Spigolature astronomiche*

A cura di Annibale D'Ercole

INAF · Osservatorio di astrofisica e scienza dello spazio di Bologna (OAS)

Teoria delle code

Annibale D'Ercole

Il tempo che la popolazione trascorre in coda è veramente notevole: uno studio effettuato nel 2018 dall'ISTAT ha evidenziato come ogni anno gli italiani passino in media circa 400 ore (16 giorni) in fila davanti a qualche sportello (di banca, di posta, di una ASL) o in qualche negozio! Questo, oltre a portare stress, frustrazione e nervosismo, si traduce in una perdita economica di oltre 40 miliardi di euro dovuta alla mancanza di operatività delle persone in attesa. Al di là delle singole esperienze quotidiane, le code sono presenti in vari tipi di industria.[1] Non stupisce, quindi, che si sia presentata ad un certo punto la necessità di studiare i meccanismi di sistemi di coda – anche molto complessi – in modo da poterli ottimizzare, ossia ridurre i tempi di attesa e incrementare il rendimento economico (come vedremo).

La teoria delle code è più complicata di quanto si possa pensare di primo acchito, sia concettualmente che matematicamente. La sua origine risale all'inizio del secolo scorso quando Agner Krarup Erlang (1878-1929), ingegnere danese della Compagnia dei telefoni di Copenaghen, pubblicò un primo lavoro su tale materia nel 1909. Egli era interessato a determinare il numero di circuiti e centralinisti necessario per ottenere un accettabile attesa da parte degli utenti. La sua analisi, culminata in un articolo del 1920 intitolato *Tempi di attesa telefonici*, descrive i primi modelli di coda ed è alla base della teoria delle code. L'unità internazionale di traffico telefonico è stata battezzata *erlang* in suo onore.[2]

La teoria delle code è uno studio matematico della formazione, gestione e congestione delle code di attesa. Nonostante la complessità dell'aspetto matematico, il "nocciolo" della teoria è alquanto semplice ed è composto da due parti:

1. qualcuno o qualcosa che richiede un servizio detto in genere *utente* o *richiedente*;
2. qualcuno o qualcosa che rilascia il servizio e che indicheremo nel seguito con il generico termine di *servente*.

A scopo illustrativo diamo due esempi. Nel primo consideriamo la fila che si forma allo sportello di una banca: gli utenti sono le persone che intendono versare o ritirare denaro mentre il cassiere è il servente (un esempio del tutto analogo è dato dalla coda davanti ad un ATM).[3] Il secondo esempio è dato da una coda costituita da documenti elettronici indirizzati a una stampante: gli utenti sono coloro che li hanno inviati mentre la stampante è il servente. La FIG. 1 rappresenta un possibile tipo di coda con più serventi (la rincontreremo più tardi).

La teoria delle code analizza l'intero sistema di attesa che comprende la *frequenza media di arrivo* λ dei clienti (il numero di clienti che giungono nell'unità di tempo, p.e. numero di clienti/ora che entrano in un negozio), la *frequenza media di effettuazione del servizio* μ (il numero di clienti serviti nell'unità di tempo, p.e. numero di clienti/ora), la *capacità media L* del sistema che indica il numero totale di utenti (considerando sia gli utenti in coda sia quelli che stanno beneficiando del servizio), il numero dei serventi *m*, il *tempo medio W* di attesa totale del singolo utente all'interno del sistema, la *disciplina della coda* (quest'ultima si riferisce alle regole della coda; ve ne sono diverse, noi ne riportiamo tre nella TABELLA 1).

* Questa rubrica – iniziata nel 1999 e che ha raggiunto quasi i 100 numeri – si propone di presentare in modo sintetico e, per quanto possibile, autoconsistente argomenti che stanno alla base della conoscenza astronomica, spesso trascurati nella letteratura divulgativa, in quanto ritenuti di conoscenza generale oppure troppo difficili o troppo noiosi da presentare a un pubblico non specialistico. Questi "fondamenti di astronomia", volutamente trattati in uno spazio limitato, possono essere letti a due livelli; eventuali approfondimenti per i lettori che desiderino ampliare la conoscenza dell'argomento vengono esposti in carattere corsivo e incorniciati. Si suggerisce questa rubrica, quindi, a studenti dei vari tipi e livelli di scuole. Le *Spigolature astronomiche* si possono trovare anche in rete, nel sito Web del «Giornale di Astronomia», http://giornaleastronomia.difa.unibo.it/giornale.html.

[1] Tra questi ricordiamo il commercio, le Telecomunicazioni, i trasporti, le banche e la Finanza, il calcolo informatico, la logistica.

[2] L'erlang (E) è adimensionale e rappresenta l'intensità di occupazione nell'unità di tempo. Ad esempio: se un utente parla al telefono per 50 minuti in un'ora esso avrà sviluppato 50/60 = 0,833 erlang in quell'ora.

[3] ATM è l'acronimo di *automated teller machine*, ossia lo "sportello automatico" che permette di effettuare operazioni bancarie.

