

Bias e Fairness in Information Sharing e Filtering

Luciano Caroprese¹, Massimo Guarascio¹, Marco Minici^{1,2}, Francesco Sergio Pisani¹, Ettore Ritacco¹, Giuseppe Manco¹

1. ICAR - Istituto di Calcolo e Reti ad Alte Prestazioni

2. Università degli Studi di Pisa

1. {nome.cognome}@icar.cnr.it 2. marco.minici@phd.unipi.it

Abstract

L'IA ricopre un ruolo di primo piano nei meccanismi che governano le attuali reti sociali e gran parte dei servizi digitali con cui interagiamo ogni giorno. Un uso improprio può alimentare disuguaglianze, polarizzare opinioni e amplificare convinzioni false/errate. In questo documento, sono delineate alcune delle linee di ricerca di maggiore interesse, sia in ambito accademico che industriale, che mirano a stimolare lo sviluppo di piattaforme di IA più eque, trasparenti ed affidabili. Successivamente sono presentate altresì alcune delle soluzioni definite dal gruppo di ricerca dell'ICAR-CNR volte ad affrontare le problematiche connesse allo sviluppo di questa nuova generazione di IA.

1 Introduzione

L'influenza degli algoritmi di IA su fenomeni sociali come la polarizzazione delle opinioni o l'amplificazione di disuguaglianze pre-esistenti ha suscitato vasto interesse sia da parte di ricercatori ed esperti - PNAS ha recentemente dedicato uno special issue al tema [Baldassarri e Page, 2021] - che dalla cosiddetta società civile di tutto il mondo, come dimostrano le politiche sull'adozione di metodi automatici basati su IA - vedi articolo 22 del GDPR europeo - o le discussioni successive alle audizioni del Senato americano che hanno coinvolto Facebook [Zorthian, 2021]. In questo documento, vengono delineate tre linee di ricerca tra di loro complementari che mirano a stimolare lo sviluppo di piattaforme che integrino componenti di IA più eque, trasparenti ed affidabili. La prima linea di ricerca ha come oggetto la creazione di framework robusti in grado di misurare il ruolo che gli algoritmi hanno nel far emergere fenomeni rischiosi per la società e l'individuo. Successivamente, sono introdotte tecniche di debiasing finalizzate ad invertire - o quantomeno a mitigare - gli effetti negativi riscontrati. Il terzo argomento si focalizza sulla trasparenza algoritmica tramite l'impiego di metodi di explainability. Nel documento, è altresì effettuata un'analisi dello stato dell'arte di queste tematiche e vengono discusse le sostanziali challenges che ognuna di loro presenta.

2 Linee di ricerca emergenti

2.1 Analisi di Echo Chambers e Polarizzazione delle opinioni

I social media hanno cambiato significativamente il modo in cui le persone consumano l'informazione, al punto tale che molti eventi di grande interesse a livello nazionale ed internazionale (ad esempio, le elezioni politiche negli USA e le primavere arabe) siano stati influenzati dal modo in cui essi operano. Una delle peculiarità delle piattaforme social è l'uso degli algoritmi come mediatore tra utenti ed informazione, al contrario dei media offline in cui la mediazione è generalmente attuata da giornalisti professionisti. Il contributo degli algoritmi di IA nel far emergere determinati fenomeni è stato sotto attenta osservazione negli ultimi anni sia dalla comunità scientifica che dalle istituzioni. Recenti studi suggeriscono come gli utenti social tendano ad accedere alle informazioni su uno spettro di opinioni limitato, formando *bolle sociali* di individui omogenei [Pariser, 2011], a causa del filtro algoritmico. Inoltre, visto che gli utenti di queste bolle consumano informazioni che tendono a confermare i loro preconcetti, i social media possono trasformarsi in *echo chambers* [Quattrociocchi *et al.*, 2016], rinforzando le loro convinzioni. Tuttavia la visione sul tema è lontana dall'essere unanime. [Groshek e Koc-Michalska, 2017] sostiene che gli elettori più attivi sui social durante le elezioni americane del 2017 siano anti-correlati col supporto alla candidatura di Trump. [Barberá *et al.*, 2015] afferma che la segregazione ideologica sui social sia stata sovrastimata e le analisi di [Bakshy *et al.*, 2015] su dati reali di Facebook suggeriscono che il filtro algoritmico abbia invece incrementato l'esposizione a punti di vista eterogenei. Altri studi hanno associato la presenza di echo chambers online a quelle geografiche/offline [Bastos *et al.*, 2018; Morales *et al.*, 2021]. Comprendere il contributo causale degli algoritmi in questi fenomeni e quali siano i meccanismi che innescano eventuali eventi avversi (ad esempio polarizzazione di opinioni ed echo chambers) rimangono ancora domande aperte che necessitano di ulteriore ricerca.

