti sociali con alto grado di segregazione). Abbastanza prevedibilmente, il *fact-checking* non mirato, effettuato da agenti selezionati in modo del tutto casuale, non ha praticamente alcuna efficacia. Al contrario, la diffusione del *debunking* iniziato da agenti molto popolari (*hub*) sembra funzionare molto meglio rispetto ad agenti "molto democratici", ovvero quegli agenti che hanno la possibilità nella loro vita quotidiana di intercettare individui con posizioni molto diverse tra di loro. La congiunzione degli ultimi due fattori potrebbe portare a strategie molto pratiche ed efficaci: gli insegnanti ed i medici, ad esempio, sono spesso dei "ponti" sociali che hanno l'onere e l'onore di interagire con porzioni della popolazione più esposte alle informazioni di scarsa qualità provenienti da fonti diverse e pertanto più vulnerabili rispetto ad altre categorie sociali. Da questo punto di vista, l'esperimento finlandese di investire molto sulla formazione scolastica nella creazione di un insieme di strumenti metodologici per l'individuazione della disinformazione[30] sembra proprio andare nella direzione suggerita dal nostro modello. Maggiore la popolarità dell'agente, maggiore la sua responsabilità: ad esempio diventa sempre più rilevante il diritto ad un'informazione accurata, tra i sei diritti aletici già proposti da D'Agostini e Ferrera[31]: il giornalista professionista e l'affidabilità delle sue fonti e della sua analisi dei fatti è pertanto un messaggero estremamente cruciale in questo dibattito. Nessuno del resto si deve sentire escluso: lo scienziato faccia uno sforzo maggiore verso la divulgazione accessibile da tutti, il politico riconosca maggiore legittimità alle competenze acquisite e non a quelle dichiarate, ed il mondo dello spettacolo interrompa il ciclo non virtuoso di mettere fatti ed opinioni sullo stesso piano.

6. Discussione finale e limitazioni

Il nostro modello resta uno tra i molti presentati negli ultimi anni per spiegare il fenomeno della diffusione delle *fake news*: il nostro contributo originale sta nel rappresentare il problema come un processo virale in cui la bufala e la smentita competono tra loro, rendendo quindi il *fact-checking* una parte integrante del processo, che può influenzarne l'evoluzione stessa. Gli ostacoli che si presentano nel raccogliere e trattare i dati in questo campo, spesso incompleti perché rimossi o modificati, rendono al momento quasi impossibile una validazione puntuale di questi modelli, principalmente perché è molto difficile stimare la porzione suscettibile di popolazione e dell'esposizione reale degli utenti. Il limite di questi modelli, infatti, è che rappresentano l'evoluzione del *belief*, ovvero dell'opinione delle persone riguardo un certo

[30] J. Henley, *How Finland starts its fight against fake news in primary schools*, in *The Guardian*, 29 Gennaio 2020, disponibile su: <https://www.theguardian.com/world/2020/jan/28/fact-from-fiction-finlands-new-lessons-in-combating-fake-news>.
[31] F. D'AgostinI - M. Ferrera, *La verità al potere: sei diritti aletici*, Torino, 2019.

argomento (in questo caso, una bufala), ma i dati (incompleti) provenienti dai *social media* riguardano l'attività parziale degli utenti e non necessariamente riflettono tutte le loro opinioni. Nonostante ciò, crediamo che analizzare da un punto di vista modellistico e teorico questi processi possa aiutarci nella loro comprensione e suggerire nuovi strumenti (o esperimenti) per tentare di arginare i danni che la propagazione di notizie false può creare. Infatti, nell'ottica di mantenere la rete una piattaforma democratica, è stato più volte ribadito che sarebbe preferibile evitare l'uso di *blacklist* o strumenti che possano limitare la libertà di espressione, fornendo invece servizi di *fact-checking* accurati e facilmente accessibili, come ad esempio il progetto SOMA (*Social Observatory for Disinformation and Social Media Analysis*[32] – Osservatorio Sociale per la Disinformazione e l'Analisi dei Social Media), lanciato nel Novembre 2018 attraverso un finanziamento Horizon 2020 della Commissione Europea. I risultati teorici delle nostre analisi possono indicare agli sviluppatori di queste piattaforme nuovi spunti per renderle ancora più efficaci, sfruttando ad esempio una propagazione mirata del *debunking* coordinata da una comunità di *fact-checker* localizzati in punti strategici della rete, oltre che suggerire ai politici e ai legislatori quali strategie incentivare e quali abbandonare perché non efficaci neanche in un contesto del tutto ideale.

[32] Progetto SOMA (*Social Observatory for Disinformation and Social Media Analysis)*, disponibile su: <https://www.disinfobservatory.org>.