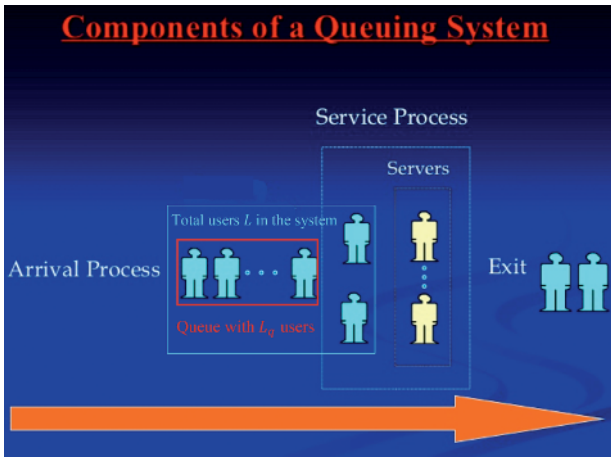

Fig. 1. Rappresentazione schematica di un sistema di coda. Il riquadro rosso evidenzia la coda composta da L_q utenti (omini azzurri) mentre il riquadro azzurro comprende tutti gli L utenti, quelli in coda e quelli alle prese con i serventi (omini gialli). Il sistema nel suo complesso comprende sia gli utenti che i serventi.

È necessario sottolineare che per poter trattare una coda è necessario che quest'ultima sia *stabile*, ossia che venga verificata la condizione $\lambda < \mu$; in altre parole, il ritmo con cui si aggiungono clienti alla fila alla cassa di un negozio deve essere inferiore al ritmo con cui il cassiere "smaltisce" i clienti.[4] In verità l'intuizione suggerirebbe che non ci dovrebbe essere alcuna fila visto che il cliente alla cassa viene servito prima che ne sopraggiunga un altro. Il paradosso deriva dal fatto che utilizziamo valori medi per λ e μ; in realtà una fila può ugualmente formarsi dal momento che il tempo di interarrivo (ossia il tempo che intercorre tra l'arrivo di un cliente e il successivo) è casuale e può accadere che occasionalmente il tasso di arrivo superi quello di servizio. Lo stesso discorso vale per il servizio al cliente (alcuni pagano rapidamente con carta di credito, altri più lentamente in contanti e in attesa di un eventuale resto). Riprenderemo questo punto nel livello avanzato.

Se per studiare sistemi molto complessi è necessario ricorrere all'utilizzo del calcolatore, esiste un'equazione tanto semplice quanto potente in grado di analizzare un sistema di code e di dare una prima rapida risposta alla richiesta di ottimizzazione del sistema, ossia di minimizzazione dei tempi di attesa. Si tratta della *legge di Little*[5] che collega tra loro la capacità *media* di un sistema (L), il tempo *medio* speso nel sistema (W) e la frequenza *media* degli arrivi (λ), senza la necessità di conoscere il numero di serventi, la disciplina della coda, o qualunque altra caratteristica. Come vedremo nel livello avanzato, questa legge vale non solo per un intero sistema ma

[4] In caso contrario la coda è instabile e tende a crescere all'infinito creando un ingorgo nel negozio.

[5] La legge è stata formulata dal fisico statunitense John Little (1929-2024) mentre era impegnato presso la General Electric per la soluzione di problemi di controllo del traffico dei segnali nell'ambito dei *problemi decisionali* tesi ad ottimizzare determinate tecniche in campo industriale per massimizzare profitti ed efficienza e minimizzare costi, rischi e perdite.

