ANTONIO SCALA

DELLA NATURA NON-EUCLIDEA DEI BIG DATA SU INTERNET E SULLE SUE CONSEGUENZE

ABSTRACT: ABOUT THE NON EUCLIDEAN NATURE OF INTERNET'S BIG DATA AND ITS CONSEQUENCES
In this paper we discuss the non Euclidean nature of navigable Big Data, hinting as it could introduce – thanks to its combinatorial richness – the possibility of multiple interpretation spaces. We also propose that, in such structures, exploration algorithms have a paradoxical nature insofar they either enhance human biases or are useless. Finally, we propose that Internet could be used as an artificial universe for observing constructivism at work.

1. *Introduzione*

Il termine "Big Data" è stato associato a un nuovo approccio metodologico secondo il quale si sarebbe smesso di prendere decisioni basandosi su istinto ed esperienza e al loro posto si sarebbe usata l'enorme mole di dati disponibili per trovare le risposte all'interno dei dati stessi. In realtà, la nascita del termine "Big Data" andrebbe piuttosto ricercata nella necessità di trovare termini accattivanti, attività divenuta ormai necessaria sia per rinvenire i fondi di ricerca necessari alla ricerca scientifica, sia per questioni di marketing. Storicamente, gli scienziati che più si sono ritrovati a maneggiare enormi moli di

dati sono stati i fisici impegnati a ricercare e rilevare eventi rari nelle reazioni fra particelle elementari, o gli astrofisici in ascolto di segnali dallo spazio; eppure, data l'omogeneità dei dati e la disponibilità/costruibilità di modelli interpretativi, mancava il senso di impotenza e di smarrimento che si dovrebbe provare davanti ai "Big Data", moli di dati che vanno oltre le capacità di comprensione ed elaborazione e riguardano spesso ambiti – come le scienze sociali – nei quali c'è scarsità, quando non mancanza, di modelli quantitativi e validabili.

La sorgente principale dei Big Data è Internet, nata per scopi militari (controllo e comunicazione), estesa poi dalla comunità scientifica a essere mezzo di immagazzinamento e scambio di conoscenza nonché ambiente collaborativo che permettesse di abbattere le distanze geografiche. Il semplice mettere insieme più basi di dati, ovvero la possibilità di immagazzinare in maniera "distribuita" dati in siti o su macchine diverse, ha di per sé enormemente aumentato la mole di dati disponibile su ogni singolo argomento; eppure la rivoluzione maggiore è stata l'introduzione del World Wide Web, ovvero l'implementazione dell'idea dei collegamenti ipertestuali. La capacità di un oggetto di riferirsi ad altri oggetti fa esplodere in modo combinatorio le possibilità di creare relazioni fra gli oggetti; i dati iniziano ad arricchirsi di significato in base ai riferimenti che hanno, ma allo stesso tempo la scelta dei riferimenti – ovvero la loro contestualizzazione – può cambiare il significato del singolo dato (vedi fig. 1). I Big Data smettono di essere collezioni di oggetti, ma vengono arricchiti dalle relazioni fra oggetti, vengono a far parte di una o più reti. È chiaro a questo punto che l'idea che i "Big Data" parlino da se è quanto meno velleitaria: al variare della rete di relazioni nella quale osservo uno stesso oggetto, il significato può cambiare. Ma anche ammettendo l'esistenza di una unica rete di relazioni, possiamo cercare di

analizzare cosa succede nel momento in cui andiamo a considerare la struttura di questa rete.

2. *Effetto "small-world" e natura non euclidea di Internet*

Internet è costruito per essere navigabile: ad esempio, le sue pagine web (il cosiddetto "World Wide Web" o WWW) contengono collegamenti ad altre pagine che permettono di "saltare" dai contenuti di una pagina all'altra. Tali collegamenti creano una rete fra le pagine (vedi fig. 2): durante la navigazione online, noi ci muoviamo quindi su una rete che – come vedremo – ha delle caratteristiche estremamente peculiari.

