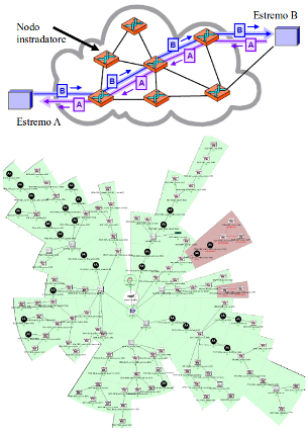
Highlighter

Analisi statistica del traffico della rete geografica wireless relativa al Progetto ADD (Anti Digital Divide).†

Augusto Pifferi,[a] Stefano Mastropasqua,[b] Gaetano Campi[a] and Andrea Lora[a]



Questo lavoro è stato realizzato con lo scopo di condurre una prima analisi statistica del traffico della rete geografica wireless gestita dall'Istituto di Cristallografia presso l'Area della ricerca Roma 1 – C.N.R., che si estende nel territorio della sabina romana e reatina. In primo luogo si è voluto verificare tramite serie temporali che il modello aderente alla realtà sperimentale fosse quello auto-similare; condizione determinata dal meccanismo di trasmissione a commutazione di pacchetto, tipico di tutte le reti informatiche. Contemporaneamente è stato accertato il fallimento della statistica di Poisson, che si adatta al comportamento dei dati nelle comuni linee telefoniche ed è associata al metodo di comunicazione tramite commutazione di circuito. E' stata in seguito valutata, tra due possibili scelte, la distribuzione di probabilità che meglio descrive analiticamente le densità sperimentali elaborate, su diverse scale dei tempi di osservazione secondo la definizione di auto-similarità. Data la particolare dinamica di diffusione dei pacchetti di broadcast, la stessa analisi è stata ripetuta separatamente per questo tipo di traffico, anche per verificare che tale componente non costituisse un fattore di degrado delle prestazioni della rete.
Keywords: Networking, Auto-similare, Poisson, Wireless.

1 Introduzione

Un importante campo di studi nell'ambito dello sviluppo delle reti di comunicazione riguarda il confronto fra modelli statistici in grado di descrivere le principali caratteristiche del flusso delle informazioni. Le analisi possono essere svolte con due finalità di base:

- diagnostiche, per studiare il comportamento del traffico o verificare come una rete reagisce quando viene sottoposta a particolari condizioni di utilizzo;
- di progetto, quando si vogliono convalidare algoritmi e protocolli per gestire i pacchetti in transito o si vuole intervenire sulla topologia di un network oppure definire le funzionalità dell'hardware di rete.

E' quindi essenziale che la teoria di riferimento rifletta il più possibile le proprietà del flusso dei dati che si suppone voglia rappresentare.

Il primo modello, basato sulla statistica di Poisson, viene concepito nell'ambito della telefonia, dove le chiamate in arrivo ad un centralino possono essere considerate indipendenti ed aventi una frequenza media costante. Sebbene inizialmente idoneo e analiticamente semplice, tale strumento si rivela non adatto a descrivere le caratteristiche del traffico nelle moderne reti locali (*LAN*), metropolitane (*MAN*) e geografiche (*WAN*), dove correlazione di eventi e rapida variabilità del flusso diventano fattori determinanti. L'utilizzo di nuovi modelli basati sull'auto-similarità e le relative distribuzioni è divenuto quindi predominante.[1]

L'indagine statistica presentata in questo documento è stata condotta su una rete wireless geografica ad accesso pubblico, per ottenere una descrizione analitica del comportamento del traffico e valutare, come specificato in seguito, l'incidenza della componente di broadcast.

2 Modellizzazione del traffico

Il traffico di una rete informatica può essere visto come una sequenza di arrivi di entità discrete, i pacchetti. Matematicamente si possono utilizzare due rappresentazioni: *counting process* e *interarrival time process*. Un *coun-*

[a] *CNR - Istituto di Cristallografia, Strada Provinciale 35/d, Montelibretti, Italia*
[b] *Università di Roma "La Sapienza", Facoltà di Fisica, P.le Aldo Moro, 9, 00185 Roma, Italia.*

† Il contenuto di questo documento costituisce una sintesi della tesi di laurea in Fisica svolta dal Dott. Stefano Mastropasqua presso la Facoltà di Scienze Matematiche, Fisiche e Naturali dell'Università degli Studi di Roma "La Sapienza" (A.A. 2012-2013).(registrato come rapporto tecnico IC/RM/2014/05 con numero di protocollo IC 1016 del 10/06/2014).

ting process {N(t)} è un fenomeno statistico funzione del tempo, a valori interi dove N(t) esprime il numero di arrivi per "t" compreso in un definito intervallo temporale. Un *interarrival time process* $\{A_n\}$ è una sequenza aleatoria di valori non negativi, dove $A_n = T_n - T_{n-1}$ indica la durata dell'intervallo che separa gli arrivi "n-1" e "n". Si supponga ad esempio di voler registrare il numero di pacchetti che transitano attraverso un determinato nodo della rete. Una rappresentazione di tipo *counting process* ad intervalli regolari di 0,25 s può essere schematizzata come segue:

n:	-	1	2	3	4	...
$t_n(s)$:	0	0,25	0,50	0,75	1,00	...
N(t):	-	3	2	4	2	...

La prova è stata realizzata definendo t = 0,25 s costante, mentre N(t) assume valori casuali ad ogni misura. Questo metodo si può adottare quando lo strumento di raccolta rileva i dati mediante un ciclo di polling cioè in tempi t_n equidistanti fra loro, oppure benché sia in grado di prelevare ogni singolo pacchetto, si desidera effettuare un campionamento ad intervalli fissati. In una rappresentazione di tipo interarrival time process la precedente sequenza viene tradotta come riportata in tabella 1.

Ad ogni singolo arrivo si registra il rispettivo tempo e si determina il valore di A_n . Sono stati evidenziati gli istanti nei quali veniva rilevato N(t) nel counting process.

Il modello statistico da impiegare per l'analisi dipende dalla natura di $\{N(t)\}$ e $\{A_n\}$. Una proprietà importante di tale modello deve essere la sua efficacia nel descrivere la forte variabilità del traffico (*traffic burstiness*). In particolare, una successione di arrivi nel tempo è fortemente variabile se gli istanti T n tendono a costituirsi in gruppi distinti (*clusters*), cioè se la corrispondente A n può essere suddivisa in sottosequenze di intervalli di arrivo alternativamente brevi ed estesi. Matematicamente, l'alta variabilità del traffico è legata ad un'autocorrelazione a lungo termine (*long term autocorrelation*) fra gli intervalli di arrivo.

Tabella 1: Risultati del interarrival time process.

n:	-	1	2	3	4	5	6	7	8	9	10	11	...
$T_n(s)$:	0	0,10	0,17	0,25	0,40	0,50	0,55	0,63	0,71	0,75	0,88	1,00	...
$A_n(s)$:	-	0,10	0,07	0,08	0,15	0,10	0,05	0,08	0,08	0,04	0,13	0,12	...

Non esiste tuttavia una singola definizione di alta variabilità generalmente accettata, a tale proposito vengono impiegati diversi parametri di riferimento come il "coefficiente di variazione" (*coefficient of variation*) $C_A = \sigma[A_n]/E[A_n]$ degli intervalli di arrivo, dove "σ" è la deviazione standard ed "E" il valor medio. In secondo luogo si può valutare "l'indice di dispersione dei conteggi" (*index of dispersion for counts, IDC*); dato un intervallo di tempo "τ" si definisce $IDC(\tau) = Var[N(\tau)]/E[N(\tau)]$, dove "Var" indica la varianza. Infine, come vedremo in seguito, il "parametro di Hurst" può essere usato come misura della variabilità in caso di traffico "auto-similare".[1]

3 Il modello di Poisson

Il modello di Poisson è stato uno dei primi strumenti statistici impiegati. Introdotto nell'ambito della telefonia da A. K. Erlang ha riscosso un enorme successo per la sua particolare attitudine nel descrivere le proprietà del traffico delle chiamate in una rete telefonica pubblica. E' stato quindi assunto come riferimento nella progettazione di nuove strutture e nell'integrazione di quelle già esistenti. Quando i canali preposti sono stati adattati per sostenere il flusso dei dati di reti informatiche in crescente evoluzione, la statistica di Poisson è stata di nuovo vista come naturale strumento di analisi, tuttavia la sua applicazione in questo ambito si è rivelata inaspettatamente inadeguata e si è resa necessaria un'indagine più approfondita sulla fenomenologia dei network in fase di sviluppo, per formulare nuove ipotesi che permettessero di individuare un modello alternativo.

3.1 Descrizione

Il traffico viene caratterizzato assumendo che gli intervalli di arrivo A n abbiano le seguenti proprietà:

- sono indipendenti;
- sono distribuiti esponenzialmente con parametro λ generalmente costante (frequenza media degli eventi); la probabilità "P" che risulti $A_n \leq t$ è data da: $P\{A_n \leq t\} = 1 - e^{-\lambda t}$.

In altre parole ciò equivale a descrivere il fenomeno attraverso un counting process associato alla distribuzione di probabilità detta "poissoniana"

$$P\{N(t) = n\} = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \qquad (1)$$

dove N(t) è il numero di arrivi nell'intervallo di tempo "t". I processi "poissoniani" godono delle seguenti

proprietà analitiche:

- la sovrapposizione di più processi indipendenti con parametri $\lambda_1, \lambda_2, \ldots, \lambda_n$ genera un nuovo processo con parametro $\lambda_1 + \lambda_2 + \cdots + \lambda_n$;
- li arrivi registrati in intervalli di tempo disgiunti sono statisticamente indipendenti. Questa proprietà viene anche detta "degli incrementi indipendenti" (*independent increments property*) e rende i processi poissoniani "senza memoria" (*memoryless processes*);
- per una distribuzione $P\{N(t) = n\}$ con parametro λ, la media e la varianza sono pari a λ. Ciò porta ad un coefficiente di variazione unitario.