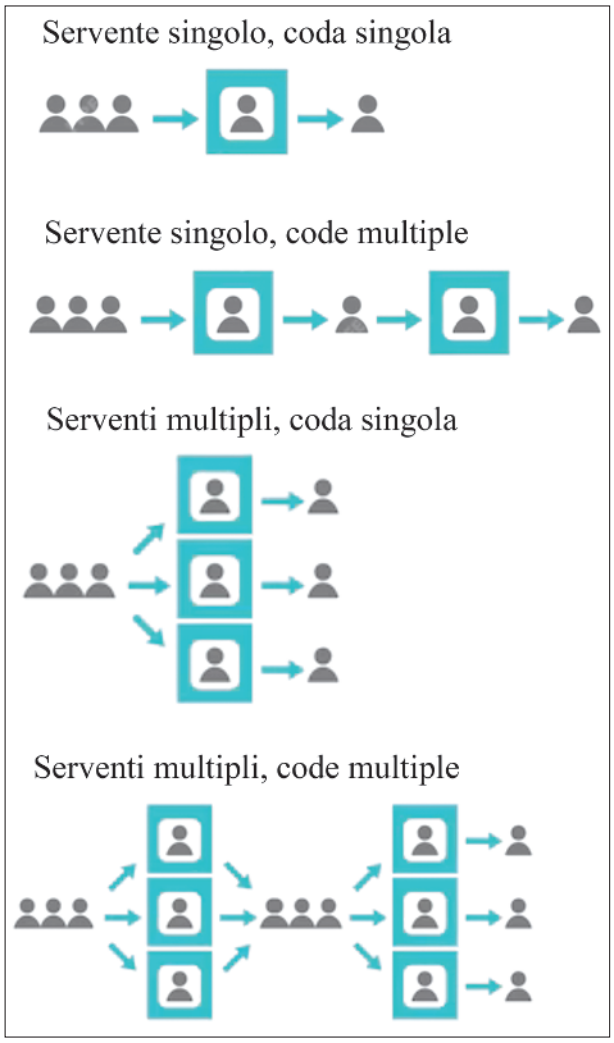

Fig. 2. La figura illustra schematicamente i diversi sistemi di coda elencati in Tabella 2.

anche per suoi sottosistemi. Inoltre, essa è valida indipendentemente dal fatto che gli intervalli temporali tra un arrivo e il successivo o tra un servizio e l'altro siano distribuiti in maniera erratica oppure uniforme o seguano una qualche particolare distribuzione statistica (si veda il livello avanzato).

La formula di Little è alquanto semplice:

$$L = \lambda\, W. \qquad (1)$$

"Tradotto" in parole questo significa che il numero medio di utenti presenti nel sistema è dato dal numero medio di utenti che arrivano nell'unità di tempo moltiplicato per il tempo medio di attesa di un singolo utente.

Come abbiamo detto, la formula vale anche per sottosistemi. Consideriamo una fila in banca servita da uno o più sportelli. Escludendo dal numero totale L di utenti quelli che vengono serviti (cioè quelli alle prese con i cassieri), ne rimangono L_q che compongono la fila (Fig. 1); anche in questo caso vale la formula

$$L_q = \lambda\, W_q, \qquad (2)$$

dove W_q è il tempo di attesa dovuto unicamente alla coda, escludendo il tempo speso mentre si è serviti

Tabella 1	
DISCIPLINA DELLA CODA	ESEMPI
FIFO / FCFS *"first-in, first-out"* / *"first-come, first-served"*	È la disciplina più diffusa. I clienti vengono serviti in ordine di arrivo, come capita in un bar, o per l'utilizzo di un ATM, ecc.
LIFO / LCFS *"last-in, first-out"* / *"last-come, first-served"*	Data una pila di piatti appena lavati, l'ultimo aggiunto viene prelevato per primo. Nel riempire un furgone di un trasportatore, generalmente l'ultimo oggetto caricato è il primo ad essere scaricato in maniera da ottimizzare il percorso.
SIRO *"service in random order"*	Questa disciplina è basata sull'ordine delle priorità, tipica dei pronto soccorso.

Tabella 2 (vedi Fig. 2)	
SISTEMA DI CODA	ESEMPI
Servente singolo, coda singola	Gli utenti sono disposti in un'unica fila con un unico servente. È quello che accade per gli utenti in coda davanti ad un ATM.
Servente singolo, code multiple	Gli utenti sono disposti in più file, ciascuna servita da un unico servente: gli utenti attendono ogni volta per ciascuna fila. In un ambulatorio l'utente attende in una prima fase alla reception e in una seconda fase per la visita.
Serventi multipli, coda singola	Gli utenti attendono in un'unica fila con più serventi a disposizione; ogni utente aspetta solo una volta e si rivolge al primo servente che si rende disponibile. È il caso della fila in banca dove sono presenti più sportelli. È anche il caso in Fig. 1.
Serventi multipli, code multiple	Gli utenti attendono in più file e sono serviti dal primo servente disponibile di ciascuna fila. In un fast-food con più file i clienti in una prima fase attendono di ordinare a una delle casse, c'è poi una seconda fase in cui i clienti si mettono in fila aspettando che alcuni addetti preparino quanto ordinato, e infine c'è una terza fase in cui diversi inservienti impacchettano e consegnano il pasto ai clienti in attesa in quest'ultima fila.

(questa distinzione ha la sua importanza, come vedremo nel livello avanzato).

Prima di proseguire è necessario sottolineare che la legge di Little è valida in caso *stazionario*, cioè quando il tasso in ingresso nella coda è pari al tasso in uscita:

$$\lambda = \lambda_{entrata} = \lambda_{uscita}.$$

In altri termini, la formula di Little esprime una legge di conservazione in quanto gli elementi che entrano nella coda devono essere pari a quelli che ne escono.

Per quanto l'eq. (1) sembri talmente semplice da apparire quasi banale, essa permette di gestire situazioni diverse a seconda dell'informazione che vogliamo ottenere. Diamo qui di seguito tre semplici esempi considerando un sistema composto da una singola coda FIFO (*first-in, first-out*).

Nel primo esempio, immaginiamo un seggio elettorale in cui vengono "smaltiti" in media 50 elettori ogni ora e supponiamo che ogni cittadino attenda 10 minuti dal momento in cui arriva a quello in cui esce dal seggio. Ci chiediamo quanto valga L, ossia quante persone siano presenti nel seggio (scrutatori esclusi). Abbiamo $\lambda = 50$ (votanti ogni ora) e $W = 1/6$ il tempo di attesa (in ore): dall'eq. (1) ci aspettiamo che la coda sia composta mediamente da $L = \lambda\, W \approx$ 8-9 persone.