2.2 Tecniche di debiasing

Con l'avvento della nuova era digitale, accanto all'aumento della capacità dei sistemi di stoccaggio dati, è possibile riscontrare una crescita esponenziale dei contenuti digitali prodotti e condivisi in rete, i quali richiedono una nuova genera-

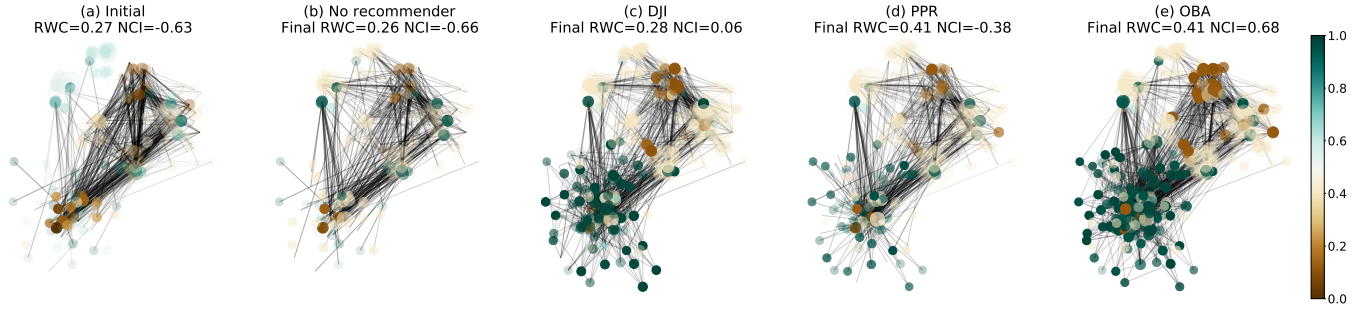


Figura 1: Simulazione di PROD. In (a) il grafo iniziale, in (b) il grafo finale senza un recommender, mentre in (c), (d) ed (e) il risultato con diversi algoritmi. La barra di colore rappresenta l’opinione dell’utente, mentre RWC e NCI sono due misure di polarizzazione ed echo chamber.

zione di AI per poter essere processati efficacemente. Questo fenomeno prende il nome di *data abundance* e coinvolge quelle piattaforme che hanno delle tecnologie di IA a sostegno dell’esperienza utente. Gli algoritmi giocano un ruolo cruciale in quanto abilitano l’utente finale alla navigazione di vasti cataloghi di dati. Gli ambiti di applicazione comprendono i social media così come piattaforme di streaming multimediale o e-commerce. Queste tecniche di IA sono più comunemente chiamate sistemi di raccomandazione, in quanto raccomandano all’utente un insieme di (pochi) item con cui interagire filtrando cataloghi di migliaia di items. L’obiettivo di questi sistemi è quello di stimare la quantità $P(Y|X)$ cioè l’evento Y che l’utente interagisca positivamente una volta esposto all’item X . Essendo i dataset di addestramento generalmente raccolti durante l’uso della piattaforma, si viene a creare un fenomeno di closed-loop [Chen *et al.*, 2021] tra umano, algoritmo e dati in cui gli items maggiormente popolari vengono più frequentemente raccomandati e cliccati non permettendo di stimare correttamente la quantità di interesse $P(Y|X)$. Le conseguenze sono il restringimento della raccomandazione ad un insieme molto ristretto di items che possono minare l’obiettivo di personalizzazione del servizio (quindi la soddisfazione dell’utente), indurre meccanismi perniciosi come le bolle sociali e le echo chambers prima descritte o ineguaglianze rispetto a popolazioni minoritarie [Fabbri *et al.*, 2020].