Ci sono diversi modi per verificare se un particolare processo è poissoniano. Osservando che

$$P\{A_n = t\} = \frac{d}{dt} P\{A_n \leq t\} = \lambda e^{-\lambda t} \qquad (2)$$

un semplice metodo visivo consiste nel disegnare un grafico in cui si riporta il tempo "t" in ascisse, la probabilità $P\{A_n = t\}$ sulle ordinate per verificare se l'andamento che ne risulta decresce esponenzialmente. Oppure, dato che dalla (2) si ricava

$$log\{P\{A - n = t\}\} = log\{\lambda\} - \lambda t \qquad (3)$$

un grafico con scala logaritmica sull'asse "y" deve mostrare un andamento lineare. In tal caso il parametro λ può essere ricavato dall'intercetta della retta o dal suo coefficiente angolare.

Un caso particolare del modello di Poisson è quello in cui il valor medio dipende dal tempo (time-dependent Poisson process), adatto a descrivere i fenomeni in cui il parametro λ viene espresso come funzione del tempo: $\lambda = \lambda[t]$.[1]

3.2 Traffico telefonico vocale e dati: potenzialità e limiti del modello di Poisson

Una delle principali proprietà del traffico telefonico vocale è quella di essere relativamente omogeneo e prevedibile, inoltre la fenomenologia correlata ha dei tempi caratteristici abbastanza lunghi. Di conseguenza diverse chiamate concorrenti possono essere gestite in modo da condividere un mezzo di collegamento comune, riservando a ciascuna di esse una determinata frazione della capacità totale di trasmissione. E' quindi semplice valutare se un dato ramo o nodo della rete è in grado di sostenere il carico di una nuova richiesta di accesso; la tecnica utilizzata per l'eventuale attivazione è detta "a commutazione di circuito" (*circuit switching*).[2] Viene stabilita e mantenuta per tutta la durata della comunicazione una connessione tra due dispositivi; un cammino fissato che il traffico, instradato da diversi nodi, percorre dalla sorgente alla destinazione e di cui si tiene costantemente traccia. Questo circuito può essere riservato esclusivamente

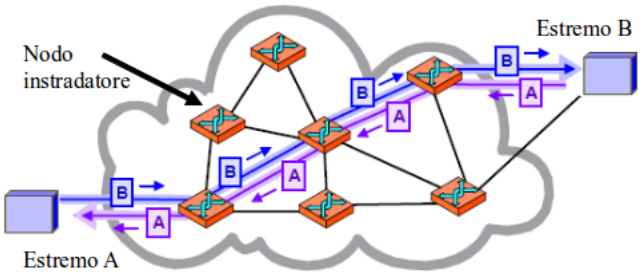


Fig. 1 Rete a commutazione di circuito.

ad un chiamante e ad un ricevente che lo usano quando si mettono in contatto, oppure rappresentare uno dei tanti percorsi attivabili in base alla distribuzione del traffico sui vari rami.[3] Nel secondo caso si parla di "circuito virtuale" poiché la rete si comporta come se fornisse un collegamento diretto fra due estremi (Fig. 1).

Il meccanismo appena descritto è alquanto semplicistico, perché rende tutte le richieste di accesso alla rete equivalenti tra loro in termini di banda di trasmissione da allocare e in merito ai percorsi necessari per soddisfare le diverse chiamate. Queste ultime inoltre sono separate da intervalli di tempo non correlati e la coppia sorgente–destinatario che si attiva in ogni istante è totalmente casuale. Una siffatta fenomenologia "senza memoria" è il motivo del successo iniziale del modello di Poisson, applicabile ad eventi che si verificano successivamente ed indipendentemente in un dato arco temporale, per i quali sia nota la frequenza media.

Diversamente dal traffico vocale, il flusso dati nei processi informatici è notevolmente variabile nei tempi e nelle velocità di trasmissione. Durante una connessione ogni informazione scambiata viene scomposta in blocchi o pacchetti inviati separatamente uno dall'altro, che possono seguire percorsi distinti attraverso i vari rami pur avendo emittente e destinatario in comune. Questo metodo di comunicazione viene chiamato "a commutazione di pacchetto" (*packet switching*) e rappresenta un punto di svolta nei meccanismi di gestione delle reti. Ciascun pacchetto, oltre ai dati, contiene un'intestazione (*header*) con i dettagli necessari al suo instradamento e utilizzati da ogni nodo, o *router*, per l'invio al successivo punto di scambio fino alla destinazione, dove i blocchi vengono assemblati per ricostruire il messaggio originale. Di conseguenza i router non tengono traccia di ogni collegamento attivo, si occupano solo di spedire il traffico in arrivo al nodo seguente e in caso di perdita di dati con annessa ritrasmissione, non hanno problemi a reinstradare informazioni già elaborate (Fig. 2).

Il passaggio dal circuit switching al packet switching ha profonde implicazioni. Ciascun pacchetto compete con gli altri per l'accesso ad uno o più rami della rete. Una connessione che si serve di un percorso lungo il quale il traffico non è particolarmente intenso, può utilizzare l'in-

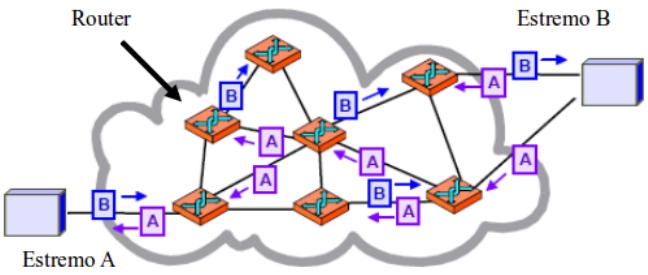


Fig. 2 Rete a commutazione di pacchetto.

tera banda di trasmissione disponibile e trasferire i suoi dati molto velocemente; se più connessioni condividono gli stessi rami la capacità viene ripartita per soddisfare tutti gli scambi. Inoltre la commutazione di pacchetto rende le reti in grado di instradare il traffico attraverso un percorso alternativo se un ramo smette di funzionare, senza interrompere i collegamenti attivi; una realtà ben diversa da quella della commutazione di circuito, in cui nell'eventualità di guasti la connessione cade e va ristabilita. Ciò non toglie che il sistema possa subire un sovraccarico quando i dati transitano ad una velocità tale da superare la capacità di trasmissione disponibile. I pacchetti in eccesso vengono temporaneamente memorizzati (buffered) nei router in attesa di poter essere spediti, tuttavia se viene raggiunta una condizione limite detta "congestione" (congestion), le unità di memoria si riempiono ed alcuni blocchi possono essere scartati o persi. Per assicurare un adeguato funzionamento dei dispositivi che inviano dati in una situazione di sovraccarico, i protocolli di comunicazione prevedono meccanismi di controllo della congestione (end-to-end congestion control) che diminuiscono automaticamente la velocità di trasmissione se viene rilevato un livello critico di attività. Ciò implica che il traffico venga modulato (shaped) in funzione dello stato in cui si trova o si è venuto a trovare ciascun segmento di rete ed introduce significative correlazioni temporali fra le connessioni attive.

Da quanto detto sul packet switching emerge chiaramente come il corrispondente flusso dei dati non sia un processo stazionario, ma costituito dall'alternarsi di momenti con basso, medio, alto utilizzo delle risorse e periodi di stasi; per descriverlo si usa comunemente l'aggettivo *bursty* (che cambia all'improvviso). Questo termine, pur avendo un significato intuitivo, si riferisce ad una proprietà che trova un riscontro concreto nel momento in cui si fissa una scala temporale nell'ambito della quale osservare il fenomeno. Tornando al traffico telefonico, si parte dal parametro λ della statistica di Poisson che descrive la dinamica delle chiamate in arrivo, se ad esempio $\lambda = 100\ s^{-1}$ la scala è dell'ordine di $1/\lambda = 10\ ms$. Ciò vuol dire che data un'intensità media del traffico, periodi di attività visibilmente superiore o inferiore si osservano con sempre maggiore difficoltà tanto più la scala

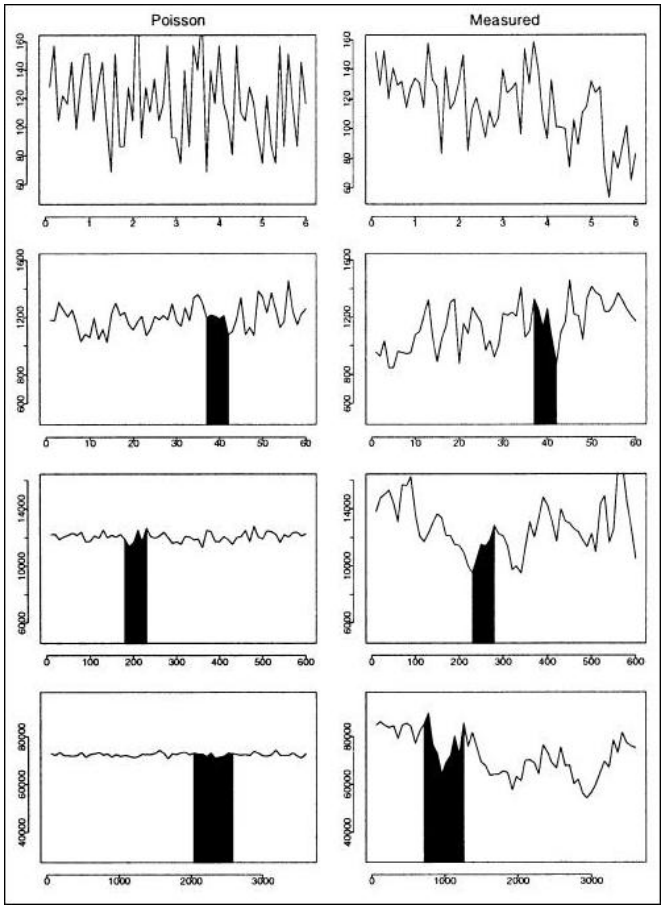


Fig. 3 Traffico di rete misurato (destra), modellizza- to secondo Poisson (sinistra).

si allontana da $1/\lambda$. In una rete informatica invece è stato ampiamente osservato che la variabilità del flusso dei dati tende ad apparire nel tempo su diversi ordini di grandezza, non adattandosi dunque al modello poissoniano. A titolo di esempio la Fig. 3 illustra visivamente tale incompatibilità.

