Casualità, causalità e Machine Learning nel contenimento epidemico

It is a mistake to think you can solve any major problem just with potatoes.
—— **Douglas Adams, *Life, the Universe and Everything***

Alfredo Braunstein *Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino*
Luca Dall'Asta *Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino*
Alessandro Ingrosso *The Abdus Salam International Centre for Theoretical Physics Strada Costiera 11, 34151 Trieste*

Le difficoltà principali nel contenimento di Covid-19 sono dovute ad alcune caratteristiche dell'infezione. La trasmissione del virus SARS-CoV-2 da parte di individui infetti avviene principalmente tramite l'emissione di goccioline respiratorie (*droplets*). Gli individui iniziano ad essere infettivi poco tempo dopo il contagio e sono molto spesso asintomatici[1]. Inoltre, nei casi in cui vi sia sviluppo di sintomi, questo avviene in un periodo dai 2 ai 14 giorni dopo l'infezione [2], permettendo quindi al virus di diffondersi per diversi giorni senza essere scoperto. Per poter controllare un focolaio epidemico è dunque fondamentale poter isolare i casi infetti ma privi di sintomi (asintomatici e presintomatici).

[1] La frazione di individui asintomatici è molto difficile da stimare, ma potrebbe aggirarsi attorno al 50%-75% [1]

Consideriamo il seguente esempio minimale di nascita di focolaio nell'Italia *post-lockdown*, in un periodo vicino a quello della pubblicazione di questo manoscritto:

- 19/7/2020, 19:00 - Al termine di una giornata di lavoro, Daniele e Carlo si incontrano per un breve aperitivo prima di rientrare alle rispettive abitazioni per cena.
- 20/7/2020, 16:20 - Ernesto ed il suo amico Daniele esultano abbracciandosi mentre guardano la partita della loro squadra del cuore.
- 22/7/2020, 19:45 - Amanda si prepara ad uscire per andare a trovare sua nonna Benedetta. Suo fratello Ernesto non andrà con lei: gliel'aveva preannunciato a pranzo tre giorni prima (il 19/7).
- 22/7/2020, 11:00 - Daniele si sveglia dopo una notte terribile, mal di testa fortissimo e

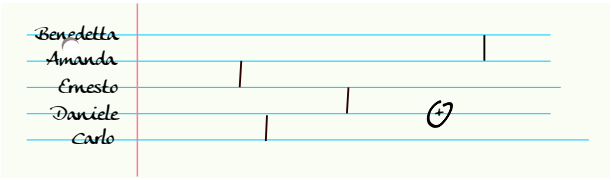

Figura 1: *Diagramma dei contatti: il tempo scorre da sinistra a destra, i segmenti verticali denotano contatti tra gli individui corrispondenti ai loro estremi. Il segno ⊕ corrisponde al momento in cui è stato prelevato il tampone, poi risultato positivo.*

febbre alta. Sarà qualcosa che ha mangiato la sera prima?

- 22/7/2020, 18:00 - Daniele decide di recarsi al pronto soccorso. Viene immediatamente sottoposto al tampone per SARS-CoV-2. Il risultato arriva dopo 40 minuti: positivo.

Il lavoro investigativo

Il giorno dopo, con in mano il risultato del test, un operatore di tracciamento (*contact tracer*) prova a ricostruire la dinamica del focolaio tramite il cosiddetto *tracciamento manuale*. Intervista Daniele, ricostruendo con lui i suoi contatti degli ultimi giorni ed determinando che ha avuto incontri che possono aver portato ad un contagio con i suoi amici Carlo ed Ernesto. Convoca dunque entrambi per effettuare dei test e (in caso di esito positivo) successiva intervista. Questo lavoro è purtroppo per sua natura lento e laborioso, e può protrarsi anche per diversi giorni. In particolare, è difficile immaginare che l'informazione possa arrivare ad Amanda in tempo per dissuaderla dal visitare Benedetta.

Supponiamo però che l'operatore tenti di portarsi avanti, svolgendo le interviste telefonicamente, senza attendere i risultati dei test. Con l'aiuto di semplice carta e penna, potrebbe poi disegnare un diagramma che rappresenta gli individui ed i loro contatti temporali, come quello riportato in Figura 1. Lo stesso operatore potrebbe, successivamente, considerare alcuni degli scenari possibili, corrispondenti a potenziali catene di infezione, a partire dai contatti avvenuti nel breve lasso di tempo costituito dai tre giorni precedenti al momento in cui è stato appreso il risultato del tampone. Per semplicità, assumerà che solo uno degli individui in questione fosse infetto il giorno 19 (siamo, dopotutto, in uno scenario *post-lockdown*, in cui i casi di infezione sono rari) e chiamerà questo individuo infetto il Paziente Zero (P0). Guardando il diagramma, l'operatore potrà subito eliminare l'ipotesi che Benedetta sia stata il paziente zero. Ecco dunque le possibilità rimaste:

Possibilità 1: Amanda è il P0: il 19 luglio contagia Ernesto, il quale a sua volta contagia Daniele durante la cena a casa sua;

Possibilità 2: Daniele è il P0;

Possibilità 3: Ernesto è il P0: il 20 luglio contagia Daniele;

Possibilità 4: Carlo è il P0: il 19 luglio contagia Daniele.

Queste possibilità potrebbero essere rappresentate come in Figura 2 (curandosi di evidenziare in ognuno dei casi tutti gli individui che potrebbero risultare infetti).

In questo piccolo esempio, tutti e cinque gli individui coinvolti possono essere stati infettati, e quattro di loro potrebbero aver ricoperto il ruolo di P0. Ogni scenario presenta possibili evoluzioni nefaste per gli attori in gioco; in particolare, Benedetta, che appartiene ad una categoria di soggetti a rischio, può essersi contagiata nell'incontro con Amanda nelle possibilità 1 e 3.

Ci troviamo in una impasse: il lavoro investigativo progressivo produce risultati puramente speculativi - per costruzione pessimistici - che possono solo fornire un'idea generica dell'estensione del focolaio. Per sua natura, inoltre, tale processo è troppo lento e non permette azioni di contenimento sufficientemente tempestive (in particolare in situazioni in cui il numero di persone coinvolte fosse significativamente maggiore rispetto all'esempio qui considerato).