Le reti "navigabili" cui siamo storicamente abituati sono le reti dei trasporti: ad esempio, la metropolitana di Londra può essere vista come una rete che collega le stazioni e ci permette di andare dall'una all'altra; non a caso le mappe della metropolitane sono schematizzate usando puntini rappresentanti le stazioni (gli oggetti "navigabili") connessi da linee che indicano la presenza di collegamenti fra gli oggetti. Le reti stradali che collegano le città, o le reti che portano il gas ad agglomerati urbani e industriali, sono ulteriori esempi di reti "navigabili" cui siamo abituati; in ogni caso però ci troviamo di fronte a reti "planari", ovvero che sono immerse[1] su una superficie bidimensionale e (localmente) euclidea. Questo significa che l'esplorazione di tali reti è soggetta a dei vincoli: in particolare, se raddoppio la distanza a cui posso viaggiare, statisticamente quadruplicherò il numero di località raggiungibili. In generale, andando fino a una distanza d, potrò raggiungere un numero di oggetti che cresce come $d \times d$. Nota bene che la legge di crescita del numero di oggetti è legata alla dimensione dello spazio che sto esplorando: se fossi in tre

[1] Usando il termine immersione vogliamo stimolare la curiosità del lettore ad approfondire il concetto di planarità, soprattutto se riferito a reti casuali immerse nel piano (cfr. M. Barthelemy, *Spatial networks*, in «Physics Reports», 499, 2011, pp. 1-101).

dimensioni, crescerebbe come *d x d x d*; in quattro, come *d x d x d x d* e così via. In realtà, essendo abituati a vivere e interagire sulla superficie della terra, per noi l'ambiente naturale è proprio quello bidimensionale: non a caso l'idea di una terra piatta, per quanto in contraddizione con le evidenze scientifiche, è quella di più semplice accettazione in quanto corrisponde alle nostre esperienze quotidiane. Insomma, le reti navigabili di cui storicamente abbiamo esperienza concreta sono oggetti bidimensionali.

Nel WWW ci troviamo in uno spazio diverso. Anche qui, come in una metropolitana si va da una stazione all'altra, possiamo ora andare da una pagina all'altra. Però in questo caso non abbiamo una mappa, ma abbiamo solo una indicazione degli altri "luoghi" dove possiamo andare; considereremo poi in seguito il problema se sia in generale possibile avere *una* (sola) mappa in uno spazio del genere. Ci troviamo quindi nella situazione in cui esploriamo lo spazio – ricordiamo che tale spazio è definito da una rete – avendo solo limitate informazioni su come tale spazio è fatto, ovvero solo informazioni sull'intorno del luogo di partenza. Per caratterizzare tale spazio, è possibile misurare le caratteristiche della rete che lo definisce tramite algoritmi che esplorano il WWW. Troviamo quindi una situazione totalmente diversa da quella cui siamo abituati: definendo la distanza fra due pagine come il numero di collegamenti che debbo attraversare, mi trovo che con pochi "click" posso passare da una pagina a qualsiasi altra pagina. Per fare un esempio, immaginiamo che il mio spazio sia fatto di 60000 pagine e che ogni pagina abbia collegamenti ad altre 3 pagine. Se la rete avesse una struttura simile al WWW, in una decina di click potrei arrivare da una pagina a qualsiasi altra. Se invece la rete fosse immersa in un mondo bidimensionale (immaginiamola ad esempio come un reticolo quadrato), per arrivare alle pagine più lontane ci vorrebbero centinaia di click. Quindi, da un lato ho una rete estremamente

navigabile (posso andare dove voglio in pochi click, mentre in una rete planare ci sono pagine che non raggiungerei mai perché mi stancherei di navigare prima), dall'altro è chiaro che senza una mappa rischio di perdermi o, peggio ancora, di non trovare mai quello che cerco.