Le attuali linee di ricerca si concentrano sul debiasing dei sistemi di raccomandazione tramite l’uso di opportune ripetizioni della loss [Schnabel *et al.*, 2016], tecniche di campionamento dell’insieme di addestramento [Boratto *et al.*, 2021] oppure, più recentemente, l’applicazione di tecniche di inferenza causale [Zhang *et al.*, 2021]. Uno dei problemi che minano l’adozione di tecniche di debiasing in scenari reali è quello di mantenere delle performances eguali o superiori alle tecniche classiche.

2.3 Explainable Embeddings

Attualmente, per poter gestire modalità di dati che non hanno una naturale rappresentazione vettoriale viene fatto uso di embeddings. Gli embeddings sono utili in svariati campi come la raccomandazione, l’analisi di social network, la diffusione di informazione e altri. La natura black-box degli odierni modelli di embeddings rende opaco il significato

dei fattori latenti individuati ed il problema è esacerbato dalla presenza di bias negli algoritmi che pone dei seri interrogativi sui possibili risvolti negativi derivanti dal loro uso. Da qui la necessità di ideare modelli che generino explainable embeddings, le cui componenti dello spazio geometrico siano quindi interpretabili.

3 Risultati di ricerca

3.1 Misurare l’effetto dei people-recommender sulla formazione di echo chambers

Tra gli algoritmi che hanno suscitato maggiormente l’attenzione ci sono i cosiddetti *people recommenders*, che abilitano funzionalità di suggerimento quali ‘*Persone che potresti conoscere*’, ‘*Persone da seguire*’. Il loro ruolo nella formazione delle reti sociali degli utenti ha una diretta conseguenza su quali informazioni e con quali persone ogni utente interagisce. Tuttavia, l’esatto contributo dei people recommenders in tal senso è ancora oggetto di ricerca. Idealmente, una possibile soluzione consisterebbe nell’implementare una sorta di RCT (Randomized Controlled Trial) nel quale ad una parte del traffico sia rimossa la componente algoritmica, mentre alla restante sia permessa la navigazione sulla piattaforma dotata del people recommender. Tale ipotesi tuttavia è ostacolata dal rischio del social media di perdere utenza a causa del mancato servizio di raccomandazione.

Per questo motivo, è stato definito PROD [Cinus *et al.*, 2021], un modello che permette di sperimentare in-vitro gli effetti degli algoritmi di raccomandazione di persone sui social media. Tramite (i) l’utilizzo di un random network model in cui è possibile simulare diversi livelli iniziali di omofilia e segregazione, (ii) diversi modelli di opinion dynamics che descrivono il modo in cui gli utenti interagiscono tra di loro e (iii) la possibilità di inserire nella procedura l’effetto di un people recommender, è possibile stimare l’effetto degli algoritmi di raccomandazione di persone in termini di polarizzazione delle opinioni e sviluppo di echo chambers. In Figura 1 è riportato un esempio di come una rete sociale iniziale si sviluppi a seguito dell’utilizzo di diversi tipi di people recommenders.

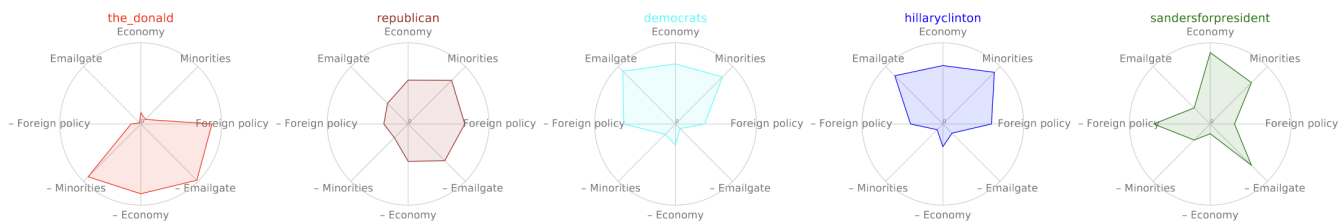


Figura 2: Esempio di embedding ideologici su determinati topics di esempio.