Il tracciamento automatico

Un'alternativa più agile recentemente proposta è `quella del tracciamento automatico` dei contatti, basata sull'assunzione che ognuno degli individui abbia previamente installato un'applicazione (*app*) di tracciamento sul proprio *smartphone* (in Italia, `Immuni` [3]). Il *software* utilizza il segnale radio `Bluetooth Low Energy` per fornire un'ipotesi di contatto tramite una misura

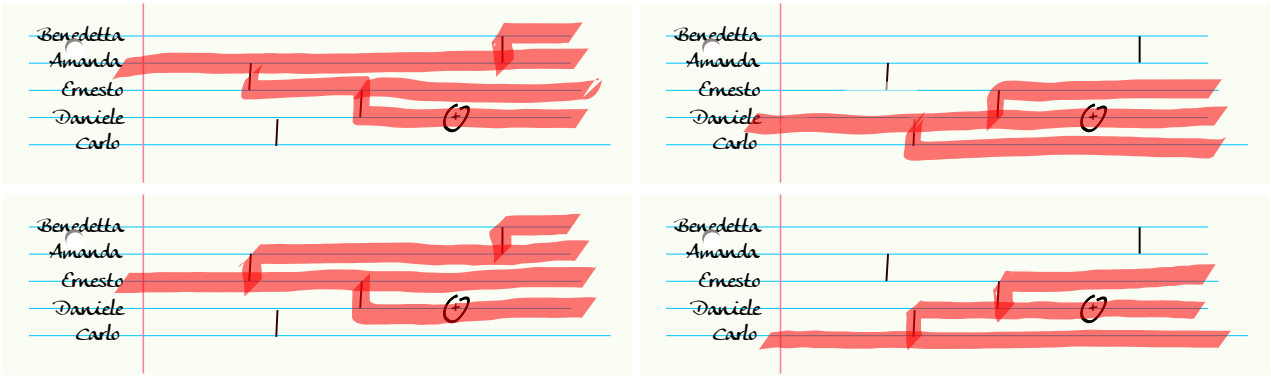

Figura 2: *Quattro delle possibili storie epidemiche (sopra: possibilità 1 e 2. sotto: possibilità 3 e 4.) che possono spiegare l'infezione di Daniele a partire dai quattro possibili pazienti zero.*

approssimata della distanza e della durata caratterizzanti la relazione di prossimità tra due dispositivi. Questo sistema permette immediatamente di evitare tutto l'iter di interviste telefoniche, avendo anche il vantaggio di rilevare contatti tra sconosciuti (ad esempio tra cliente e commerciante). Inoltre, il risultato di un test positivo viene inserito nel sistema dall'operatore, per cui Ernesto potrebbe ricevere immediatamente[2] una `notifica di esposizione`, cioè di un possibile contatto rischioso con una persona potenzialmente positiva. Parliamo di "possibile contatto" perché le caratteristiche dell'`hardware` non garantiscono grande precisione nel rilevamento della prossimità e non permettono di determinare se c'è stato veramente un contatto, e "potenzialmente positiva" semplicemente perché non è possibile sapere se Daniele fosse già positivo al momento dell'incontro. Grazie a questa notifica, Ernesto potrebbe contattare immediatamente la sua famiglia, ed Amanda potrebbe decidere, per precauzione, di non fare visita a sua nonna.

C'è un rovescio della medaglia, però: il numero di falsi allarmi generato da questo sistema può essere enorme. Se ogni individuo a cui è arrivata una segnalazione chiedesse di sottoporsi ad un test, il numero di test necessari (e la rapidità con cui questi dovrebbero poi essere processati) crescerebbe velocemente con il numero di contatti, fino a rendere il sistema stesso sostanzialmente inutilizzabile. Dovendo restringere il numero di persone da sottoporre al test, è necessario avere a disposizione un ordine o classifica di rischio d'infezione, così da sottoporre per primi al test gli individui più a rischio e solo in un secondo tempo gli altri, allorché vi sia sufficiente disponibilità di risorse per svolgere ed analizzare tutti i test.

[2]Si noti che questa descrizione non rispecchia il funzionamento di Immuni [3]. In particolare, le notifiche di esposizione in Immuni non sono immediate.

Il ruolo fondamentale della casualità

La trasmissione del virus in corrispondenza di un contatto non è un evento deterministico, così che ognuna di queste possibilità per il P0 presenterà molteplici ramificazioni: cosa è successo ad Amanda nel suo incontro con Ernesto nella Possibilità 3?

La vera natura della stima del rischio epidemico (in inglese *epidemic risk assessment*) è in realtà probabilistica. I modelli matematici considerati più accurati sono probabilistici, racchiudendo in un semplice evento stocastico (per esempio la possibile trasmissione del virus durante un contatto tra due persone) un'infinità di variabili, la maggior parte impossibili da determinare (come alcune caratteristiche del contatto e del sistema immunitario del potenziale ricevente, l'abbigliamento o eventuale protezione delle persone coinvolte, ecc.) e molte di esse ancora ignote (come sono tuttora molte delle caratteristiche della diffusione del virus SARS-CoV-2).

L'approccio più utilizzato nella letteratura scientifica per descrivere matematicamente la diffusione di Covid-19 (e più in generale di molte altre epidemie) si basa su modelli epidemici a compartimenti. In questi modelli, ogni individuo è caratterizzato ad ogni istante di tempo da uno stato interno appartenente ad un insieme finito X di possibilità (compartimenti). Nel

caso più semplice, l'individuo può essere nello stato S (sano) o I (infetto). È spesso ragionevole aggiungere almeno un terzo stato R (rimosso), per tenere conto sia della risposta immunitaria, che può rendere l'individuo immune al contagio, sia di un suo potenziale decesso; in entrambi i casi, l'individuo non è più parte del processo di diffusione del virus.

Denotiamo con $x_i^t \in X$ lo stato dell'individuo i al tempo t (assumeremo qui per semplicità t discreto, per esempio nel caso in cui si consideri una risoluzione temporale su scala giornaliera). Se denotiamo con $\partial i(t)$ l'insieme degli individui che hanno avuto contatto con i al tempo t e la configurazione dei loro stati con $x_{\partial i}^{t-1} = \{x_j^{t-1}\}_{j\in\partial i(t)}$, possiamo scrivere la distribuzione di probabilità per le traiettorie collettive del sistema $\mathbf{x} = \mathbf{x}^{0:T}$ nei tempi $0, \dots, T$ dove $\mathbf{x}^{0:t} = \{x_i^s\}_{i=1,\dots,N}^{s=0,\dots,t}$ come:

$$p(\mathbf{x}) = \prod_i p\left(x_i^0\right) \prod_{t=0}^{T-1} p\left(x_i^{t+1} | x_{\partial i}^{0:t}, x_i^{0:t}\right) , \quad (1)$$

in cui $p\left(x_i^{t+1} | x_{\partial i}^{0:t}, x_i^{0:t}\right)$ è la probabilità condizionata che l'individuo i si trovi in uno stato x_i^{t+1} al tempo $t+1$ dato lo stato suo e di tutti gli individui con cui ha avuto contatto fino al tempo t, mentre $p\left(\mathbf{x}^0\right) = \prod_i p\left(x_i^0\right)$ è la distribuzione di probabilità dello stato iniziale $\mathbf{x}^0$ del processo. Possiamo assumere di avere per quest'ultima una forma fattorizzata sugli individui, dal momento che i rari casi di infezione (pazienti zero) che danno inizio a focolai epidemici hanno solitamente origine diversa (per esempio, vengono importati da un'altra comunità) e possono pertanto essere considerati eventi indipendenti.