I grafici sul lato destro rappresentano il traffico registrato in un'ora sulla connessione ad Internet della rete locale di una grande azienda. Supponendo che tali dati siano propri di un sistema poissoniano adattato al valor medio ed alla varianza sperimentali, sul lato sinistro è riportato il flusso che verrebbe di conseguenza generato. Le quattro righe dello schema corrispondono ad altrettante scale temporali o intervalli di campionamento. La prima in alto mostra in 6 secondi di registrazione scelti a caso il traffico campionato su una scala di 100 ms; ovvero la coordinata verticale di ciascun punto indica il numero di pacchetti osservati in un intervallo di 100 ms. La seconda riga si riferisce ad una scala temporale dieci volte più grande; ogni punto rappresenta il numero di pacchetti registrati in 1 s, su un totale di 60 s. Le aree scure mostrano quale sottoinsieme di dati ha generato il grafico della riga superiore. E' da notare che ad aumentare di un fattore 10 non è solo la scala sull'asse "x", anche quella sulle ordinate cambia in modo analogo.

Nella terza riga questo incremento viene ripetuto, mentre la quarta è dilatata di un fattore 6 per coprire l'intera ora.

La discrepanza tra il modello di Poisson e i dati sperimentali è lampante. Al crescere del tempo di campionamento il flusso poissoniano tende a regolarizzarsi fin quasi ad appiattirsi, rivelando la sua natura di processo senza memoria, ovvero scarsamente dipendente dalla successione degli eventi e degli stati che il sistema attraversa. Da un punto di vista statistico ciò si traduce in un preciso comportamento della funzione di autocorrelazione del traffico (numero dei pacchetti in transito). Fissato un intervallo di campionamento si registra infatti una decrescita esponenziale, quindi molto rapida, in funzione del tempo. Aumentando la scala di osservazione l'andamento degenera verso un valore nullo. Sul lato opposto il flusso reale mostra inequivocabilmente una notevole e persistente variabilità. Questo risultato ha importanti conseguenze pratiche. Un traffico che si comporta come illustrato nella colonna di sinistra è abbastanza semplice da controllare: oltre una data scala dei tempi la conoscenza dell'intervallo medio di arrivo dei pacchetti è sufficiente per poterlo descrivere. Non sono necessarie memorie di supporto (*buffer*) per i dispositivi di rete o strategie particolari per garantire un funzionamento efficiente. In netto contrasto la colonna di destra mostra un flusso difficilmente prevedibile su più scale temporali, ciò suggerisce l'impiego di memorie temporanee per tamponarne le improvvise fluttuazioni e di un meccanismo per prevenire la saturazione dei collegamenti che, incidendo sulle prestazioni, non sempre ne assicura un livello minimo.[2]

4 Modello auto-similare

La caratteristica principale del traffico telefonico, che rende efficiente il modello di Poisson, è la sua debole variabilità nello spazio e nel tempo; gli eventi coinvolti nei processi statistici sono indipendenti o hanno una correlazione temporale e una densità di probabilità che decadono esponenzialmente.

Ciò che si osserva nelle reti informatiche è invece un'estrema variabilità nel flusso dei dati. Quella spaziale viene descritta da distribuzioni, come quella di Pareto, con lunghe code (*heavy tailed distributions with infinite variance*) che assegnano probabilità non trascurabili anche a risultati lontani dal valore medio. Quella temporale si traduce in una dipendenza a lungo termine (*long range dependences*), ovvero in una funzione di autocorrelazione che decresce come una potenza di "t" (*power law decay*). Quando questi fattori si combinano, le relative grandezze manifestano in genere proprietà frattali, ovvero caratteri statistici che si ripetono su diverse scale di osservazione.[1]

In questa sezione introduciamo un modello alternativo detto "auto-similare" (*self-similar model*), riportandone un'analisi dettagliata e mostrando qualitativamente come riesce a descrivere il traffico di rete.

4.1 Definizioni e proprietà

Il concetto di auto-similarità nell'ambito delle comunicazioni fu introdotto da Mandelbrot a metà degli anni '60, tuttavia solo dopo il 1980 venne interpretato come potenziale strumento per la creazione di modelli in grado di descrivere la variabilità del traffico di rete. Quelle che seguono costituiscono le definizioni e le basi analitiche dell'auto-similarità.[4]

a) Un processo statistico stazionario $\{X_t\}$ (*stationary stochastic process*) è un fenomeno aleatorio la cui distribuzione di probabilità non cambia se traslata nello spazio o nel tempo, di conseguenza parametri come il valor medio $\mu = E[X_t]$ e la varianza $\sigma^2 = E[(X_t - \mu)^2]$, se esistono, rimangono costanti. In caso di non stazionarietà $\{X_t\}$ può assumere ad esempio valori X_{t_1} distribuiti con media μ_{t_1} e varianza σ_{t_1} all'istante t_1, valori t_2 con parametri μ_{t_2} e σ_{t_2} all'istante $t_2 > t_1$. Ciò implica che la funzione di autocorrelazione $r(t_1,t_2)$ dipende da due variabili ed è definita come

$$r(t_1,t_2) = \frac{E[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})]}{\sigma_{t_1}\sigma_{t_2}} \qquad (4)$$

Se il fenomeno è stazionario, avendo un'unica distribuzione di riferimento, l'autocorrelazione degli eventi dipende solo dalla differenza $\tau = t_2 - t_1$ e la (4) può essere riformulata nel modo seguente:

$$r(t_1,t_2) = \frac{E[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})]}{\sigma_{t_1}\sigma_{t_2}} \qquad (5)$$

b) Sia $X = (X_t : t = 1,2,3,\dots)$ un processo statistico stazionario (o serie temporale) con valor medio μ, varianza σ^2 e funzione di autocorrelazione $r(\tau)$, $\tau \geq 0$. In particolare si assuma che $r(\tau)$ abbia la forma

$$r(\tau) = \tau^{-\beta} L(\tau) \text{ per } \tau \to \infty \qquad (6)$$

dove $0 < \beta < 1$ e L è una funzione lentamente variabile all'infinito ovvero

$$\frac{L(\tau x)}{L(\tau)} = 1 \text{ per ogni } x > 0 \qquad (7)$$

c) Per ogni $m = 1,2,3,\dots$ sia

$$X^{(m)} = (X_k^{(m)} : k = 1,2,3,\dots)$$

un nuovo processo stazionario con funzione di autocorrelazione $r^{(m)}$, ottenuto mediando (o aggregando) la serie originale X su intervalli disgiunti di

dimensione “m”:

$$X_k^{(m)} = \frac{X_{(k-1)m+1} + \cdots + X_{km}}{m} \text{ con } k \geq 1 \qquad (8)$$

l’indice “m” rappresenta l’ordine di aggregazione dei dati iniziali mentre “k” è la cardinalità dell’intervallo.

d) Il processo X viene definito “esattamente auto-similare” (exactly second-order self-similar) con parametro $H = 1 - \beta/2$ se per ogni m=1, 2, 3,... risulta

$$Var[X^{(m)}] = \sigma^2 m^{-\beta} \text{ e } r^{(m)}(k) = r(k) \qquad (9)$$

X è detto “asintoticamente auto-similare” (*asymptotically second-order self-similar*) con parametro $H = 1 - \beta/2$ se per grandi valori di “k” risulta

$$r^{(k)}(k) \to r(k) \text{ con } m \to \infty \qquad (10)$$

con $r(k)$ che soddisfa la (6). In altre parole X è esattamente o asintoticamente auto-similare se il corrispondente processo $X(m)$ è indistinguibile da X o lo diventa per grandi valori di “m”, almeno in base alle rispettive funzioni di autocorrelazione. La quantità H (compresa tra 0,5 e 1) definisce il grado di auto-similarità del processo e viene anche detta “parametro di Hurst”.