Tecnicamente lo spazio di navigazione del WWW ha la struttura di una rete random[2] o, più precisamente, di una famiglia di reti random dette invarianti di scala[3]. Una caratteristica delle reti random è di essere spazi con una struttura localmente iperbolica, ovvero l'intorno di un oggetto ha un numero di vicini che cresce più rapidamente di quanto farebbe in uno spazio euclideo (ricordiamo che noi viviamo in uno spazio che è localmente euclideo). Questo significa che, nel momento in cui iniziamo a esplorare il WWW, ci stiamo muovendo in uno spazio alieno, totalmente diverso dallo spazio quasi bidimensionale nel quale ci siamo evoluti. Uno spazio iperbolico è uno spazio per il quale non abbiamo naturali mezzi o organi di orientamento, uno spazio nel quale affidarsi all'intuito per esplorare può portare a risultati opposti a quelli desiderati.

3. *Reti esplicite, reti implicite*

Il WWW è un esempio in cui la rete di navigazione è esplicita, ovvero i suoi oggetti (le pagine) contengono i collegamenti ad altre pagine. Analogo è il caso di Wikipedia, in cui è possibile andare da un argomento alle voci correlate; ultimamente anche le pubblicazioni scientifiche si stanno spostando online e contengono una bibliografia "ipertestuale" che collega direttamente agli articoli citati (se anch'essi pubblicati online). Partendo da una pagina nota, si può esplorare il suo circondario cliccando sui link; chiunque abbia un minimo di esperienza online, sa bene che

[2] Cfr. G. Caldarelli, M. Catanzaro, *A very short introduction to networks*, Oxford University Press, Oxford 2012.
[3] Cfr. A. Barabasi, A. Reka, *Emergence of scaling in random networks*, in «Science», 286, 1999, pp. 509-512.

la quantità di materiale che si trova a soli due click di distanza è già tale da richiedere giorni, settimane o anche mesi di lettura.

Su Internet i business model si basano invece spesso su reti *implicite*. In una rete implicita il legame è potenziale: ad esempio, la simiglianza fra i testi di due pagine o le caratteristiche di due prodotti creano un legame implicito che può essere sfruttato nel momento in cui lo esplicito e lo uso per navigare, per collegare le due entità. I legami impliciti creano quindi nuove reti, utili ad esempio per ordinare gli elementi a seconda dell'interesse di una persona: sono quelli che permettono di proporre a un acquirente oggetti simili a quello di cui è interessato (o crede di esserlo).

È interessante notare che il meccanismo delle reti implicite annulla la differenza tra utente osservatore e oggetti osservabili: poiché su Internet rimangono tracce sia delle operazioni fatte sia di chi le fa, è possibile stabilire collegamenti impliciti fra due persone in base a similitudini come una comune ideologia, un hobby, un interesse o delle preferenze negli acquisti. Queste reti implicite sono oggetto del marketing: è sfruttando la similitudine fra individui che gli algoritmi di presentazione propongono non solo oggetti simili a quello che si cerca, ma anche tutto ciò che hanno comprato persone con caratteristiche simili alle proprie.

Quello che è utile per il marketing è automaticamente utile alla politica: così come l'eldorado del marketing è la segmentazione del consumatore (ovvero la suddivisione in classi, per ognuna della quali si conosca prodotto e strategia di marketing ideale), la profilazione dell'elettorato è la pietra filosofale del politico: sapere cosa dire, come dire, a chi dirlo e quando. La complessità dello spazio in cui ci si muove, la possibilità di

avere più reti[4] e la nostra ignoranza dell'animo umano fanno sperare che la profilazione non possa essere un compito ben definito e per il quale esista una ricetta definitiva; ad esempio, il contributo di Cambridge Analytica alle elezioni americane, oltre a non essere misurabile, sembra più millantato che realtà, visto che la stessa company aveva lavorato durante le primarie per l'avversario di Trump. Allo stesso tempo, la prova quantitativa e su larga scala dell'esistenza delle "echo-chambers"[5], ovvero gruppi di utenti isolati in cui circolano e si amplificano posizioni ideologiche e convinzioni monolitiche, introduce un possibile vulnus nelle stesse basi liberali delle democrazie occidentali[6].

4. *Algoritmi di navigazione*

Appurato quindi che una quantità grande a piacere di dati è inutile se i dati non sono navigabili, e che la navigazione avviene in uno spazio astratto per il quale non siamo evolutivamente attrezzati, bisogna chiedersi quanto ciò che si trova possa dipendere dalla struttura di relazioni che si usa per la ricerca.