Prediamo come esempio il modello SIR più semplice (che chiameremo «SIR standard»), che è Markoviano in quanto lo stato di un individuo a tempo $t+1$ dipende solamente dallo stato del sistema a tempo t. Un individuo nello stato I può diventare R con probabilità μ (potremmo rappresentare concisamente questa situazione con $I \xrightarrow{\mu} R$). Inoltre, ogni individuo nello stato I può contagiare indipendentemente con probabilità λ ogni altro individuo nello stato S con cui ha un contatto (rappresentato come $IS \xrightarrow{\lambda} II$). Questo ci fornisce le seguenti espressioni

$$p(x_i^{t+1} = S | x_{\partial i}^{0:t}, x_i^{0:t}) = \delta_{x_i^t,S} \prod_{j\in\partial i(t)} (1 - \lambda\delta_{x_i^t,I})$$
$$p(x_i^{t+1} = R | x_{\partial i}^{0:t}, x_i^{0:t}) = \delta_{x_i^t,R} + \mu\delta_{x_i^t,I},$$

in cui $\delta_{a,b} = 1$ se $a = b$, e zero altrimenti. La probabilità di essere nello stato I si può ricavare analogamente o per normalizzazione. L'interpretazione di queste espressioni è semplice: nella prima equazione, per essere S al tempo $t+1$, l'individuo deve trovarsi già nello stato S al tempo t e non essere contagiato a causa dei contatti tra t e $t+1$. Nella seconda, per essere R al tempo $t+1$, o l'individuo lo era già al tempo t, oppure era I ma è stato rimosso con probabilità μ.

Il metodo Monte Carlo

Un'idea per la determinazione stocastica del rischio è quella adottata da `ViraTrace` [4], alla base dell'*app* ufficiale di tracciamento dei contatti attualmente in uso in India [5]. L'idea è semplice ed allettante. La diffusione del virus avviene nell'ombra, nascosta ai nostri occhi. Quindi, perché non replicarla in maniera visibile, usando una diffusione sintetica (che riproduca il più possibile le caratteristiche infettive del virus)? Secondo questo approccio, una volta ottenuto il risultato del test che riveli la positività di un individuo, la sua *app* inizierebbe a trasmettere l'infezione sintetica ad altre *app* di individui con cui è entrato in contatto dopo il prelievo. Data la natura stocastica del processo, ovviamente, le infezioni sintetiche possono seguire storie diverse, discostandosi tra loro e, quel che è più grave, dall'evoluzione dell'epidemia reale. La propagazione di un numero sufficientemente grande di epidemie sintetiche indipendenti permette tuttavia di esplorare gli scenari più probabili ed assegnare una probabilità di infezione ad ogni individuo (queste propagazioni possono avvenire portando avanti in parallelo $m > 1$ sistemi epidemici). A tutti gli effetti, l'approccio equivale al metodo di campionamento Monte Carlo per calcolare distribuzioni marginali $p(x_i^t)$ dalla distribuzione in (1), in cui gli individui per cui l'infezione è nota saranno considerati dei P0 al tempo del prelievo del test, e per gli individui che hanno avuto un risultato di un test negativo si assume un'immu-

nità artificiale nel periodo precedente al test (per esempio cancellando tutti i loro contatti in quel periodo). Questo tipo di campionamento non presenta particolari difficoltà e permette di produrre campioni indipendenti solo riproducendo la dinamica. Avendo a disposizione m campioni $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ distribuiti secondo (1), calcoleremo in modo approssimato il rischio infettivo dell'individuo i al tempo t come

$$p(x_i^t = I) \approx \frac{1}{m} \sum_{\mu=1}^{m} \delta_{x_i^{(\mu),t},I} \tag{2}$$

cioè la frazione di campioni in cui i risulta infetto. Un individuo che è stato sufficientemente a contatto con uno o più individui infetti noti risulterà spesso infetto nei campioni ottenuti mediante la propagazione sintetica, cosicchè anche i suoi contatti futuri saranno relativamente frequentemente infetti, e via dicendo. L'*app* di un individuo potrebbe allora quantificare questo rischio tramite una probabilità e, se la stima dovesse superare una certa soglia, potrebbe allertare il sistema sanitario nazionale oppure semplicemente suggerire al proprietario di farlo.

Questa potrebbe sembrare una soluzione non medica quasi ideale, che permetterebbe di contenere un focolaio sia di Covid-19 che di altre malattie infettive con caratteristiche simili. Purtroppo, come vedremo in seguito, ci sono diverse problematiche associate a questa soluzione. In particolare, il metodo può solo inferire infezioni da una sorgente infetta avvenute a seguito del prelievo del tampone. Dal momento che l'individuo dovrebbe essere sottoposto a quarantena immediata alla notifica di un esito positivo, l'efficacia del metodo si vanifica se l'analisi viene effettuata in tempi rapidi. In realtà, i problemi sono conseguenza del fatto che stiamo calcolando i marginali della distribuzione sbagliata!

La stima a posteriori

Nel contesto probabilistico appena descritto abbiamo sorvolato su un ingrediente fondamentale: come vengono tenuti in considerazione i risultati dei *test*? In termini Bayesiani, il risultato dei *test* ci fornisce un'evidenza $\mathcal{O}$ (ad esempio, l'individuo i è infetto al tempo t e l'individuo j è sano al tempo t'). In questo contesto, l'equazione (1) descrive una distribuzione di probabilità *a priori*, cioè precedente all'acquisizione dell'evidenza; noi siamo però interessati alla distribuzione a posteriori di $\mathbf{x} = \mathbf{x}^{0:T}$:

$$p\left(\mathbf{x}|\mathcal{O}\right) = p\left(\mathbf{x}\right) p\left(\mathcal{O}|\mathbf{x}\right) p(\mathcal{O})^{-1} \tag{3}$$

Oltre al termine in (1), ne abbiamo dunque altri due. Il termine $p(\mathcal{O})$ non è per il momento rilevante, essendo una costante che non dipende dallo stato del sistema (sarà però importante più avanti). Il termine $p\left(\mathcal{O}|\mathbf{x}\right)$ è legato alle caratteristiche dei test: per un test ideale senza difetti, è deterministico (prende valori 0 e 1) ed agisce semplicemente come un filtro, cancellando tutte quelle traiettorie collettive incompatibili con i risultati dei test; nel caso generale, attribuirà invece solo un peso minore alle traiettorie incompatibili. In ogni caso, è ragionevole assumere che i risultati dei test siano indipendenti per individui diversi, così che

$$p\left(\mathcal{O}|\mathbf{x}\right) = \prod_i p\left(\mathcal{O}_i|x_i\right) \tag{4}$$

Possiamo notare inoltre come, in presenza di evidenza di infezione (almeno un test con risultato positivo), l'ipotesi di indipendenza dello stato iniziale copra anche il caso di singolo P0: se la probabilità di essere infetto a tempo 0 tende a zero, la misura di probabilità a posteriori tenderà ad essere completamente concentrata su traiettorie con esattamente un singolo P0.