Matematicamente l’auto-similarità si manifesta in diversi modi equivalenti tra i quali:

- La varianza di X (m) decresce più lentamente del reciproco di “m” (*slowly decaying variances*), ad esempio $Var[X^{(m)}] \sim a_1 m^{-\beta}$ per $m \to \infty$, con $0 < \beta < 1$ (a_1 è una costante positiva e limitata);
- Per “m” fissato, la funzione di autocorrelazione decresce con legge iperbolica anziché esponenziale ed è pertanto non sommabile (si veda sempre la (6)): $\sum_k r(k) = \infty$ (long range dependence).

un’altra evidente proprietà dei processi (esattamente o asintoticamente) auto-similari è che all’aumentare di “m” la funzione di autocorrelazione di $X^{(m)}$ non tende ad annullarsi, ovvero non è degenere. Questo aspetto viene qualitativamente mostrato nei grafici della colonna a destra in Fig. 3: se X rappresenta il numero di pacchetti rilevati ogni 100 ms (prima riga in alto), la seconda, terza e quarta riga riportano segmenti delle serie temporali $mX^{(m)}$ con m=10, 100, 600 (pacchetti registrati ogni 1 s, 10 s, 60 s) rispettivamente. E’ da notare che i relativi andamenti si distinguono chiaramente dal cosiddetto “rumore di fondo” (*pure noise*), il segnale che si prende come esempio di insieme di valori non correlati, dal quale cioè non emergono sequenze che si ripetono. In netto contrasto, i grafici della colonna a sinistra si riferiscono a processi $X^{(m)}$ per i quali $r^{(m)}(k) \to 0$ per $m \to \infty$; l’ultimo

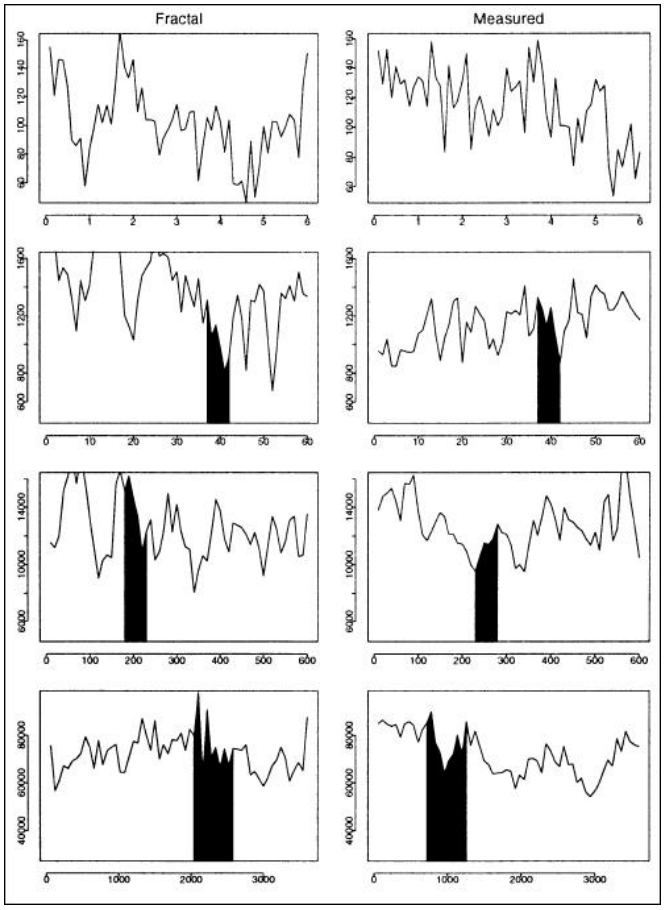


Fig. 4 Traffico di rete misurato (destra), descritto con un modello auto-similare (sinistra).

può essere considerato un tipico rumore di fondo. Questi fenomeni statistici sono caratterizzati dalle seguenti proprietà:

- La varianza di $X^{(m)}$ decresce come il reciproco di “m”: $Var[X^{(m)}] \sim a_2 m^{-1}$ per $m \to \infty$;
- La funzione di autocorrelazione decresce con andamento esponenziale ($r(k) \sim \rho^k$ con $0 < \rho < 1$) e risulta sommabile:$\sum_k r(k) < \infty$ (*short range dependence*).

Il modello auto-similare ha il notevole pregio di essere minimale, ovvero introduce l’unica ulteriore quantità H a quelle descrittive già utilizzate dalla statistica di Poisson. La Fig. 4 riporta lo stesso campione rappresentato in Fig. 3, ma i grafici sul lato sinistro mostrano il traffico generato da un sistema auto-similare che ha ricevuto in ingresso valor medio, varianza e parametro di Hurst dei dati sperimentali.

Come si osserva, la variabilità è ora preservata su tutte le scale temporali; ciò rispecchia il comportamento del flusso originale.

Uno dei metodi per determinare il parametro di Hurst si può dedurre dalla relazione (9) $Var[X^{(m)}] = \sigma^2 m^{-\beta}$ equivalente a:

$$\log(Var[X^{(m)}]) = \log(\sigma^2) - \beta \log(m) \qquad (11)$$

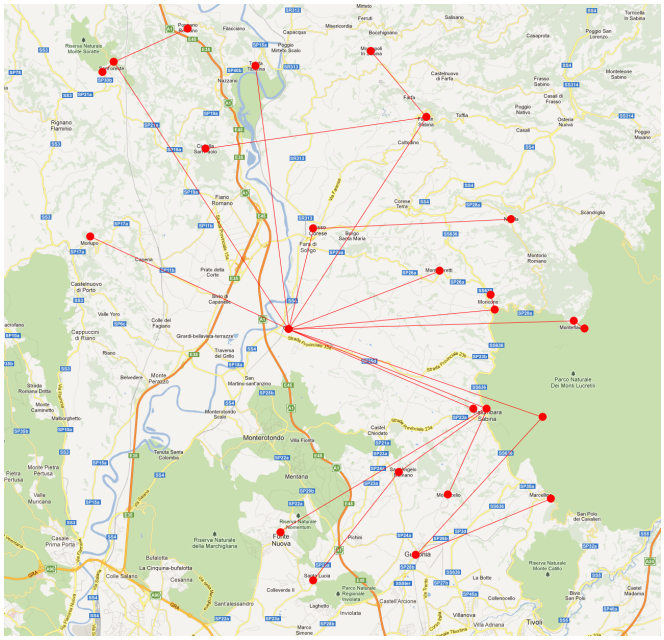


Fig. 5 Mappa geografica della WAN.

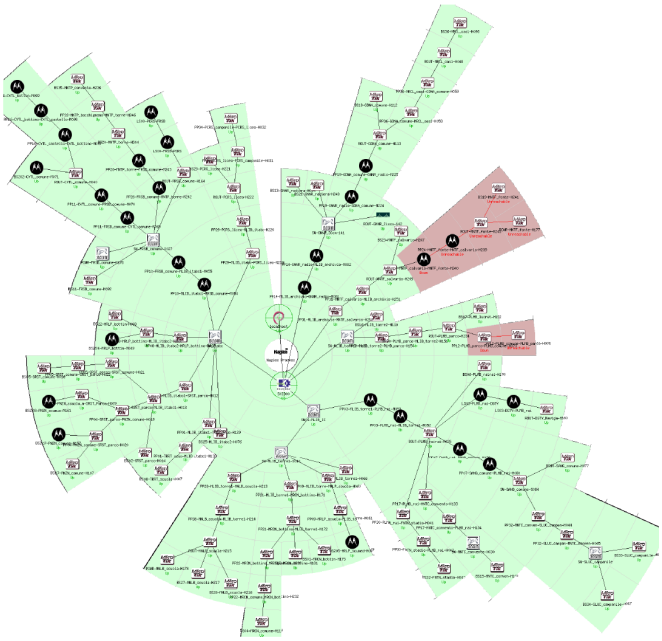


Fig. 6 Mappa tecnica della WAN.

Riportando in un grafico il logaritmo della varianza del processo $X^{(m)}$ in funzione del logaritmo dell'ordine di aggregazione "m", si ottiene un cosiddetto *variance-time plot* (diagramma tempo-varianza). In queste variabili, ricordando che $0 < \beta < 1$, i fenomeni auto-similari sono identificati da una retta con pendenza negativa compresa tra -1 e 0 ed intercetta data dal logaritmo della varianza del processo originale X. Dopo aver ricavato β con un procedimento di regressione lineare si può ottenere il parametro di Hurst dalla relazione $H = 1 - \beta/2$. In conclusione, quanto più piccolo è β (la retta è poco inclinata e H tende a 1) tanto maggiore è il grado di auto-similarità.

5 Descrizione della WAN relativa al Progetto Anti Digital Divide

L'Istituto di Cristallografia del Consiglio Nazionale delle Ricerche (CNR), Area della Ricerca Roma 1, situata presso il comune di Montelibretti (RM), gestisce una rete geografica wireless ad accesso pubblico che serve utenze dislocate nel territorio della sabina romana e reatina. Tra i comuni che delimitano la zona interessata si possono citare Guidonia, Palombara Sabina, Fara Sabina, Ponzano Romano, Morlupo, Fonte Nuova, in un'area di circa 1500 km^2 . Lo scopo di questa struttura è di fornire servizi di rete a banda larga ad utenze private ed istituzioni pubbliche come scuole ed uffici comunali, con particolare riguardo per l'accesso ad Internet.

La Fig. 5 mostra la copertura territoriale dei rami della dorsale, mentre in Fig. 5.2 viene riportata una mappa tecnica dettagliata con la disposizione degli apparati che supportano la rete nei vari nodi; come si può osservare è stata adottata una topologia a stella con centro nella sede del CNR.

I client che attualmente accedono al traffico Internet sono oltre 600. Le antenne per la trasmissione punto-punto lungo la dorsale sono rappresentate da quei nodi connessi a due soli rami o ad un singolo ramo nel caso dei dispositivi di frontiera. Il segnale che raggiunge un'antenna viene ricevuto anche da una base station posta nelle immediate vicinanze che, tramite un collegamento punto-multipunto lo invia alle utenze finali. I nodi da cui hanno origine più rami rappresentano switch che smistano il traffico verso i client o lo concentrano se è diretto all'origine della stella, ovvero nel Centro Elaborazione Dati (CED) del CNR da cui può raggiungere il gateway per l'accesso ad Internet.