Supponiamo ora di essere in coda in un caffè aspettando di essere serviti e vogliamo stimare il tempo di attesa. Assumiamo che ci siano $L = 15$ avventori presenti, e che entrino (ed escano, per quanto detto più sopra sulla stazionarietà) $\lambda = 2$ persone ogni minuto. Dalla eq. (1) abbiamo $W = L/\lambda = 7{,}5$ minuti.

Ipotizziamo adesso che l'ufficio prestiti di una banca impieghi mediamente 6 giorni per espletare una pratica e che 100 pratiche in diversi stadi di elaborazione siano in attesa di essere completate. Ci chiediamo quante pratiche vengono perfezionate

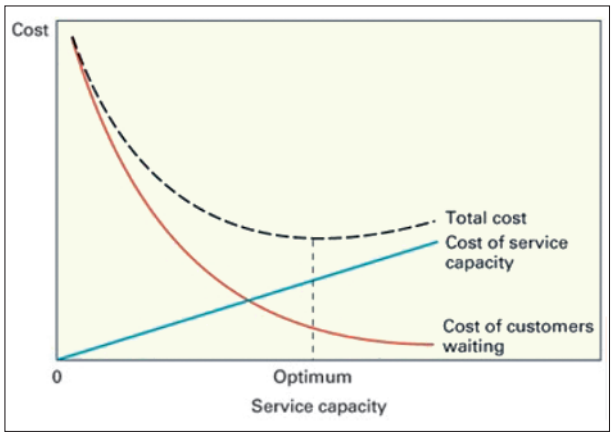

FIG. 3. Minimizzazione del costo di un sistema di code. Con riferimento all'esempio del negozio dato nel testo, in ascissa abbiamo il numero di cassieri e in ordinata i costi. La linea rossa rappresenta il costo di attesa e quella blu il costo del servizio: quest'ultimo aumenta al crescere dei cassieri, mentre il primo diminuisce per il diminuire del tempo di attesa (si veda il testo). La curva tratteggiata rappresenta il costo totale dovuto alla somma dei due precedenti: il suo punto più basso individua (segmento tratteggiato) il numero ottimale di cassieri per ridurre i costi.

ogni mese. In questo caso abbiamo $L = 100$ e $W = 6/30 = 0{,}2$ mesi; di conseguenza vengono completate $\lambda = L/W = 500$ pratiche al mese.

Ma la legge di Little, nella sua semplicità, è estremamente versatile e si rivela utile nei processi di ottimizzazione come quello, preso come esempio dall'ambito militare, che ci accingiamo a descrivere. Consideriamo i costosissimi 19 bombardieri statunitensi B-2 in grado di risultare (quasi) "invisibili" ai radar. La loro complessità richiede una manutenzione frequente che può durare un lasso di tempo W variabile dai 18 ai 45 giorni. La legge di Little ci aiuta a bilanciare al meglio il numero degli aerei operativi e di quelli in manutenzione. In base all'analisi dei piani di volo, risulta che ad ogni istante sono mediamente in manutenzione $L = 3$ bombardieri e che ogni B-2 entra in manutenzione circa ogni 7 giorni, ossia $\lambda = 1/7$. Dall'eq. (1) otteniamo allora $W = L/\lambda = 21$ giorni. Questo è dunque il numero ottimale di giorni di manutenzione per conciliare il numero di aerei operativi con i piani di volo regolari.

Quello appena descritto è un esempio alquanto semplice di ottimizzazione. In generale, scopo di una buona progettazione di un sistema di code è quello di minimizzare i costi. Questi sono di due tipi: il *costo di attesa* e il *costo di servizio*. Nell'esempio già visto in precedenza della fila alla cassa di un negozio il costo di attesa è dovuto al numero di clienti che, scoraggiati dalla lunghezza della fila, rinunciano all'acquisto e vanno via; il costo di servizio è dato (per semplificare) dal costo del cassiere. Naturalmente, è possibile aumentare il numero di cassieri riducendo i tempi di attesa e la "fuga" dei clienti impazienti, ma il costo del servizio aumenta per il maggior numero di stipendi per i cassieri. Scopo dell'ingegnere della coda (il matematico che si occupa della teoria delle code) è quello di minimizzare i costi (senza ridurre i guadagni), come illustrato in FIG. 3. Ne riparleremo nel livello avanzato.

La teoria delle code si occupa anche (e soprattutto) di casi ben più complessi di quelli qui considerati.[6] Noi abbiamo sorvolato a volo d'uccello la superficie della teoria; nel livello avanzato daremo una grattatina a tale superficie per intravvedere cosa c'è sotto, ma senza alcuna velleità di "scavare".