Difficoltà computazionale

Quanto è difficile stimare il rischio epidemico? In principio, basterebbe elencare tutte le possibili storie epidemiche $\mathbf{x}$ compatibili con i risultati dei test. Ad ognuna di esse è associata una probabilità secondo l'Eq. (3) (una di queste storie è quella vera, da cui la somma di tutte queste probabilità è esattamente 1). Per calcolare la probabilità a posteriori che un dato individuo sia stato infettato (o sia infetto), basterebbe mediare rispetto a questa misura:

$$p(x_i^t = I|\mathcal{O}) = \frac{\sum_{\mathbf{x}} p(\mathcal{O}|\mathbf{x}) p(\mathbf{x}) \delta_{x_i^t,I}}{\sum_{\mathbf{x}} p(\mathcal{O}|\mathbf{x}) p(\mathbf{x})} \tag{5}$$

Purtroppo, il numero di storie epidemiche $\mathbf{x}$ cresce esponenzialmente con il numero di individui,

rendendo questa soluzione proibitiva. Forse, però, non tutto è perduto: chi ci assicura che non esista una soluzione più intelligente rispetto ad elencare tutte le possibili storie epidemiche? La questione di determinare se per un dato problema esista una soluzione algoritmica praticabile è fondamentale in molti ambiti della scienza. Per la maggior parte dei problemi considerati difficili, non esiste purtroppo una dimostrazione matematica di questa difficoltà. Esiste però una vasta classe di problemi, chiamati $\mathcal{NP}$*-complete*, composta di tanti problemi notevoli (tra cui, per esempio il Problema del Commesso Viaggiatore) a cui non si è mai giunti ad una soluzione algoritmica soddisfacente (i.e. che non sia esponenzialmente lenta!), che però sono intimamente legati: se venisse scoperta (o fosse già stata scoperta) una soluzione soddisfacente per uno di essi, allora questa soluzione potrebbe essere adattata per risolvere ciascuno di questi problemi! Dimostrare che un problema appartiene a questa classe ci fornisce automaticamente una dimostrazione, se non matematica, almeno storica, della difficoltà della sua soluzione. Una soluzione efficiente di uno di questi problemi – oltre ad essere un'impresa scientifica straordinaria con enormi implicazioni – probabilmente non potrebbe rimanere nascosta a lungo: ad esempio, permetterebbe di generare un'immensa quantità di BitCoin con relativa facilità, o di superare senza difficoltà le barriere crittografiche che difendono i sistemi bancari online (il sistema RSA). Siamo dunque abbastanza convinti del fatto che una soluzione del genere non esista.

Il problema della stima del rischio epidemico non è $\mathcal{NP}$*-complete*, ma è relativamente facile dimostrare che essere capace di risolverlo efficientemente fornirebbe una soluzione efficiente anche per un problema $\mathcal{NP}$*-complete* e quindi, per tutti gli elementi della classe! Abbozzeremo qui informalmente tale dimostrazione. Il problema a cui facciamo riferimento è chiamato *Unweighted Minimum Steiner Tree* (UMST) [6]: si tratta di trovare, dato un grafo $G = (V, E)$ (dove V è l'insieme di vertici e E è l'insieme di archi) ed un sottoinsieme $Q \subset V$ di vertici (chiamati `terminali`), un sotto-albero di G che connetta tutti i terminali utilizzando un numero di archi minore di una costante predeterminata. È relativamente semplice dimostrare matematicamente (in grande generalità rispetto al modello epidemico adottato) che, per una probabilità di contagio λ sufficientemente piccola, considerando il grafo formato da tutti gli individui e dai loro contatti (che assumiamo per semplicità ripetersi nel tempo) ed un sottoinsieme Q di individui con test positivo, le storie epidemiche più probabili corrisponderanno a quelle in cui il numero di individui contagiati k è minimo e la probabilità di queste storie sarà proporzionale a λ^{k-1} (corrispondente a $k-1$ contagi). Qualunque soluzione con un numero di contagi maggiore conterrà almeno un fattore λ in più, e sarà dunque molto meno probabile. Se λ è sufficientemente piccolo, la probabilità di questo insieme di soluzioni sarà vicina ad 1! Se avessimo a disposizione un `oracolo` algoritmico efficiente per decidere se la probabilità di un individuo di essere stato infettato sia maggiore di una costante, potremmo dunque usarlo per risolvere il problema dell'UMST. Si potrebbe obiettare che non c'è ragione per cui λ debba essere così piccolo (valori ragionevoli di λ per Covid-19 potrebbero ad esempio aggirarsi tra 10^{-1} e 10^{-2} a seconda del tipo e durata del contatto) e che forse la difficoltà sia associata esclusivamente a valori estremamente piccoli e non realistici. Sostituendo però ogni arco con una catena di r archi, la probabilità che un'infezione attraversi tutta la catena risulterà un multiplo di λ^r, numero che si può rendere piccolo a piacere aumentando r. La soluzione del problema epidemico in questo grafo aumentato ci fornirebbe una soluzione del UMST nel grafo originale. In sintesi, dunque, se sapessimo risolvere efficientemente il problema della stima del rischio su un qualunque (qui, la parola "qualunque" è fondamentale) grafo di contatti, potremmo risolvere altrettanto efficientemente diversi dei problemi computazionali più difficili della storia. Questa conclusione, purtroppo, lascia ben intendere la difficoltà computazionale del problema. Ma non ci dobbiamo scoraggiare: la rete di contatti interpersonale non è una rete completamente arbitraria e non è stata disegnata con malvagità per rendere difficile la soluzione. Possiamo ancora sperare di trovare soluzioni algoritmiche, magari approssimate, che funzionino in modo accettabile nei casi reali.

L'effetto dell'evidenza

È interessante a questo punto tornare al nostro esempio per mostrare come anche esiti negativi dei test possano modificare la nostra stima a posteriori dello stato infettivo di un individuo, in modo non completamente intuitivo. Per semplicità dei calcoli, assumeremo un modello SIR standard con $\mu = 0$, i.e senza lo stato R, e con probabilità di contagio λ in ogni contatto. Calcoliamo per prima cosa la probabilità a posteriori che Benedetta sia infetta, sapendo che il test di Daniele è risultato positivo. Per far ciò, è sufficiente partire dalle storie di infezione precedenti e considerare le catene infettive che contengano Benedetta. Dopo aver diviso per la probabilità che Daniele sia infetto, otteniamo la probabilità condizionata che cerchiamo:

$$p\left(x_B^{t_1} = I | x_D^{t_2} = I\right) = \frac{2\lambda^3}{1 + \lambda^2 + 2\lambda} \qquad (6)$$