Lo standard adottato per la propagazione del segnale radio è l'*HiperLAN* (High Performance Radio LAN), nato come risposta europea al *Wi-Fi* statunitense (*IEEE 802.11*), successivamente orientato anche all'utilizzo di reti più estese e sviluppato dall'*ETSI* (European Telecommunications Standards Institute), un organismo internazionale indipendente ufficialmente responsabile della definizione e dell'emissione di standard nel campo delle telecomunicazioni in Europa. In merito all'hardware impiegato e ai metodi operativi, l'HiperLAN è molto simile ad un'altra tecnologia nota come **WiMAX** (Wireless Interoperability for Microwave Access o *IEEE 802.16*), si affida al posizionamento di ponti radio sul territorio da servire garantendo una copertura di 30-40km intorno a ciascuna antenna. Le base station si collegano da un lato alla dorsale di rete del provider a sua volta connessa ad Internet, dall'altro attraverso l'etere ai dispositivi riceventi. Ovviamente il segnale portante deve garantire alte prestazioni a tutti gli utenti distribuiti in un'area dalle dimensioni lineari almeno 10 volte più grandi di quelle previste da una rete 802.11, di conseguenza le base station HiperLAN operano a potenze nettamente superiori

rispetto ai tipici access point Wi-Fi. In Italia vengono adottate frequenze comprese nell'intervallo 3,4-3,6GHz private, ovvero acquistate da specifici provider, oppure libere intorno ai 5,4GHz. La velocità massima alla quale una base station può trasmettere è di circa 70Mbps, relativa tuttavia alle distanze già menzionate, oltre le quali si ha un sensibile degrado delle prestazioni dovuto ad una non trascurabile incidenza degli errori. Gli utenti che si trovano ai limiti della zona servita riescono a connettersi, in condizioni di medio carico, con velocità comprese tra 1Mbps e 4Mbps. Occorre però considerare che, come in tutti i sistemi wireless, la banda disponibile viene condivisa dai dispositivi attivi in una determinata area, quindi le singole velocità diminuiscono all'aumentare delle stazioni riceventi collegate; con un numero ed un posizionamento adeguato delle antenne è possibile garantire delle prestazioni minime nell'intervallo di 4-8Mbps.

Dovendo svolgere funzioni di supporto per le attività di rete, l'HiperLAN è stato progettato per trasmettere pacchetti *IP* (Internet Protocol) attraverso l'etere e per connettersi ad una struttura cablata preesistente; le informazioni scambiate possono riguardare chiamate *VoIP* (Voice over IP), traffico peer-to-peer o streaming e qualsiasi altro servizio comunemente accessibile tramite Internet.

5.1 Limiti della rete

La topologia a stella impiegata non prevede le ridondanze tipiche di una rete a maglia, che fornirebbero una maggiore tolleranza alla caduta di uno o più rami oppure la possibilità per gli apparati di non occupare ulteriormente linee in cui si riscontra un'intensa attività, instradando i dati verso percorsi alternativi. Tuttavia l'aspetto da tenere più in considerazione è costituito dal traffico di broadcast, ovvero da tutti quei pacchetti che non hanno un destinatario specifico, ma contengono un particolare indirizzo che permette loro di arrivare a tutti i nodi raggiungibili a partire da quello che li ha generati, diffondendosi quindi sul maggior numero possibile di rami con il rischio di influire negativamente sulle prestazioni della rete se non adeguatamente controllati.

Tra le informazioni contenute in un pacchetto esiste un campo riservato al suo indirizzamento, cioè all'identificazione del destinatario. Occorre però separare due tipi di indirizzo:

- il *MAC address* (o indirizzo fisico), formato da sei coppie di cifre esadecimali (es. 1A EB 54 D1 69 CC), legato all'hardware di rete ed assunto quindi come riferimento da tutti quei dispositivi instradatori che operano al livello *data link* (il secondo) del modello ISO/OSI (hub, bridge, switch);
- l'*indirizzo IP*, rappresentato in forma decimale da quattro numeri compresi tra 0 e 255, separati da un punto (es. 79.204.151.36), assegnato all'intero host e valutato dai router, che operano al livello *network* (il terzo).

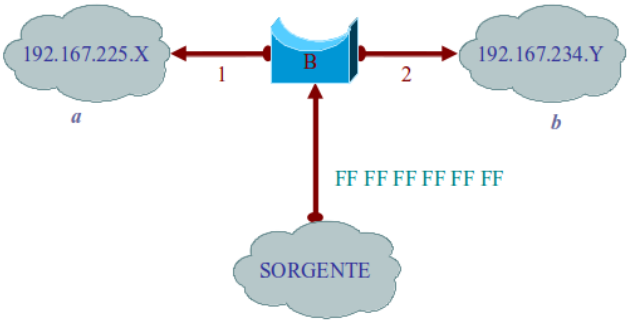


Fig. 7 Trasmissione di broadcast data link gestita da un bridge.

Il primo codice è presente in ogni pacchetto, il secondo può non esserlo in alcuni. I router instradano il traffico basandosi sull'indirizzo IP e bloccando i dati contenenti il solo MAC address, mentre hub, bridge e switch consentono la propagazione di tutti i pacchetti, in particolare di quelli che riportano un indirizzo di broadcast data link: FF FF FF FF FF FF. E' proprio questa la componente che va attentamente monitorata e possibilmente circoscritta se si vuole salvaguardare il buon funzionamento della rete.

Si può chiarire meglio il concetto considerando a titolo di esempio una LAN non molto estesa partizionata in tre sottoreti messe in comunicazione da un bridge, due delle quali identificate dai seguenti indirizzi IP: 192.167.225.X, 192.167.234.Y (X e Y distinguono gli host). Dalla terza subnet, una sorgente avvia una trasmissione di broadcast data link (Fig. 7).

Elaborando l'indirizzo fisico, il bridge non può stabilire in quale sottorete (la "a" o la "b") si trova il destinatario ed invia il messaggio ad entrambe, propagandolo anche in quella sicuramente non coinvolta, rallentandone il traffico ed occupando i rami "1" e "2".

Se si sostituisce il bridge con un router qualsiasi pacchetto che non abbia un indirizzo IP verrebbe bloccato. Potrebbe comunque verificarsi una trasmissione di tipo broadcast ma al livello network, designata da "255" come ultimo gruppo di cifre decimali (Fig. 8).

Leggendo l'indirizzo IP, il router invierebbe il messaggio solo alla subnet con cui trova una corrispondenza (la ".225" attraverso il ramo "1"), escludendo l'altra (la ".234" attraverso il ramo "2"); ne risulterebbe un traffico ottimizzato rispetto al primo caso. Se si estende l'esempio appena fatto ad una MAN o ad una WAN, si comprende come reti realizzate con dispositivi che operano al livello data link possono essere sottoposte ad un carico assai elevato, con pacchetti che invadono segmenti nei quali sicuramente i rispettivi destinatari non si trovano. E' proprio questo il caso della rete geografica oggetto del paragrafo. Gli apparati che la supportano potrebbero be-

nissimo instradare il traffico riferendosi ai protocolli del terzo livello, ma per motivi di connettività sono impostati per operare al secondo. Un pacchetto di broadcast emesso in qualsiasi istante da qualunque nodo si propagherebbe sull'intera WAN, che costituisce quindi un unico "dominio di broadcast". A questo proposito, tornando all'esempio citato, è di uso comune affermare che i router "spezzano" i domini di broadcast (Fig. 8 ogni sottorete costituisce un insieme separato dagli altri).

Leggendo l'indirizzo IP, il router invierebbe il messaggio solo alla subnet con cui trova una corrispondenza (la ".225" attraverso il ramo "1"), escludendo l'altra (la ".234" attraverso il ramo "2"); ne risulterebbe un traffico ottimizzato rispetto al primo caso.

Se si estende l'esempio appena fatto ad una MAN o ad una WAN, si comprende come reti realizzate con dispositivi che operano al livello data link possono essere sottoposte ad un carico assai elevato, con pacchetti che invadono segmenti nei quali sicuramente i rispettivi destinatari non si trovano. E' proprio questo il caso della rete geografica oggetto del paragrafo. Gli apparati che la supportano potrebbero benissimo instradare il traffico riferendosi ai protocolli del terzo livello, ma per motivi di connettività sono impostati per operare al secondo. Un pacchetto di broadcast emesso in qualsiasi istante da qualunque nodo si propagherebbe sull'intera WAN, che costituisce quindi un unico "dominio di broadcast". A questo proposito, tornando all'esempio citato, è di uso comune affermare che i router "spezzano" i domini di broadcast (Fig. 8 ogni sottorete costituisce un insieme separato dagli altri).

6 Raccolta ed elaborazione dati

Viene ora presentata la raccolta e l'elaborazione statistica dei dati di traffico relativi alla rete wireless descritta nella sezione precedente. L'acquisizione riguarda quattro giorni non consecutivi, due feriali e due domeniche, in ciascuno dei quali il traffico è stato monitorato nell'arco di 24 ore. Le informazioni su cui l'elaborazione si concentra sono, in ordine di priorità, il numero e le dimensioni dei pacchetti transitati in un nodo nel quale è stato possibile concentrare il flusso dati dell'intera WAN, il protocollo utilizzato per trasmettere ciascun pacchetto.