Un effetto banale – una volta che lo si è compreso – è quello per il quale eventuali oggetti iperconnessi (detti anche "hub") occorrono più spesso nelle ricerche di quelli isolati. Infatti, se ci si muove in maniera più o meno casuale, sono proprio gli hub quelli che con maggiore probabilità si finisce per "vedere" all'interno di una ricerca (o di un vagabondaggio). Ciò non è un problema se chi effettua la ricerca conosce il modo in cui le relazioni fra gli oggetti vengono create e *sceglie* il modo in cui si muove: se ad esempio sta effettuando una misura statistica,

[4] L'ambiguità nel definire cosa determini una similitudine implica la possibilità di molteplici reti.

[5] Cfr. M. Del Vicario et al., *The spreading of misinformation online*, https://www.pnas.org/content/113/3/554.

[6] Cfr. G. Pondrano Altavilla, A. Scala, *Ripensare i fondamenti della liberal-democrazia nell'era di Internet*, in «Micromega», 7, 2018, pp. 124-136.

deve solo tenere conto degli effetti dell'algoritmo con il quale recupera informazioni dai Big Data per evitare di introdurre errori sistematici.

Il problema sorge nel momento in cui l'ordinamento dei collegamenti possibili mi è dato dall'esterno, da un "oracolo" che cerca magari di pormeli nell'ordine che più mi faciliti. In tal caso, la ricerca viene assolutamente influenzata dall'ordine di presentazione: ciò avviene non solo per un algoritmo automatico, che comunque risente dell'ordine in cui vengono presentate le opzioni (vedi la ricca letteratura sui paradossi delle preferenze), ma a maggior ragione per un essere umano, che come sanno bene i sondaggisti può essere "guidato" a risposte diverse a seconda di come gli vengano poste le domande.

Non bisogna pensare che il pericolo derivi da un complotto di colui[7] che gestisce gli algoritmi di esplorazione: basta solo che l'algoritmo sia adattativo e che consideri un oggetto più importante quante più volte esso viene raggiunto/osservato, perché – per un meccanismo di rinforzo naturale[8] – vengano a determinarsi degli hub, degli oggetti collegati con quasi tutti gli altri. È chiaro a questo punto che, anche se gli hub fossero stati generati a caso nelle fasi iniziali di addestramento, l'algoritmo finirà per riproporli sempre nelle ricerche.

Questo meccanismo di "rafforzamento" tipico di molti algoritmi di ricerca (se una cosa ti è piaciuta, te la ripropongo) è stato spesso accusato di essere alla base della formazione delle "echo-chambers" sui social media e di essere quindi il responsabile – insieme al (o più del) bias di conferma[9] – della polarizzazione delle opinioni che inficia il processo deliberativo democratico. Una delle possibili soluzioni proposte è stata quella di cambiare

[7] Sarebbe comunque interessante poter analizzare l'evoluzione degli algoritmi di ricerca/presentazione dal punto di vista degli operatori.

[8] Cfr. H.A. Simon, *On a class of skew distribution functions,* in «Biometrika», 42, 1955, p. 425.

[9] Per bias di conferma si intende la tendenza delle persone di ricercare informazioni in accordo con la propria visione del mondo.

gli algoritmi di esplorazione in modo che la presentazione fosse più "democratica" ed esponesse agli utenti dei social media informazioni più diversificate: detto in parole povere, un algoritmo che facesse muovere gli utenti in maniera casuale sulla rete dei Big Data presenti nei social. L'esperimento non ha avuto – a questo punto direi ovviamente – molto successo: una esplorazione casuale non può che generare una successione incoerente di oggetti, mentre noi nelle nostre ricerche per lo più abbiamo scopi razionali ben determinati (ad esempio se stiamo usando lo strumento per lavoro), o scopi ricreativi. Una esplorazione casuale non può che raggiungere allegramente l'obiettivo di vanificare entrambi gli scopi della ricerca.