Cosa succederebbe se Carlo, dopo aver incontrato Daniele, venisse testato con risultato negativo? Seguendo la medesima logica, la probabilità a posteriori che Benedetta sia infetta è la seguente:

$$p\left(x_B^{t_1} = I | x_D^{t_2} = I, x_C^{t_3} = S\right) = \frac{2\lambda^3}{1 + \lambda^2} \qquad (7)$$

Vediamo come il risultato negativo del test eseguito su Carlo costituisca evidenza a favore del contagio subito da Benedetta! La differenza è di 2λ nel denominatore, che corrisponde alle due situazioni impossibili evidenziate nella Figura 3. Essendo λ positivo, la quantità in (7) sarà sempre maggiore di quella in (6): il fatto che Carlo non sia infetto aumenta il peso a posteriori delle storie infettive in cui Amanda o Ernesto sono i P0 e che dunque Benedetta possa essere infettata da Amanda. Si noti come questo fenomeno non possa essere assolutamente catturato dalla dinamica Monte Carlo come descritta precedentemente [4]. Quel modo di tenere conto dell'informazione acquisita permette infatti solo di influenzare gli eventi che seguono causalmente l'evidenza stessa, non consente invece di porre alcun condizionamento sulle relazioni di causalità che l'hanno preceduta temporalmente e che hanno realmente portato a tale evidenza. Così facendo, non si è in grado di migliorare la stima della probabilità di eventi correlati alla possibile

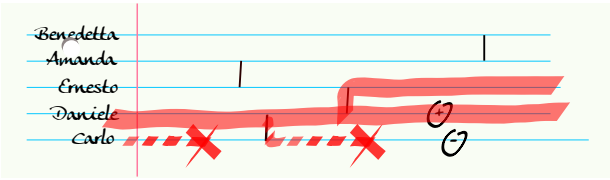

Figura 3: *L'evidenza di un test negativo di Carlo aumenta la probabilità a posteriori dell'evento che Benedetta sia infetta: confrontando* (7) *con* (6)*, due addendi λ risultano eliminati dal denominatore, corrispondenti agli eventi "Carlo è il P0 e contagia Daniele" e "Daniele è il P0 e contagia Carlo". Le infezioni di Carlo e di Benedetta sono dunque anti-correlate.*

infezione di un individuo, ma non strettamente causati da questa. Nell'esempio in questione, l'informazione addizionale del test negativo porterebbe nell'approccio Monte Carlo solo ad una diminuzione del livello di infezione degli altri individui, non potrebbe mai causarne l'aumento. Per osservare tale fenomeno, l'informazione aggiunta dall'evidenza del test negativo ha dovuto viaggiare indietro nel tempo (limitando l'insieme di possibili P0) per poi ritornare aggiornando l'infezione di Benedetta. Qualunque metodo che non permetta questo tipo di trasporto temporale dell'informazione a doppio senso non potrà mai catturare correttamente il fenomeno.

Aggiustamenti al Monte Carlo

È possibile aggiustare il metodo Monte Carlo per campionare dalla distribuzione a posteriori (3)? Una possibilità sarebbe semplicemente quella di ripesare i campioni in (2) con il termine mancante in (1) ma presente in (3):

$$p(x_i^t = I | \mathcal{O}) \approx \frac{\sum_{\mu=1}^m \delta_{x_i^{(\mu),t}, I} p(\mathcal{O}|\mathbf{x}^{(\mu)})}{\sum_{\mu=1}^m p(\mathcal{O}|\mathbf{x}^{(\mu)})} \qquad (8)$$

Questa strategia, anche se corretta, purtroppo non porta a buoni risultati perché il campionamento diventa estremamente inefficiente. Per esempio, assumendo test ideali, il termine mancante sarà diverso da zero con una frequenza esponenzialmente piccola; richiedendo un numero di campioni esponenzialmente grande per poter ottenere una stima ragionevole.

Un'altra possibilità consiste nel campionare direttamente dalla (3). In [7] viene proposto un

campionamento di Gibbs, in cui si campiona iterativamente la traiettoria di ogni singolo individuo $x_i^{0:T}$ condizionata allo stato del resto del sistema. Gli autori dello studio sostengono di ottenere buoni risultati fino a sistemi di dimensioni corrispondenti ad una piccola città (10^4 individui). Tuttavia, le proprietà di convergenza potrebbero degradarsi notevolmente al crescere della dimensione del sistema, come spesso accade nel campionamento Monte Carlo di distribuzioni complicate.

Algoritmi ispirati alla Meccanica Statistica

Considerando la dipendenza rispetto allo stato del sistema $\mathbf{x} = \mathbf{x}^{0:T}$, la distribuzione a posteriori in Eq.(3) può essere riscritta nella seguente forma:

$$p\left(\mathbf{x}|\mathcal{O}\right) = \frac{1}{Z} e^{-H(\mathbf{x})} \tag{9}$$

in cui $Z = p(\mathcal{O})$ e

$$H\left(\mathbf{x}\right) = -\log p\left(\mathbf{x}, \mathcal{O}\right) = \sum_i H_i\left(\mathbf{x}_i, \mathbf{x}_{\partial i}\right) \tag{10}$$

dove $H_i\left(\mathbf{x}_i, \mathbf{x}_{\partial i}\right) = -\log p(\mathcal{O}_i|\mathbf{x}_i) - \log p(x_i^0) - \log \prod_{t=0}^{T-1} p\left(x_i^{t+1}|x_{\partial i}^{0:t}, x_i^{0:t}\right)$. L'espressione (9) rappresenta la distribuzione di probabilità in un *ensemble* canonico (distribuzione di Boltzmann) per un modello meccanico-statistico in cui le variabili sono le traiettorie epidemiche $\{\mathbf{x}_i\}$ dei singoli individui. Ad ogni storia epidemica $\mathbf{x}$ è associata un'energia $H(\mathbf{x})$, costituita da contributi $H_i\left(\mathbf{x}_i, \mathbf{x}_{\partial i}\right)$, locali nel grafo dei contatti, cioè tali da coinvolgere la traiettoria epidemica $\mathbf{x}_i$ di un individuo i e dei suoi vicini ∂i (tutti e soli gli individui che sono venuti in contatto con esso nell'intervallo temporale preso in esame). La costante di normalizzazione

$$Z = \sum_{\mathbf{x}} e^{-H(\mathbf{x})} \tag{11}$$

rappresenta la funzione di partizione del modello. Esistono vari metodi per stimare, almeno in modo approssimato, sia la funzione di partizione che medie rispetto alla distribuzione stessa (comprese le sue distribuzioni marginali $p(\mathbf{x}_i|\mathcal{O})$, a cui siamo interessati). In particolare, mediante metodi variazionali di campo medio, si possono ottenere numericamente espressioni approssimate per queste quantità. Una possibilità è la cosiddetta approssimazione di Bethe, il cui associato schema computazionale viene chiamato algoritmo di *Belief Propagation* (BP) [8]. L'algoritmo di BP risolve una equazione di punto fisso per un vettore di dimensione elevata tramite iterazioni. Una delle caratteristiche di BP è che può essere implementato tramite dei calcoli locali, rendendo la soluzione particolarmente appetibile in un contesto distribuito: l'app di ogni individuo potrebbe in principio effettuare una parte del calcolo in cooperazione con quelle degli individui con cui è stato in contatto, evitando dunque di scambiare informazione con un server centrale. Vedremo in seguito un confronto tra diverse strategie.