Osservando preliminarmente l'andamento del traffico sono stati isolati in ciascun giorno due intervalli temporali di alta e bassa intensità, scelti in modo da individuare altrettanti processi stocastici ragionevolmente stazionari. Nell'ambito di ogni intervallo è stata condotta un'indagine tesa a verificare la potenziale autosimilarità del flusso ed a confrontarla con una ipotetica natura poissoniana, per valutare quale modello descrive più accuratamente il traffico reale. Le funzioni scelte per interpolare i dati secondo la statistica auto-similare sono la distribuzione log-normale e la distribuzione di Weibull.[5] La prima si adatta alle variabili casuali che hanno un carattere molti-

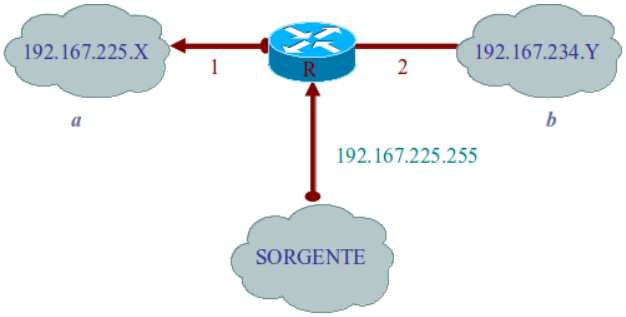


Fig. 8 Trasmissione di broadcast IP gestita da un router.

plicativo, assumono cioè valori come prodotto di più fattori da cui hanno origine. La seconda descrive bene quegli eventi in cui il rapporto tra successi ed insuccessi in un intervallo di tempo "t" dipende dall'intervallo stesso. L'elaborazione ha quindi due obiettivi:

- stabilire quale teoria è più efficace nel modellizzare il flusso dei dati: quella poissoniana o quella autosimilare;
- nel secondo caso definire quale forma analitica descrive meglio le frequenze sperimentali: la distribuzione log-normale o la distribuzione di Weibull.

L'ultima parte dell'indagine riguarda il traffico di broadcast. Questa componente è naturalmente presente in tutte le reti, a prescindere dall'estensione, ma per sua natura va debitamente circoscritta poiché rischia di saturare le linee di trasmissione se aumenta oltre una certa soglia. La WAN analizzata non adotta meccanismi di controllo in tal senso, rappresenta cioè un unico dominio di broadcast (si veda la sezione 5). Tuttavia al suo interno come verrà messo in evidenza, questo sottoinsieme del traffico costituisce meno di qualche unità per mille del flusso totale. Nonostante la ridotta presenza, l'elaborazione condotta sui soli pacchetti di broadcast ha fornito risultati quasi in netto contrasto con quelli ottenuti sull'intera popolazione, ma logicamente deducibili se si pensa al meccanismo di propagazione di tale componente.

6.1 Strumenti hardware e software

All'interno del Centro Elaborazione Dati (CED) dell'Area della Ricerca Roma 1 sono stati allestiti due server: uno per l'acquisizione e la conservazione, l'altro per l'analisi dei dati. Il primo è costituito da due processori fisici con due core logici ciascuno, frequenza pari a 2,66GHz, 2GB di memoria RAM, un hard disk locale con 36GB di capacità e due schede di rete da 1Gbps: una per il controllo (management) del server stesso da una postazione remota tramite protocollo SSH (*Secure Shell*) e per l'invio dei dati alla macchina elaboratrice con protocollo SFTP (*Secure File Transfer Protocol*), l'altra dedicata alla cattura del traffico della WAN (monitor). I pacchetti prelevati so-

no stati memorizzati in tempo reale su un secondo hard disk da 1,8TB collegato al server attraverso un'interfaccia di tipo USB. Come piattaforma locale è stato utilizzato un sistema operativo Debian 3.2.46 con kernel Linux 3.2.0. Sul secondo server è stato installato un gestore di macchine virtuali VMware 5.5 per l'emulazione di un sistema operativo Windows 7 Professional a 64 bit supportato da due processori fisici, ciascuno dotato di quattro core logici, frequenza pari a 2GHz, 8GB di RAM, un hard disk locale da 180GB ed una scheda di rete da 1Gbps per la ricezione tramite protocollo SFTP dei dati da elaborare e per il controllo remoto con protocollo RDP (*Remote Desktop Protocol*).

Il software utilizzato per l'acquisizione del traffico sul primo server, TCPdump 4.3.0 con librerie libpcap 1.3.0, rientra nella categoria dei *packet sniffer* ovvero quei programmi che, ponendosi in ascolto su una determinata porta di un dispositivo di rete (ad esempio uno switch), riescono a prelevare le informazioni contenute nei pacchetti che transitano per quel nodo. I dati così ottenuti sono stati archiviati in un insieme di file binari classificati con estensione .pcap (*packet capture*) e successivamente tradotti in documenti di testo attraverso funzionalità accessorie dello stesso software. Per una analisi preliminare finalizzata ad ottenere i primi riscontri sperimentali è stato invece utilizzato, sul secondo server, un programma scritto in linguaggio Visual Basic 6.0, in grado di ricevere in ingresso i suddetti file di testo e tramite opportune scelte dell'utente elaborare i dati ivi contenuti e produrre in uscita: informazioni di carattere generale relative al traffico (numero di pacchetti catturati, dimensioni, tempo di acquisizione ecc.), rappresentazioni grafiche della cronologia del flusso o di distribuzioni delle frequenze di grandezze di interesse statistico. Per il confronto fra gli andamenti sperimentali ed i modelli teorici è stato utilizzato il software Origin Pro 8.5, ma i grafici presenti nelle successive immagini sono stati costruiti solo per motivi pratici con Microsoft Excel 2003.

6.2 Acquisizione dei dati

Come già accennato, la raccolta dati è stata effettuata prelevando i pacchetti da un nodo in cui è stato possibile concentrare il traffico dell'intera WAN. La Fig. 9 mostra uno schema sintetico dei dispositivi principali coinvolti nella connessione di un gruppo di client alla rete wireless e del metodo adottato per collegare il server di acquisizione senza alterare il funzionamento del sistema.

I client comunicano tramite connessioni punto-multipunto con gli access point (base station) della WAN, collegati agli switch che concentrano il traffico verso il centro della rete (si veda la sezione V). Il server PPPoE (*Point to Point Protocol over Ethernet*) elabora le richieste di connessione punto-punto ed invia i dati ad uno switch modello Foundry SX800, uno dei principali apparati di

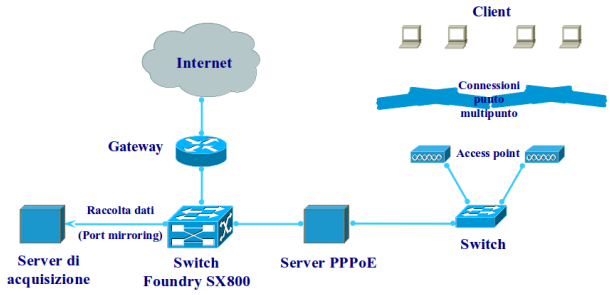


Fig. 9 Inserimento del server di acquisizione nelle connessioni della WAN.

supporto, installato nel CED.

Successivamente i pacchetti raggiungono il gateway per l'accesso ad Internet. Ovviamente il flusso è bidirezionale: il traffico può anche provenire da Internet ed arrivare ai client seguendo un percorso opposto a quello appena descritto. Il Foundry SX800 è il nodo dal quale sono stati prelevati i dati mediante una tecnica di *port mirroring* cioè di replica del traffico dell'intera WAN, che attraversa una particolare porta dello switch, su un secondo canale monitorato dal packet sniffer installato nel server di acquisizione. Occorre precisare che sono stati catturati sia i pacchetti diretti che quelli provenienti da Internet, così come quelli che si sono propagati solo all'interno della rete wireless.

Da un punto di vista software le raccolte sono state programmate tramite opportuni comandi di TCPdump inseriti nei registri di pianificazione dell'ambiente Linux (*crontab*), per evitare la necessità di un controllo diretto dell'utente sulle operazioni svolte. Durante le acquisizioni il packet sniffer ha creato in tempo reale dei file binari con estensione .pcap in cui sono state memorizzate tutte le informazioni relative ai pacchetti catturati. Precisamente, i file sono stati suddivisi in modo che in ciascuno di essi venissero riversati i dettagli del traffico monitorato in un intervallo di 3 minuti, quindi nell'arco di 24 ore sono stati generati 480 archivi. Aumentare il tempo di suddivisione avrebbe ridotto il numero di elementi, ma fatto crescere le loro dimensioni rendendo non sempre possibile in caso di necessità, l'apertura di un file .pcap in ambiente Windows attraverso un packet sniffer con interfaccia grafica come Wireshark (TCPdump è stato avviato con la riga di comando). D'altra parte diminuire eccessivamente tale tempo avrebbe avuto come conseguenza troppi file da gestire; l'intervallo di 3 minuti si è dimostrato un buon compromesso. Terminata la raccolta è stata avviata (sempre attraverso crontab) la traduzione dei dati in altrettanti documenti di testo con estensione .txt; per ogni pacchetto sono stati estratti dai file iniziali solo i tre parametri dai quali hanno avuto origine le successive elaborazioni:

1. l'istante di arrivo dall'inizio dell'acquisizione, espres-

so in secondi ed arrotondato al microsecondo (10^{-6} s);
2. le dimensioni, un valore intero espresso in byte;
3. il nome del protocollo di trasmissione.

L'ultimo dato è stato utile per valutare la natura prevalente del traffico della WAN, non per finalità statistiche particolari.