In questo livello cercheremo di capire meglio il funzionamento di una coda in base ad alcune sue grandezze dette indici di prestazione *(p.e. L, L_q, W, W_q). Nel livello base abbiamo accennato al fatto che i tassi di arrivo e di servizio in una coda variano in maniera casuale in base ad una qualche distribuzione di probabilità. La relazione tra le statistiche degli arrivi e dei servizi e gli indici di prestazione si basa su modelli matematici alquanto complessi che non è possibile riportare in questa sede. Riteniamo tuttavia istruttivo illustrare ugualmente i rudimenti statistici alla base della teoria delle code. Il lettore non interessato può saltare gli argomenti esposti nei prossimi quattro paragrafi e passare direttamente agli indici di prestazione.*

La distribuzione di probabilità che in genere si verifica per gli arrivi è quella di Poisson[7] (detta poissoniana) che calcola la probabilità $P(n)$ di realizzazione di n eventi in un intervallo temporale di prefissata durata T:[8]

$$P(n) = \frac{(\lambda T)^n}{n!} e^{-\lambda T}, \qquad (3)$$

dove λ è il tasso medio di arrivi e il punto esclamativo indica il numero fattoriale.[9]

A titolo di esempio riportiamo uno studio della NASA secondo cui asteroidi più grandi di 1 km colpiscono la Terra mediamente una volta ogni

[6] In una fabbrica di automobili vi sono diverse code. Ad esempio, c'è ne una che indirizza le scocche al reparto verniciatura, mentre in un'altra (una catena di montaggio) le scocche vengono "arredate" man mano che avanzano. C'è poi un'altra catena di montaggio (tra le tante che qui trascuriamo) dove vengono assemblati i motori. È necessario che tutte queste code siano coordinate al meglio evitando la presenza di "colli di bottiglia", ossia la presenza anche di una sola coda progettata male che rallenti la produzione di tutta la fabbrica. Un altro esempio di complessità è dato da un incrocio particolarmente complicato verso cui convergono numerose strade, ognuna con un flusso automobilistico diverso. In questo caso è necessario calcolare il tempo di permanenza del verde per ogni semaforo in modo da ridurre il tempo di attesa generale evitando la formazione di ingorghi.

[7] Senza soffermarci sul perché, diciamo qui che la distribuzione di Poisson si riscontra spesso in natura e descrive bene la statistica di eventi discreti tra i più disparati, dalla probabilità che in un prefissato intervallo di tempo un atomo radioattivo decada alla probabilità che una lampadina si fulmini.

[8] Questa formula venne proposta dal matematico e fisico francese Siméon-Denise Poisson (1781-1840) e verificata dall'economista e statistico russo Ladislaus Bortkiewicz (1868-1931) analizzando i dati sugli incidenti mortali da cavallo presso i 14 reparti della cavalleria prussiana nell'arco di 20 anni.

[9] $n! = 1 \times 2 \times 3 \times \ldots \times (n-1) \times n$. Ad esempio, $5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$. Per definizione, $0! = 1$.