Modelli ad agente

L'utilizzo pratico di procedure d'intervento basate su un approccio inferenziale richiede una validazione di tali metodi su modelli realistici di diffusione epidemica. Ad oggi, sono disponibili diversi modelli ad agenti per la simulazione su larga scala della diffusione di patogeni come il virus SARS-CoV-2 [9, 10, 11, 12, 13]. In questi modelli, ad ogni individuo è associato uno stato (p.es. Sano-Infetto-Rimosso) che evolve nel tempo, con una risoluzione temporale tipicamente giornaliera (ma alcuni modelli ammettono descrizioni a tempo continuo, p.es. [11]). L'infezione di un individuo può avvenire a causa di trasmissione diretta da parte di individui infetti o di contaminazioni ambientali, sempre però all'interno di reti di contatti realistiche, che riproducono correttamente le proprietà statistiche delle interazioni sociali, dall'ambito famigliare a quello lavorativo. Una caratteristica fondamentale di tutti i modelli realistici di diffusione del SARS-CoV-2 è il lungo periodo di incubazione e la presenza di un compartimento in cui l'agente è infettivo senza sintomi apparenti. L'analisi di dati reali ha altresì mostrato come l'infettività non sia costante nel tempo (come assunto negli usuali modelli Markoviani) ma sia una funzione del tempo con un picco a 5 giorni [9]. Ci focalizzeremo qui sul modello OpenABM, recentemente sviluppato da Ferretti et al [9], in cui 1 milione di individui interagiscono in una rete urbana a

tre differenti livelli (contatti domestici, lavorativi e casuali), la cui struttura è costruita a partire da dati demografici del Regno Unito.

Il problema del contenimento epidemico

Lo strumento fondamentale per il contenimento epidemico è quello del isolamento parziale o totale (ad es. quarantene coatte o restrizioni alla mobilità di individui o unità domestiche). Limitare il numero di persone sottoposte a quarantena è fondamentale ed in generale è auspicabile che questa sia preceduta da un risultato positivo ad un test dell'individuo interessato o di un coinquilino o familiare. Nel nostro contesto, il numero di test N_{test} che possono essere effettuati/analizzati ogni giorno è limitato (dovuto essenzialmente alla capacità di analizzarli). Quando un individuo risulta positivo viene immediatamente messo in quarantena (assumiamo per semplicità che il risultato del test sia disponibile lo stesso giorno), assieme ai componenti della sua unità domestica.

Il problema di decidere giornalmente quali individui sottoporre a test per minimizzare il numero totale (atteso) di individui infetti è un problema di *controllo stocastico*. La soluzione ottimale di problemi di controllo stocastico come questo è tipicamente estremamente complessa (ancora più difficile della valutazione del rischio, che in sostanza corrisponde al solo calcolo della funzione obiettivo!). In questo lavoro ci limiteremo dunque a proporre strategie euristiche.

Come esempi di strategie *non* probabilistiche, consideriamo due semplici metodi per la scelta degli N_{test} individui da sottoporre a test giornalmente:

- *random*: N_{test} individui scelti a caso tra quelli che non sono mai risultati positivi;
- *tracing*: gli N_{test} individui con il più alto numero di precedenti contatti con altri individui testati positivi.

La strategia `probabilistica` più diretta è quella di scegliere ogni giorno gli N_{test} individui con probabilità di infezione maggiore stimata dall'algoritmo di inferenza. Una piccola modifica, che permette di migliorarne l'efficacia, consiste nel concentrarsi sugli individui recentemente infettati (per esempio utilizzando la stima della probabilità di essere stato contagiato negli ultimi 10 giorni). Questi hanno infatti una maggiore potenzialità infettiva e si trovano vicino al «bordo» della propagazione epidemica, pertanto il loro isolamento consente di contenerne più facilmente l'avanzata. Per fortuna, la strategia probabilistica basata su Belief Propagation [8] permette anche di valutare questi aspetti più specifici del rischio infettivo.

Un confronto tra diverse strategie d'intervento in un esempio di propagazione epidemica nel modello OpenABM è riportato in Figura 4. I risultati mostrano che la stima probabilistica del rischio tramite BP (assumendo per l'inferenza un modello epidemico SIR, che è estremamente più semplice di OpenABM) consente una più accurata identificazione di casi non sintomatici (valutata dall'area sotto la ROC), risultando in un contenimento estremamente più efficace (si veda [8] per statistica e più dettagli).

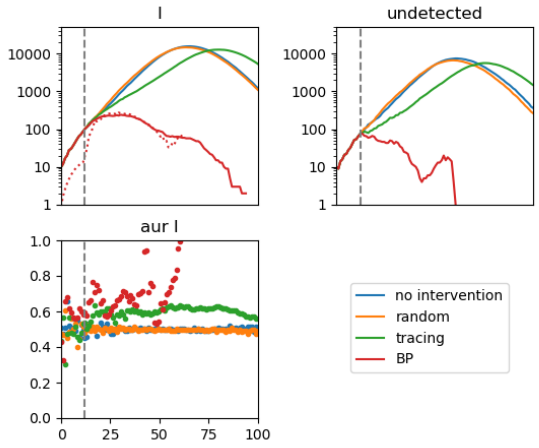

Figura 4: *Confronto tra strategie di contenimento in un esempio di una popolazione con 50K individui in cui la diffusione segue il modello OpenABM. Assumiamo che gli individui con sintomi gravi e il 50% di quelli con sintomi di media gravità vengano immediatamente isolati (o ricoverati) alla comparsa dei sintomi. Inoltre, i $N_{test} = 200$ individui selezionati dalle strategie random, tracing e BP sono sottoposti a test ogni giorno, e i casi positivi vengono isolati. L'asse orizzontale rappresenta il tempo (in giorni). Il sistema parte a tempo 0 con 10 infetti (P0), gli interventi iniziano al decimo giorno (linea tratteggiata verticale). (I): numero di individui infetti (per BP, anche la sua stima, in linea tratteggiata), (undetected): numero di infetti non testati, (aur I): area sotto la curva ROC.*

L'apprendimento automatico del modello epidemico

Nella discussione svolta sinora abbiamo trascurato un aspetto fondamentale: per essere in grado di ottenere buone predizioni, il modello utilizzato per l'inferenza deve rispecchiare il più possibile la reale dinamica di diffusione epidemica, rendendo indispensabile un processo di calibrazione dei parametri. Tale calibrazione richiede la scelta dei valori di un numero (potenzialmente molto grande) di parametri, che indicheremo congiuntamente con il vettore $\boldsymbol{\theta}$. Nel caso del modello SIR standard, ad esempio, i parametri sono solo due $\boldsymbol{\theta} = (\lambda, \mu)$: la probabilità λ che un contatto porti ad un contagio e la probabilità di guarigione istantanea μ di un individuo infetto.