6.3 Analisi preliminare

Prima del confronto con un qualsiasi modello teorico i dati sono stati analizzati servendosi di un programma scritto in linguaggio Visual Basic 6.0 apposta per questo scopo. Come già accennato, il programma riceve in ingresso uno o più file di testo con le informazioni sui pacchetti e dopo una scansione iniziale fornisce alcuni dettagli di carattere generale relativi al traffico come:

- l'intervallo temporale di raccolta corrispondente ai file selezionati;
- il numero di pacchetti e le dimensioni totali espresse in byte o suoi multipli (kB, MB, GB);
- il throughput medio, cioè il rapporto tra durata dell'acquisizione e dimensioni complessive dei dati, espresso in bps (bit per secondo) o suoi multipli (kbps, Mbps, Gbps);
- l'intervallo di tempo minimo, medio e massimo trascorso tra l'arrivo di due pacchetti consecutivi. • Le Fig. 10 a) e b) mostrano la finestra di selezione ed un esempio di quella relativa al rapporto sul traffico.

Per quanto riguarda il throughput medio, va detto che i valori indicati si riferiscono al traffico dell'intera WAN concentrato sull'unico canale di acquisizione creato attraverso il port mirroring, quindi ogni dispositivo di rete nell'intervallo di acquisizione ha sostenuto un carico medio nettamente inferiore. Terminata la scansione che permette di passare dalla prima finestra alla seconda, si può visualizzare direttamente la distribuzione dei protocolli di trasmissione; un istogramma che riporta sull'asse orizzontale i nomi dei protocolli di scambio rilevati per i vari pacchetti, sull'asse verticale le corrispondenti frequenze. Se si vogliono costruire gli altri quattro grafici è necessario prima campionare il traffico, inserendo un tempo di campionamento nello spazio in alto a destra. Chiamando ad esempio T l'intervallo scelto, si ottiene una sequenza di campioni i cui tempi di arrivo T_k (k=1, 2, 3,...) sono: $T_1 = T, T_2 = 2T, T_3 = 3T$ ecc. Ad ogni campione vengono attribuiti:

1. il numero di pacchetti raccolti per $0 < t \leq T_1$ se k=1, per $T_{k-1} < t \leq T_k$ se k>1 (popolazione del campione);
2. le dimensioni dei pacchetti per $0 < t \leq T_1$ se k=1, per $T_{k-1} < t \leq T_k$ se k>1 (dimensione del campione).

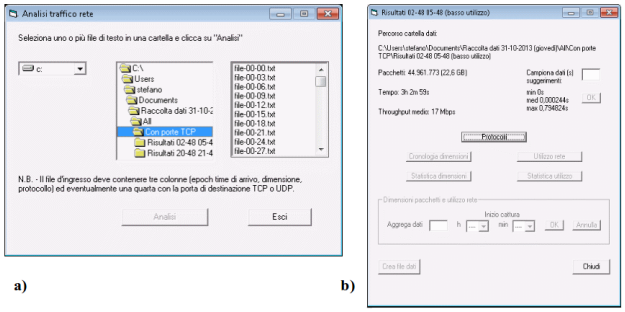


Fig. 10 Finestre del programma per l'analisi preliminare. a) Selezione del file. b) Rapporto sul traffico.

In questo modo vengono costruite le due serie temporali $X = (X_k : k = 1,2,3,\ldots)$, secondo la definizione data nella sezione 4, che saranno oggetto dell'elaborazione. E' importante notare che i valori X_k risultanti dal campionamento, per la natura dei dati e la precisione con cui opera il packet sniffer, non sono affetti da errore. Dopo questa fase si passa alla scelta del grafico.

- *Cronologia dimensioni*
  Riporta in ascisse i tempi di arrivo dei campioni, sulle ordinate le rispettive dimensioni.
- *Utilizzo rete*
  E' simile al precedente, ma sull'asse verticale viene indicata la popolazione dei campioni.
- *Statistica dimensioni*
  Riporta in ascisse le dimensioni dei campioni, in ordinate le frequenze con cui ciascuno di questi valori è stato rilevato.
- *Statistica utilizzo*
  Come la precedente è una distribuzione delle frequenze, ma sull'asse orizzontale viene indicata la popolazione dei campioni.

Prima di cliccare sul pulsante "OK" in basso a destra si deve impostare l'ordine di aggregazione "m" definito sempre nella sezione 4. Per m=1 si ricavano le serie temporali già viste, per m>1 si ottengono le serie X(m). Solo per le opzioni "Cronologia dimensioni" ed "Utilizzo rete" si può indicare come informazione facoltativa l'ora di inizio dell'acquisizione. A partire da ciascun grafico si possono salvare i relativi dati su file di testo, in modo da icostruire gli andamenti con software alternativi. Questo trasferimento è stato effettuato solo per motivi pratici adottando come programma di destinazione Microsoft Excel 2003.

In Fig. 11 viene mostrato lo schema a blocchi con le fasi di funzionamento del programma.

6.4 Elaborazione statistica dell'intero traffico

I dati di traffico raccolti si riferiscono a quattro giorni non consecutivi, due feriali e due festivi, in ciascuno dei

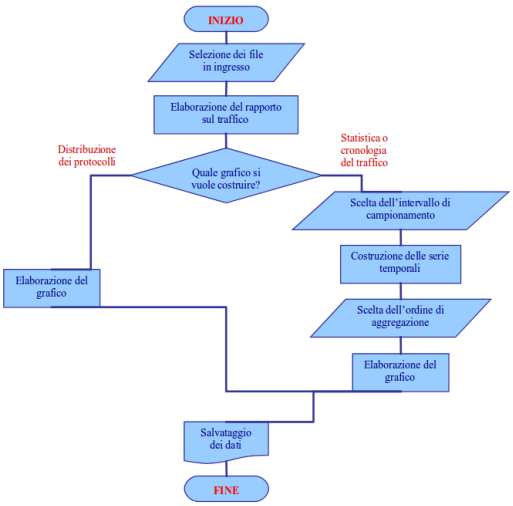


Fig. 11 Schema a blocchi del programma per l'analisi preliminare.

quali l'acquisizione è stata effettuata nell'arco di 24 ore: domenica 29-09-2013, giovedì 31-10-2013, domenica 3-11-2013, martedì 5-11-2013. La scelta delle date non è stata dettata da motivi particolari. Nel periodo trascorso tra la prima e la seconda è stato necessario risolvere alcuni problemi in merito all'ottimizzazione del programma per l'analisi preliminare ed al processo di memorizzazione in tempo reale per limitare al massimo il drop, cioè l'occasionale perdita di pacchetti da parte dell'hardware, verificatasi nei momenti di traffico intenso principalmente a causa della mancata scrittura dei dati sull'hard disk collegato al server di acquisizione. Sono state incluse due festività per verificare l'eventuale differenza di comportamento del traffico rispetto alle giornate lavorative, dovuta ad esempio alla chiusura degli uffici pubblici. Nella Tabella 1 vengono indicati i giorni di raccolta e gli intervalli di analisi.

Per ogni giorno sono stati individuati due periodi di bassa (L) ed alta (H) attività scelti in modo che al loro interno il traffico risultasse il più possibile stazionario, ovvero con valor medio approssimativamente costante e con piccole fluttuazioni intorno ad esso.

Questo per rendere ragionevole il confronto tra il modello auto-similare e quello poissoniano, cioè per soddisfare due condizioni:

- rientrare nella definizione di stazionarietà di un processo stocastico (si veda la sezione 4);
- rispettare il principio di applicabilità della distribuzione di Poisson sulla frequenza media degli eventi (si veda la sezione 3).

Gli intervalli di analisi non hanno quindi la stessa durata; per completezza sono stati indicati il numero di pacchetti e le dimensioni totali del traffico, ma il livello di attività è definito dal throughput medio.

6.5 Informazioni generali sul traffico

Per avere un'idea della tipologia di traffico che attraversa la rete, prima dell'elaborazione statistica vera e propria sono stati esaminati i protocolli di trasmissione contenuti nei pacchetti raccolti. E' stata ricavata, nell'arco di 24 ore per ciascun giorno di acquisizione, la distribuzione di tali protocolli sotto forma di istogramma.

Le Fig. 12 a), b), c), d) mostrano graficamente i risultati ottenuti. Vengono riportate esplicitamente solo le voci con le trenta maggiori frequenze, quelle restanti sono state raccolte nella colonna ALTRI sulla destra. Appare evidente come circa l'85% dei pacchetti venga trasmesso attraverso i protocolli TCP (*Transmission Control Protocol*) e UDP (*User Datagram Protocol*), che appartengono al livello di trasporto (il quarto) del modello ISO/OSI. Il primo fa da supporto a gran parte delle applicazioni della rete Internet, richiede che mittente e destinatario stabiliscano una connessione prima di avviare lo scambio dei dati e rende affidabile la trasmissione tramite funzioni di controllo degli errori, ordinamento ed eventuale reinvio dei pacchetti danneggiati o persi. Il secondo non è orientato alla connessione e sacrifica l'affidabilità a vantaggio della rapidità. E' adatto per tutte quelle applicazioni nelle quali la velocità è alla base dell'efficienza del servizio come la trasmissione di informazioni audio-video in tempo reale, dove eccessive operazioni di controllo rallenterebbero la riproduzione dei contenuti e si è disposti a tollerare occasionali imperfezioni che non pregiudicano la visione o l'ascolto.