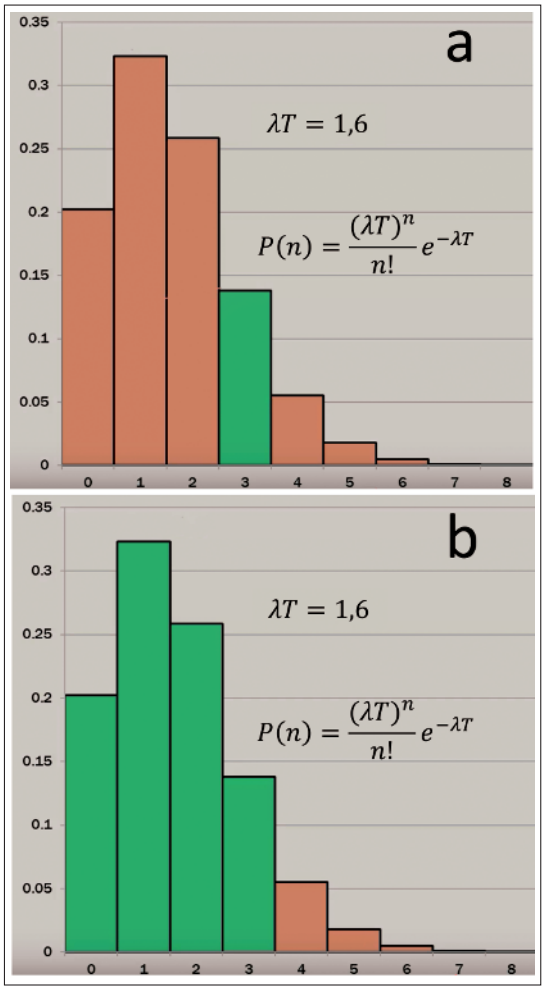

FIG. 4. **a**) Distribuzione di Poisson (la cui formula è riportata in figura) illustrata tramite un istogramma. In ascissa abbiamo il numero di eventi che possono realizzarsi (in un intervallo di tempo prefissato T). In questo particolare esempio abbiamo $\lambda T = 1{,}6$, con riferimento al caso discusso nel testo in cui l'evento è rappresentato dalla caduta di un meteorite con $\lambda = 1$. Ogni barra è centrata in ascissa ad un numero n di eventi, mentre la sua altezza è pari alla probabilità [calcolata mediante l'eq. (3)] che questo numero si verifichi. Dal momento che la larghezza di una barra è pari a 1, l'area della barra indica proprio la probabilità di n. La barra colorata in verde che indica la probabilità che si realizzino 3 eventi evidenzia questo punto; **b**) Dal momento che un possibile valore di n (compreso $n = 0$, ossia nessun evento) deve necessariamente verificarsi, la somma di tutti i $P(n)$, ossia l'area totale dell'istogramma, deve essere pari a 1. Pertanto, se vogliamo sapere qual è, p.e., la probabilità che si verifichino più di 3 eventi [$P(>3)$], dobbiamo calcolare l'area arancione, mentre l'area verde $P(\leq 3) = 1 - P(>3)$ ci dà la probabilità che si verifichino non più di 3 eventi (si veda il testo).

Myr[10] *($\lambda = 1$, utilizzando come unità di misura temporale il milione di anni). Consideriamo un lasso di tempo $T = 1{,}6$ Myr ($\lambda T = 1{,}6$). Mediante la poissoniana definita dall'eq. (3) (e illustrata in FIG. 4) possiamo calcolare le quantità*

$$P(0) \approx 0{,}20,\ P(1) \approx 0{,}32,\ P(2) \approx 0{,}26,$$
$$P(3) \approx 0{,}14,\ P(4) \approx 0{,}06,\ P(5) \approx 0{,}02$$

[10] Per brevità, utilizziamo l'acronimo inglese Myr per indicare il milione di anni.

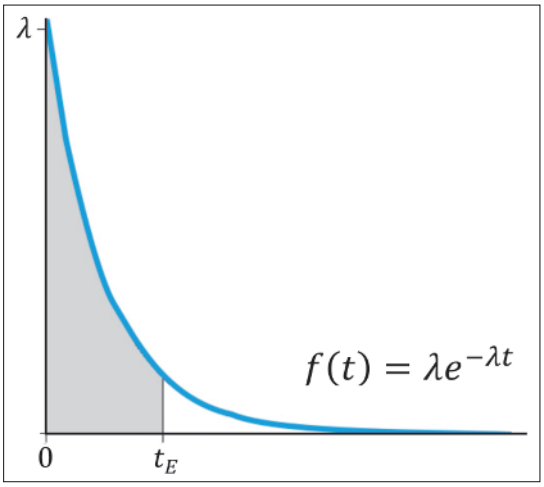

FIG. 5. La distribuzione esponenziale delle probabilità dei tempi di interarrivo (la cui formula è riportata in figura) è rappresentata dalla linea blu che si estende (teoricamente) fino a $t = \infty$. In base a semplici nozioni di calcolo integrale, l'intera area delimitata dalla linea blu e l'ascissa risulta essere pari a 1. D'altra parte è assolutamente certo che prima o poi si verifichi un evento successivo al precedente (a partire dal quale viene calcolato il tempo indicato in ascissa); pertanto, la probabilità di realizzazione di quest'ultimo evento vale 1. La coincidenza numerica tra area e probabilità non è casuale perché, come abbiamo visto in FIG. 4, il collegamento tra queste due grandezze scaturisce quando abbiamo a che fare con distribuzioni di probabilità (esponenziale in questa figura, poissoniana nella FIG. 4). Possiamo allora fissare un tempo t_E e calcolare la probabilità che l'evento accada a tempi superiori a questo. Per quanto detto, tale probabilità è data dalla regione bianca dell'area totale e vale $e^{-\lambda t_E}$; pertanto $P(t > t_E) = e^{-\lambda t_E}$. Naturalmente, la probabilità che l'evento si verifichi entro t_E è data dall'area grigia che si ottiene sottraendo l'area bianca da quella totale: $P(\leq t_E) = 1 - e^{-\lambda t_E}$.

che rappresentano, rispettivamente, le probabilità che in 1,6 Myr avvengano 0 oppure 1 o 2 o 3 o 4 o 5 impatti. Si noti che la somma delle varie probabilità si avvicina ad 1 all'aumentare del numero di eventi n; questo è abbastanza intuitivo perché è certo (probabilità pari a 1) che un qualche valore di n (compreso $n = 0$) deve verificarsi nell'intervallo di tempo prefissato. Possiamo quindi facilmente calcolare, p.e., qual è la probabilità che si verifichino più di tre impatti sottraendo alla somma di tutte le probabilità le probabilità per

$$n = 0,\ 1,\ 2,\ 3:$$
$$P(>3) = 1 - P(0) - P(1) - P(2) - P(3) \approx 0{,}08;$$

pertanto, la probabilità che in 1,6 Myr si verifichino più di 3 impatti è pari all'8%. Naturalmente la probabilità che avvengano al massimo tre impatti è pari a $P(\leq 3) \approx 1 - 0{,}08 = 0{,}92$, ossia $P(\leq 3) = 92\%$.