Data un'osservazione parziale $\mathcal{O}$ del sistema, si vorrebbero trovare i parametri di massimo potere predittivo. Ma come misurare il potere predittivo? Definito un indice di rischio individuale – ad esempio la probabilità a posteriori di essere infetto $p(x_i = I|\mathcal{O})$ – si vorrebbe in generale valutare positivamente un metodo che assegna una stima di rischio relativamente più elevata agli individui infetti (rispetto a quelli sani). In tal senso, il potere predittivo del metodo di inferenza può essere valutato, per esempio, mediante il calcolo dell'area cumulata al di sotto della curva Receiver Operating Characteristic (ROC). Essa corrisponde infatti alla frazione di tutte le coppie di individui sano-infetto in cui il metodo di inferenza assegna un indice di rischio più alto a quello infetto. La curva ROC non può essere utilizzata direttamente per la scelta dei parametri perché utilizza dati inaccessibili (il vero stato di infezione degli individui non osservati), ma può essere utilizzata per valutare l'efficacia di altri metodi di inferenza.

Un approccio frequentemente utilizzato per l'apprendimento automatico dei parametri è quello di estendere ad essi il contesto probabilistico, assumendo che tutte le distribuzioni previamente definite siano condizionate al valore del vettore $\boldsymbol{\theta}$. Questo ci permette di definire i parametri più verosimili come quelli che massimizzano la loro probabilità a posteriori date le osservazioni $p(\boldsymbol{\theta}|\mathcal{O}) \propto p\left(\mathcal{O}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta})$, oppure, in assenza di informazioni a priori per $\boldsymbol{\theta}$, semplicemente la log-likelihood

$$\mathcal{L} = \log p\left(\mathcal{O}|\boldsymbol{\theta}\right) = \log \sum_{\boldsymbol{x}} p\left(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta}\right). \tag{12}$$

Nel contesto meccanico-statistico $-\mathcal{L}$ corrisponde alla cosiddetta energia libera del modello $-\log Z$ in (9). Come esempio dell'efficacia della massimizzazione di $\mathcal{L}$, mostriamo in figura 5 la sua approssimazione di Bethe come funzione di due parametri di un modello SIR non Markoviano, in cui il tempo di guarigione di un individuo varia con un profilo temporale caratteristico non monotono dal momento del contagio. L'inferenza è stata effettuata a partire da un insieme di osservazioni $\mathcal{O}$ su una cascata epidemica generata dal modello OpenABM (definito in [9]). Anche se l'esatto punto di massimo delle due misure non coincide, la log-likelihood (che, ricordiamo, utilizza solo i dati accessibili dell'osservazione $\mathcal{O}$) è quasi ottima in corrispondenza di un sottoinsieme di parametri che hanno massimo potere predittivo nel modello OpenABM. Il risultato è notevole, soprattutto se si tiene in considerazione che il modello probabilistico di inferenza assume una dinamica di diffusione SIR a tre stati rispetto agli 11 del modello che genera i dati [9].

La figura 5 è stata realizzata tramite il calcolo *esaustivo* di un gran numero di punti fissi di BP corrispondenti ai diversi valori combinati dei parametri. Per ognuno di essi è stata calcolata la corrispondente approssimazione di Bethe dell'energia libera. All'aumentare del numero di parametri – in un'ottica di calcolo distribuito – questa soluzione risulta ovviamente troppo lenta per essere messa in pratica. Diventa allora essenziale trovare un metodo *non esaustivo* per la ricerca dei parametri ottimali, massimizzando direttamente $\mathcal{L}$. Purtroppo, il problema della massimizzazione di $\mathcal{L}$ è per sua natura particolarmente ostico, per via della sommatoria su un numero esponenziale di stati epidemici in (12). Un approccio spesso utilizzato in questa situazione è quello di *Expectation Maximization* (EM). In EM, si sfrutta l'espressione variazionale dell'energia libera del sistema canonico

$$-\log Z = \langle H \rangle_p - S(p) \tag{13}$$

$$= \min_q \langle H \rangle_q - S(q) \tag{14}$$

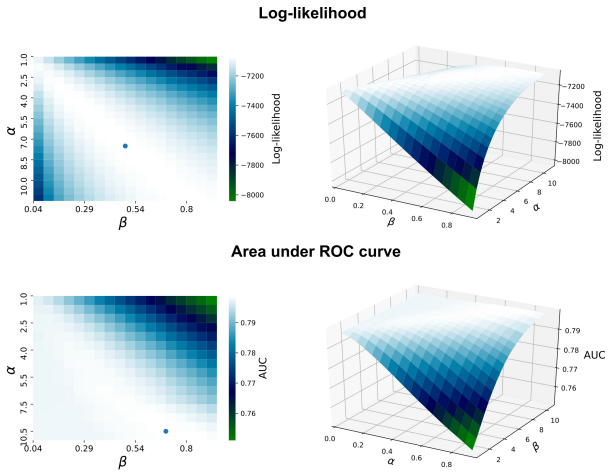

Figura 5: *Log-likelihood e area cumulata dalla curva ROC in funzione di due parametri (α, β) caratteristici della funzione temporale di guarigione (distributione gamma con forma α e scala inversa β) su un'epidemia di 10 000 individui generata da OpenABM. Si veda [14] per ulteriori dettagli.*

dove $p = p(\mathbf{x}|\mathcal{O}, \boldsymbol{\theta})$, $q(\mathbf{x})$ è una distribuzione arbitraria, $\langle\cdot\rangle_q$ denota una media rispetto alla probabilità q, ed S è l'entropia di Shannon $S(q) = -\sum_{\mathbf{x}} q(\mathbf{x}) \log q(\mathbf{x})$. Nel nostro caso specifico, massimizzando su $\boldsymbol{\theta}$ otteniamo

$$\max_{\boldsymbol{\theta}} \mathcal{L} = \max_{\boldsymbol{\theta}, q} \langle \log p(\mathbf{x}, \mathcal{O}|\boldsymbol{\theta}) \rangle_q + S(q) \qquad (15)$$

Partendo da una stima iniziale (anche rozza o banale) del vettore dei parametri $\boldsymbol{\theta}_0$, l'idea di EM è quella di risolvere la doppia ottimizzazione con una massimizzazione alternata e reiterata su q (a $\boldsymbol{\theta}$ fissato) e su $\boldsymbol{\theta}$ (a q fissato). Notando che nella k-esima iterazione l'ottimizzazione su q a $\boldsymbol{\theta}_k$ fissato comporta che q debba banalmente essere uguale a $p_k(\mathbf{x}) = p(\mathbf{x}|\mathcal{O}, \boldsymbol{\theta}_k)$ grazie a (13)-(14) e sfruttando il fatto che nell'espressione a q fissato S non dipende da $\boldsymbol{\theta}$, l'iterazione si semplifica in