Abbiamo a questo punto un paradosso: un algoritmo che non generi bias produce ricerche non interessanti, mentre un algoritmo che soddisfa i nostri criteri di ricerca non può che accrescere i nostri bias.

Vorrei infine chiudere con una interessante proposta di ricerca: data la ricchezza combinatoria di un set di Big Data, quante realtà (anche incompatibili fra di loro) posso costruirvi? È possibile, analizzando i comportamenti e le scelte delle persone, ancorché all'interno dello spazio limitato degli osservabili della rete, osservare come esse si creano una visione/interpretazione/modello del mondo virtuale? Forse con Internet abbiamo per la prima volta la possibilità di osservare in un mondo "oggettivo" il costruttivismo all'opera[10].

[10] Per il costruttivismo, cfr. P. Watzlawick, *La realtà inventata. Contributi al costruttivismo,* tr. it. Feltrinelli, Milano 2006.

Figure

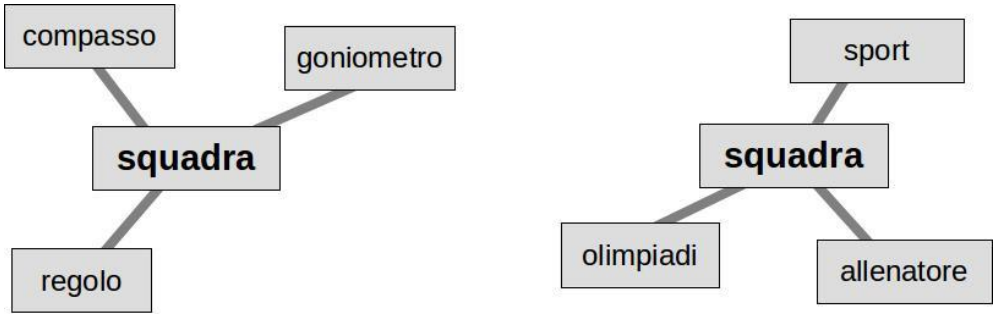

Fig. 1 : La polisemia è un esempio tipico in cui il significato di un dato dipende dal contesto (ovvero dagli oggetti con i quali è collegato).

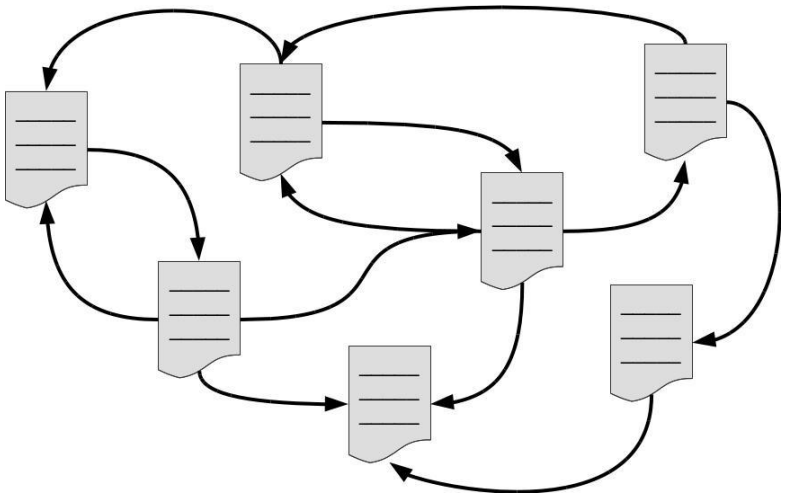

Fig. 2 : Gli oggetti su Internet formano delle reti. Ad esempio, le pagine web contengono rimandi ad altre pagine web, permettendo una fruizione "navigabile" di dati, notizie e informazioni.

ANTONIO SCALA è coordinatore di APPLICO Lab sulle applicazioni della complessità e responsabile per il progetto AMOFI sull'analisi dei flussi informativi nei social media online. Svolge attività di ricerca presso il CNR, Istituto Sistemi Complessi e il LIMS – the London Institute for Mathematical Sciences
antonio.scala@isc.cnr.it