Ovviamente, un impatto ogni Myr è una media; in realtà si può dimostrare (ma noi non lo faremo) che i tempi di interarrivo (ossia i tempi tra due eventi consecutivi) legati ad una poissoniana hanno una distribuzione probabilistica esponenziale negativa (FIG. 5):

$$f(t) = \lambda e^{-\lambda t}. \qquad (4)$$

Possiamo allora chiederci qual è la probabilità che il successivo impatto avvenga dopo un'attesa dal precedente maggiore di, p.e., 0,8 Myr. Da quanto discusso nella didascalia di Fig. *5 abbiamo* $P(t > 0{,}8) = e^{-0{,}8} \approx 0{,}45$, *ossia 45%. La probabilità che invece l'impatto avvenga dopo un'attesa inferiore o pari a 0,8 Myr è ovviamente* $P(t \leq 0{,}8) = 1 - e^{-0{,}8} \approx 0{,}55$, *ossia 55%.*

In conclusione, l'eq. (3) conta il numero di eventi discreti in un intervallo temporale definito, mentre l'eq. (4) misura il tempo di attesa tra due eventi successivi. Quanto detto per l'esempio dell'asteroide vale anche per il numero di entrate in una fila (con tasso medio λ*) e per il numero di servizi (con tasso medio* μ*).*

Veniamo finalmente agli indici di prestazione che descrivono le caratteristiche di un sistema di coda. Essi dipendono essenzialmente da tre parametri che qui riassumiamo:

λ = *numero medio di arrivi in un determinato intervallo di tempo;*
μ = *numero medio di utenti serviti nello stesso intervallo di tempo;*
m = *numero di serventi.*

Riportiamo ora solo alcuni degli indici di prestazione (quelli utili per gli esempi che daremo) derivanti dallo studio del sistema più semplice, caratterizzato da una sola coda FIFO *stabile e stazionaria (ossia* $\lambda < \mu$, *come abbiamo visto nel livello base) e da un unico servente (*$m = 1$*):*

1) *Il numero L medio di utenti nel sistema*

$$L = \frac{\lambda}{\mu - \lambda};$$

2) *Il numero medio* L_q *di utenti in coda che attendono di giungere al servente per essere serviti*

$$L_q = L - \frac{\lambda}{\mu};$$

3) *Il tempo medio W che un utente spende nel sistema*

$$W = \frac{L}{\lambda};$$

4) *Il tempo medio* W_q *che un utente spende aspettando nella coda* L_q

$$W_q = \frac{L_q}{\lambda}.$$

La prima equazione, malgrado la sua semplicità, deriva, come detto, da una matematica alquanto complessa che riposa sul tipo di statistica in atto (poissoniana nel nostro caso) ed è fondamentale per la comprensione delle prestazioni di una coda con servente singolo. La seconda equazione è piuttosto intuitiva. Essendo μ *il tasso medio di servizio per singolo servente,* $1/\mu$ *è il tempo medio per servire il singolo utente. Se gli utenti arrivano ad un tasso* λ *e ciascuno di essi necessita di un tempo* $1/\mu$ *per essere servito, mediamente abbiamo* $\lambda \times (1/\mu) = \lambda/\mu$ *utenti che vengono serviti e vanno sottratti alla totalità degli utenti nel sistema per ottenere quelli in attesa in coda. Le ultime due equazioni sono sostanzialmente l'espressione della legge di Little che, ricordiamo, si applica sia a tutto il sistema (L) che ad una sua parte (*L_q*).*

Chiariamo con un esempio pratico l'utilità delle formule appena elencate e la loro efficacia nel migliorare il sistema di code a fini economici. Immaginiamo che un gestore di un'officina abbia alle sue dipendenze un meccanico (il servente) in grado di istallare 3 marmitte in 1 ora e che i clienti giungano ad un tasso di 2 in un ora. Abbiamo dunque:

λ = *2 automobili in arrivo ogni ora;*
μ = *3 automobili sistemate ogni ora;*
m = *1 meccanico.*

In base alle equazioni precedenti abbiamo allora:

L = *2 auto presenti mediamente nel sistema;*
L_q = *1,33 auto mediamente in fila in attesa di essere servite;*
W = *1 ora spesa mediamente da una auto nel sistema;*
W_q = *2/3 di ora spesa mediamente da una auto in attesa di essere servita.*