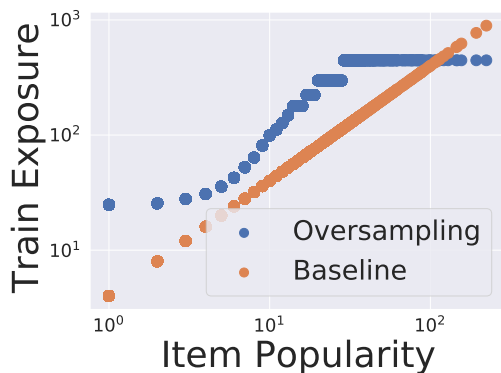


Figura 3: Distribuzione degli esempi con una strategia classica (arancio) ed una di ricampionamento (blu). È possibile osservare un incremento bilanciato dell'esposizione degli item meno popolari durante l'addestramento. Grafico in scala log-log.

3.2 Tecniche di debiasing per raccomandazioni eque

I modelli di Collaborative Filtering sono frequentemente utilizzati per affrontare problemi di raccomandazione di contenuti in quanto coniugano efficacia e semplicità. Difatti, richiedono esclusivamente lo storico delle interazioni utente-item, senza alcun bisogno di informazioni ausiliarie. Come detto precedentemente, anche questi tipi di algoritmi sono soggetti al *popularity bias* in quanto gli item molto popolari sono più frequentemente suggeriti nelle liste di raccomandazione e meglio valutati rispetto a quelli meno popolari. Per una discussione sulle conseguenze si rimanda all'introduzione o al tutorial [Chen *et al.*, 2021]. L'adozione di tecniche di debiasing in scenari reali è ostacolata dal trade-off tra performances e fairness. Lo studio di tecniche che possano mantenere una qualità predittiva alta e garantire la fairness è quindi un problema ancora aperto.

In uno studio preliminare condotto dall'ICAR-CNR [Caroprese *et al.*, 2021], sono presentate tre tecniche: una basata su una opportuna ripesatura della funzione di perdita, una seconda sul ricampionamento di esempi poco popolari e un'altra che propone di combinare tramite un ensemble un modello noto in letteratura con un altro specializzato nella predizione di item poco popolari. I risultati ottenuti sono promettenti in quanto, in alcuni scenari, la tecnica di ricampionamento permette di incrementare notevolmente le performances globali oltre quelle ottenute sugli oggetti meno popolari. Un esempio di come il ricampionamento modifichi la distribuzione originale del dataset è riportato in Figura 3.

3.3 Embeddings ideologici

Per realizzare modelli di raccomandazione, una delle tecniche che ha trovato largo impiego in letteratura consiste nel proiettare ogni utente ed item in uno spazio geometrico tramite un embedding. Ogni embedding consente di caratterizzare l'entità, quindi è possibile raccomandare contenuti simili ricercando nello spazio geometrico gli embeddings più vicini. Tuttavia, il significato di ogni componente dell'embedding non è conosciuto per cui non è possibile stabilire perché due entità siano geometricamente vicine. In [Monti *et al.*, 2021] è stato proposto un modello che, partendo dall'assunzione che una coppia di utenti interagisca su uno stesso item se interessati allo stesso topic, permette di inferire embeddings su spazi ideologici ben definiti. Una sperimentazione effettuata su un dataset estratto da Reddit ha permesso di estrapolare l'allineamento ideologico di cinque diversi subreddit stimati col metodo proposto (vedi Figura 2).

4 Attività progettuali

Nell'ambito dello sviluppo di sistemi di raccomandazione evoluti l'ICAR-CNR è coinvolto nella proposta progettuale *CATCH 4.0*, presentata nel contesto dei PON I&C 2014-2020 FESR del MISE. Uno degli obiettivi principali del progetto è appunto sviluppare strumenti mirati di *advanced content delivery*, in grado da un lato di variegare il contenuto informativo dei cataloghi digitali, arricchendo l'informazione disponibile con altre informazioni reperibili da sorgenti eterogenee o desumibili tramite tecniche di intelligenza artificiale e machine learning, dall'altro di adattare i contenuti da proporre all'utente in base alle sue preferenze contestuali desumibili in maniera implicita (tramite l'analisi del suo storico) ed esplicita (tramite interazione).