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} \langle \log p(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta}) \rangle_{p_k} . \qquad (16)$$

In un punto fisso di (16), la log-likelihood è stazionaria rispetto al vettore di parametri $\boldsymbol{\theta}$. Il grosso vantaggio di EM è l'apparentemente miracoloso scambio effettivo tra logaritmo e somma (diventata media) da (12) a (16). Essendo nel nostro caso $\log p(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta})$ una somma di termini locali (si veda Eq. (10)), la media in (16) risulterà facile da stimare con BP. Purtroppo la dipendenza dai parametri $\boldsymbol{\theta}$ è nel nostro caso ancora troppo complicata per poterla ottimizzare in modo esplicito. Un modo efficace per implementare una strategia di ottimizzazione è quello di percorrere successivamente piccoli passi nella direzione di massima crescita, usando l'informazione direzionale del gradiente $\nabla_{\boldsymbol{\theta}} \langle \log p(\boldsymbol{x}, \mathcal{O}|\boldsymbol{\theta}) \rangle_{p_k}$ – questo non è altro che il metodo della Discesa del Gradiente (o più propriamente nel nostro caso, ascesa), classicamente utilizzato in ambito di ottimizzazione convessa, ora divenuto la tecnica algoritmica standard nell'Apprendimento Automatico dei *Deep Networks*. Previo scambio di ∇ e $\langle\cdot\rangle$, questo gradiente diventa una somma di termini locali, che può essere calcolata facilmente tramite scambio di messaggi tra nodi vicini. Nella pratica, si è visto che è sufficiente alternare un singolo step nelle equazioni di BP ad un piccolo aggiornamento dei parametri, utilizzando la stima corrente del gradiente [15]. Questa procedura viene ripetuta in maniera iterativa sino al raggiungimento di un punto fisso, in corrispondenza del quale l'approssimazione di Bethe della log-likelihood è stazionaria rispetto ai parametri. Come quelli precedenti, anche questo calcolo può essere effettuato in modo distribuito tramite scambi locali tra le app di individui che sono stati in contatto.

Conclusioni

Le tecniche di inferenza probabilistica e di Machine Learning possono avere un ruolo primario nel contrasto alla diffusione dei patogeni. Il caso del virus SARS-Cov-2 è emblematico: buona parte della dinamica diffusiva è nascosta all'osservatore a causa dell'alta percentuale di individui asintomatici infettivi. Il ricorso al paradigma probabilistico pone, come si è visto, formidabili problemi computazionali, che possono essere affrontati ricorrendo a metodi di approssimazione sviluppati nel contesto della Meccanica Statistica. L'inferenza su modelli stocastici a compartimenti consente di individuare soggetti ad alto rischio di infezione e di guidare procedure di test e quarantene. Inoltre, l'evidenza dei test, per sua stessa natura parziale e rumorosa, permette di acquisire importanti informazioni sulla meccanica stessa della diffusione del virus in popolazione, attraverso una costante ri-calibrazione automatica del modello inferenziale. Per concludere, fac-

ciamo notare che la procedura algoritmica qui descritta – essendo basata su scambio di messaggi tra individui e non prevedendo il ricorso ad un'unità centralizzata di elaborazione ed immagazzinamento dati – risulta un ottimo candidato per lo sviluppo di un sistema scalabile ed attento alla privacy degli individui. Avvertiamo tuttavia il lettore che gli argomenti discussi riguardano solo aspetti epidemiologici e che uno sviluppo concreto di questi sistemi presenta sicuramente altre sfide di carattere tecnologico.

Ringraziamenti

Ringraziamo vivamente i nostri amici e colleghi del Politecnico di Torino I. Biazzo, G. Catania, F. Mazza e A.P. Muntoni che hanno svolto con noi la ricerca sugli argomenti qui presentati (in particolare A.P.M. per la Figura 5). Inoltre, A.B. ed L.D. ringraziano F. Ricci, E. Marinari e L. Ferretti per interessanti discussioni.

[1] M. Day: *Covid-19: identifying and isolating asymptomatic people helped eliminate virus in italian village.*, BMJ, 368 (2020) m1165.

[2] Centers for disease control and prevention. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html. Aggiornato a Settembre 2020.

[3] Immuni contact tracing app. https://www.immuni.italia.it/, 2020.

[4] I. Bestvina. Viratrace. https://github.com/ViraTrace/InfectionModel, 4 2020.

[5] L. Simmonds. Could croatian startup rescue tourism from coronavirus consequences? https://www.total-croatia-news.com/made-in-croatia/43538-coronavirus, (2020)

[6] M. R. Garey and D. S. Johnson.: *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)* W. H. Freeman, New York (1979).

[7] R. Herbrich, R. Rastogi, R. Vollgraf: *Crisp: A probabilistic model for individual-level Covid-19 infection risk estimation based on contact data*, arXiv:2006.04942 (2020).

[8] A. Baker at al. : *Epidemic mitigation by statistical inference from contact tracing data, 2020*, arXiv:2009.09422.

[9] L. Ferretti et al.: *Quantifying SARS-Cov-2 transmission suggests epidemic control with digital contact tracing*, Science, 368 (2020) eabb6936.

[10] M. Chinazzi et al.: *The effect of travel restrictions on the spread of the 2019 novel coronavirus (Covid-19) outbreak*, Science, 368 (2020) 395.

[11] L. Lorch at al. : *A spatiotemporal epidemic model to quantify the effects of contact tracing, testing, and containment*, arXiv:2004.07641 (2020).

[12] C. C. Kerr at al. : *Covasim: an agent-based model of Covid-19 dynamics and interventions.* medRxiv (2020). https://doi.org/10.1101/2020.05.10.20097469

[13] J. A. Moreno López et al. : *Anatomy of digital contact tracing: role of age, transmission setting, adoption and case detection*, medRxiv (2020) https://doi.org/10.1101/2020.07.22.20158352

[14] Sibyl website. https://github.com/sibyl-team (2020).

[15] A. Braunstein, A. Ingrosso: *Inference of causality in epidemics on temporal contact networks*, Sci. Rep, 6 (2016) 27538.

Alfredo Braunstein: è Professore Associato presso il Politecnico di Torino, e si occupa di applicazioni della Meccanica Statistica in particolare a problemi di inferenza ed ottimizzazione.

Luca Dall'Asta: è Professore Associato presso il Politecnico di Torino. Si occupa di Fisica Statistica e dei Sistemi Complessi e delle sue applicazioni interdisciplinari.

Alessandro Ingrosso: è *Postdoctoral Fellow* presso l'*Abdus Salam International Centre for Theoretical Physics*. Si occupa di Neuroscienze Computazionali, Machine Learning ed applicazioni interdisciplinari della Meccanica Statistica.