Immaginiamo ora che il nostro gestore voglia ottimizzare il sistema per incrementare i guadagni. Come abbiamo accennato nel livello base, vi sono due tipi di costi nel sistema: il costo di servizio e il costo di attesa dovuto al numero di clienti insofferenti che rinunciano a mettersi in fila o la abbandonano prima di giungere ad usufruire del servizio. Il primo è dato da mC_S *dove* C_S *è il costo orario di un singolo servente. Il costo di attesa orario si esprime mediante la formula* $\lambda W_q C_W$, *ossia il costo di attesa per unità di tempo di un singolo cliente* C_W *moltiplicato per il tempo di attesa e per il numero di arrivi per unità di tempo.*[11] *Pertanto, il costo totale è* $mC_S + \lambda W_q C_W$. *Ipotizzando approssimativamente* $C_W = 50$ *euro/ora e un costo del meccanico di* $C_S = 15$ *euro/ora, si ottiene un costo di attesa giornaliero (8 ore lavorative) pari a* $8 \times 2 \times (2/3) \times 50 \approx 533$ *euro, mentre quello del meccanico è* $15 \times 8 = 120$ *euro. Pertanto il costo totale del sistema è di* ~ 653 *euro giornalieri.*

Per migliorare il sistema, il gestore dell'officina pensa di sostituire il meccanico con un altro più

[11] Il costo di attesa potrebbe essere dato anche da $\lambda W C_W$, ma è irragionevole che un cliente abbandoni il sistema quando ha ormai raggiunto l'agognato servente.

veloce, capace di istallare 4 marmitte in un'ora. Avremo pertanto

$\lambda = 2$
$\mu = 4$
$m = 1$

da cui otteniamo:

$L = 1$
$L_q = 0{,}5$
$W = 0{,}5$
$W_q = 0{,}25.$

Il costo orario di attesa rimane di 50 euro e quello giornaliero diventa $8 \times 2 \times 0{,}25 \times 50 = 200$ euro; il nuovo meccanico è più costoso del precedente ($C_S =$ 20 euro/ora) e il costo di servizio giornaliero diventa $20 \times 8 = 160$ euro. Il costo totale è pertanto pari a $200 + 160 = 360$ euro. L'assunzione del nuovo meccanico (anche se più costoso) porta quindi ad un risparmio di $653 - 360 \approx 293$ euro al giorno.

Ma ridurre il tempo di attesa aumentando C_S non è l'unico modo per diminuire il costo totale del sistema. Il gestore potrebbe mantenere C_S al livello più basso ($C_S = 15$ euro/ora), ma aggiungere un secondo meccanico, anche se questo aumenta il costo del servizio a causa della paga da versare al nuovo assunto. Ciò nonostante, questa via può rivelarsi più efficiente. Prima di dimostrarlo, però, è necessario sottolineare che cambiare il numero di serventi (ossia m) cambia le caratteristiche della coda. Scriviamo gli indici di prestazione in questo caso:

$\lambda = 2$
$\mu = 3$
$m = 2.$

1) *Il numero medio L di utenti nel sistema*

$$L = !;$$

2) *Il numero di utenti L_q che rappresenta il numero di utenti in coda che attende di giungere al servente per essere serviti*

$$L_q = L - \frac{\lambda}{\mu};$$

3) *Il tempo medio W che un utente spende nel sistema*

$$W = \frac{L}{\lambda};$$

4) *Il tempo medio W_q che un utente spende aspettando nella coda L_q*

$$W_q = \frac{L_q}{\lambda}.$$

Apparentemente, sembra che sia cambiato ben poco dal caso precedente se non fosse per il punto esclamativo nella prima equazione. In effetti, per $m > 1$ la formula per ottenere L è assai complicata e noi evitiamo di riportarla (mediante l'escamotage del punto esclamativo) sia per evitare di spaventare il lettore, sia perché è inutile (a questo livello) perdersi in tecnicalità eccessive; quel che per noi conta è capire come sia possibile migliorare il rendimento di un sistema di code "giocando" con i parametri di base.

Torniamo ora a fare i conti in tasca al nostro gestore:

$L = 0{,}75$
$L_q = 0{,}083$
$W = 0{,}375$
$W_q = 0{,}0415.$

(si noti che abbiamo sostituito il punto esclamativo con il valore derivante dalla formula per L valida in questo caso). Fermo restando un costo di attesa orario di 50 euro, si ottiene un costo di attesa giornaliero (8 ore lavorative) pari a $8 \times 2 \times 0{,}0415 \times 50 =$ 33 euro. Il costo totale giornaliero del servizio $8 \times 2 \times 15 = 240$ euro. Pertanto il costo totale del sistema è $33 + 240 = 273$. Dunque l'assunzione di un secondo meccanico risulta essere l'opzione più conveniente (il lettore può verificare che la convenienza rimarrebbe – anche se assai minore – pure nel caso in cui il gestore mantenesse il costo $C_S =$ 20 euro/ora).

Annibale D'Ercole si è laureato in Fisica all'Università di Roma "La Sapienza". Astronomo associato presso l'INAF · Osservatorio di astrofisica e scienza dello spazio di Bologna (OAS), si occupa di simulazioni numeriche di idrodinamica, applicate alle nebulose e al gas interstellare delle galassie. È autore di numerosi articoli divulgativi pubblicati presso questa e altre riviste.