Nell'ambito della call europea Call H2020-ICT-2018-20, periodo 2020-2023, l'ICAR-CNR è coinvolto nel progetto HumaneAI-Net. L'obiettivo di questo progetto è quello di definire soluzioni teoriche e tecnologiche di frontiera capaci di modellare la rivoluzione dell'AI in corso, affinché (a) possano essere di beneficio per gli esseri umani, visti sia come singoli individui che come società, e (b) aderiscano ai valori etici, alle norme sociali, culturali, legali e politiche europee. In particolare, il progetto si propone di affrontare le seguenti sfide: (1) comprendere il comportamento umano, (2) interagire in scenari sociali complessi, (3) migliorare le capacità umane e potenziare individui e società nel loro insieme e (4) rispetto dell'autonomia e dell'autodeterminazione umana.

Riferimenti bibliografici

- [Bakshy *et al.*, 2015] E. Bakshy, S. Messing, e L. A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, June 2015.
- [Baldassarri e Page, 2021] Delia Baldassarri e Scott E Page. The emergence and perils of polarization. *Proceedings of the National Academy of Sciences*, 118(50), 2021.
- [Barberá *et al.*, 2015] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, e Richard Bonneau. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10):1531–1542, October 2015.
- [Bastos *et al.*, 2018] Marco Bastos, Dan Mercea, e Andrea Baronchelli. The geographic embedding of online echo chambers: Evidence from the Brexit campaign. *PLOS ONE*, 13(11):e0206841, November 2018.
- [Boratto *et al.*, 2021] Ludovico Boratto, Gianni Fenu, e Mirko Marras. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management*, 58(1):102387, 2021.
- [Caroprese *et al.*, 2021] Luciano Caroprese, Giuseppe Manco, Marco Minici, Francesco Sergio Pisani, e Ettore Ritacco. Unbiasing collaborative filtering for popularity-aware recommendation. 2021.
- [Chen *et al.*, 2021] Jiawei Chen, Xiang Wang, Fuli Feng, e Xiangnan He. Bias Issues and Solutions in Recommender System: Tutorial on the RecSys 2021. In *Fifteenth ACM Conference on Recommender Systems*, pages 825–827, Amsterdam Netherlands, September 2021. ACM.
- [Cinus *et al.*, 2021] Federico Cinus, Marco Minici, Corrado Monti, e Francesco Bonchi. The effect of people recommenders on echo chambers and polarization. *arXiv preprint arXiv:2112.00626*, 2021.
- [Fabbri *et al.*, 2020] Francesco Fabbri, Francesco Bonchi, Ludovico Boratto, e Carlos Castillo. The effect of homophily on disparate visibility of minorities in people recommender systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 165–175, 2020.
- [Groshek e Koc-Michalska, 2017] Jacob Groshek e Karolina Koc-Michalska. Helping populism win? social media use, filter bubbles, and support for populist presidential candidates in the 2016 us election campaign. *Information, Communication & Society*, 20(9):1389–1407, 2017.
- [Monti *et al.*, 2021] Corrado Monti, Giuseppe Manco, Cigdem Aslay, e Francesco Bonchi. Learning ideological embeddings from information cascades. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1325–1334, 2021.
- [Morales *et al.*, 2021] Gianmarco De Francisci Morales, Corrado Monti, e Michele Starnini. No echo in the chambers of political interactions on reddit. *Scientific Reports*, 11(1):1–12, 2021.
- [Pariser, 2011] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin UK, 2011.
- [Quattrociocchi *et al.*, 2016] Walter Quattrociocchi, Antonio Scala, e Cass R. Sunstein. Echo Chambers on Facebook. SSRN Scholarly Paper ID 2795110, Social Science Research Network, Rochester, NY, June 2016.
- [Schnabel *et al.*, 2016] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, e Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.
- [Zhang *et al.*, 2021] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, e Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. *arXiv preprint arXiv:2105.06067*, 2021.
- [Zorthian, 2021] Julia Zorthian. Washington wants to regulate facebook’s algorithm. that might be unconstitutional. <https://time.com/6106643/facebook-algorithm-regulation-legal-challenge/>, 2021. [Online; accessed 18-January-2022].