Tra gli altri protocolli che seguono per diffusione, i principali sono:

- HTTP (*HyperText Transfer Protocol*)
  Usato come sistema per la trasmissione di informazioni sul web o in una generica architettura client-server.
- ICMP (*Internet Control Message Protocol*)
  Si occupa di inviare messaggi di controllo o riguardanti malfunzionamenti nella rete.
- PPP Comp (*Point to Point Compression Protocol*)
  Serve per la compressione dei dati nei pacchetti inviati per le trasmissioni da punto a punto.
- SSH (*Secure Shell*)
  Permette di stabilire una connessione remota cifrata tramite interfaccia a riga di comando con un altro host.
- TLS (*Transport Layer Security*)
  E' un protocollo crittografico che permette una comunicazione sicura tra sorgente e destinazione, fornendo autenticazione, integrità dei dati e loro cifratura.

Dopo questi dettagli qualitativi, il secondo passaggio è stato quello di ricostruire l'andamento giornaliero del traffico per individuare gli intervalli di analisi già elencati nella Tabella 1. Dovendo necessariamente campionare i

Tabella 1 Giorni di raccolta ed intervalli di analisi del traffico

Elenco Acquisizioni				
Periodi di raccolta ed analisi		Numero di pacchetti	Dimensioni del traffico	Throughput medio
29-09-2013 (domenica)	L: 04.15 - 05.30	17.421.336	9,9 GB	17 Mbps
	H: 15.33 - 16.24	38.200.824	18,1 GB	45 Mbps
31-10-2013 (giovedì)	L: 02.48 - 05.48	44.961.773	22,6 GB	17 Mbps
	H: 20.48 - 21.48	36.138.437	26,0 GB	55 Mbps
03-11-2013 (domenica)	L: 03.00 - 05.57	38.944.330	25,5 GB	19 Mbps
	H: 17.39 - 19.15	66.226.458	44,6 GB	60 Mbps
05-11-2013 (martedì)	L: 03.18 - 05.21	27.515.346	16,9 GB	18 Mbps
	H: 18.27 - 19.12	28.109.551	19,6 GB	55 Mbps

dati, è stata scelta la variabile casuale su cui concentrare la successiva analisi statistica, tra le due già introdotte: popolazione del campione, dimensione del campione.

Diverse prove effettuate hanno evidenziato come la prima fosse più adatta, non per motivi concettuali o legati alla bontà dei risultati ottenuti, ma perché le elaborazioni condotte dal programma per l'analisi preliminare sono state più rapide, riducendo il rischio di blocco con conseguente riavvio dell'intera procedura. Tuttavia è importante sottolineare come i caratteri peculiari del traffico sono emersi anche osservando le dimensioni dei campioni. Le Fig. 13 a), b), c), d) mostrano le ricostruzioni della cronologia dei pacchetti ottenute con la funzione Utilizzo rete.

l traffico è stato campionato con un intervallo di 3 minuti, quindi l'ordinata di ogni punto rappresenta il numero di pacchetti transitati nel nodo di raccolta in 180 secondi. Come si vedrà è un tempo elevato, adatto se si vuole ottenere una panoramica del fenomeno, poiché per le indagini statistiche approfondite sono stati scelti intervalli molto più brevi. La caratteristica comune ai quattro grafici è data dal minimo di attività raggiunto durante la notte e le prime ore del mattino, in seguito il carico cresce ed ha un andamento variabile in base agli eventi verificatisi durante il giorno. Sono stati evidenziati rispettivamente in rosso e in verde i periodi di alto (anche se non massimo) e basso utilizzo all'interno dei quali sono state condotte le analisi, che corrispondono a quelli riportati nella Tabella 1.

E' bene precisare che queste prime informazioni ottenute, così come quelle che seguiranno, riguardano tutto il traffico della WAN, compreso quello di broadcast, sebbene tale componente verrà successivamente isolata e trattata in una sezione specifica.

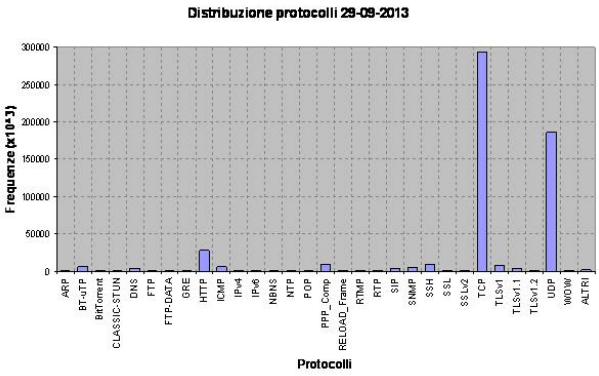
6.6 Primi indizi di autosimilarità e diagramma tempo-varianza (05-11-2013)

Nelle sezioni 3 e 4 è stato più volte sottolineato come, a differenza di quanto avviene in una rete telefonica, il flusso dati di un network informatico resta altamente variabile quando osservato su diverse scale temporali. Per avere un primo riscontro sperimentale di questa affermazione è stata esaminata una parte del traffico rilevato il giorno 05-11-2013. Sono stati inizialmente considerati 50.000 secondi (degli 86.400 corrispondenti a 24 ore) campionati con un intervallo di 50 secondi (unità di tempo) ed è stata costruita la cronologia delle dimensioni e della popolazione dei campioni così ottenuti. In seguito il procedimento è stato ripetuto altre quattro volte riducendo di un fattore 10 ad ogni passaggio sia il tempo di osservazione che quello di campionamento. Sono stati quindi riportati graficamente i seguenti andamenti:

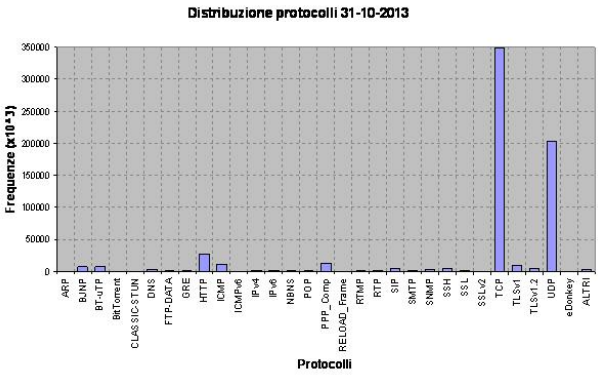
1. 50.000s di traffico con campioni di 50s
2. 5.000s di traffico con campioni di 5s
3. 500s di traffico con campioni di 0,5s
4. 50s di traffico con campioni di 0,05s
5. 5s di traffico con campioni di 0,005s

I dati considerati a partire dal punto 2 costituiscono un sottoinsieme di quelli visti nel passaggio precedente; nelle Figure 14 vengono mostrati i risultati ottenuti. I grafici nella colonna di destra sono concettualmente simili a quelli riportati nelle Figure 13, ma i valori in ascissa sono stati normalizzati rispetto all'unità di tempo. Lo stesso criterio è stato adottato nei grafici a sinistra, costruiti tramite la funzione *Cronologia dimensioni* del programma per l'analisi preliminare, dove l'ordinata di ciascun punto rappresenta la somma delle dimensioni dei pacchetti rilevati. al nodo di raccolta nell'intervallo di campionamento impostato. Partendo dall'alto, le aree colorate indicano le porzioni di traffico mostrate nelle righe successive. E' evidente come gli andamenti ottenuti non inducano ad ipotizzare un comportamento poissoniano del fenomeno che non tende a stabilizzarsi, ma conserva la propria variabilità su diverse scale dei tempi di osservazione.

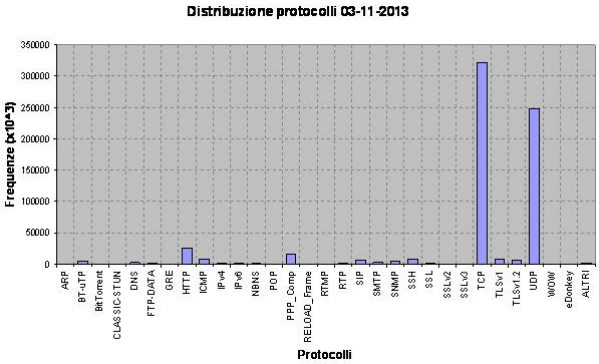
Un metodo diretto per verificare la presenza di un carattere potenzialmente auto-similare, consiste nel costruire un diagramma tempo-varianza (*variance-time plot*) come suggerito nella sezione 4. Questo procedimento è subordinato sia al campionamento dei dati che alla loro aggregazione, cioè alla costruzione delle serie temporali $X^{(m)} = (X_k^{(m)} = 1,2,3,\dots)$ di ordine m (intero positivo), in cui $X_1^{(m)}$ si ottiene mediando i valori assunti dalla variabile aleatoria scelta sul primo gruppo di m



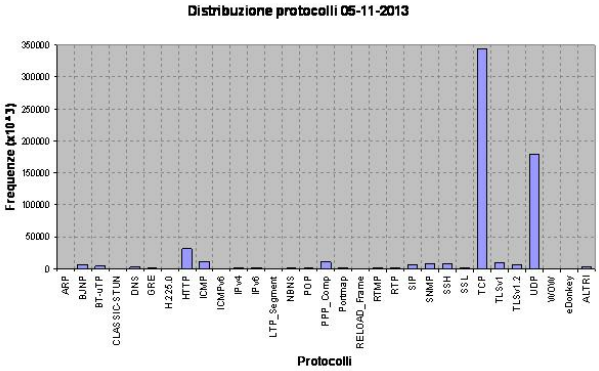
(a) Distribuzione dei protocolli di trasmissione il 29-09-2013.



(b) Distribuzione dei protocolli di trasmissione il 31-10-2013.



(c) Distribuzione dei protocolli di trasmissione il 03-11-2013.



(d) Distribuzione dei protocolli di trasmissione il 05-11-2013.

Fig. 12

campioni, $X_2^{(m)}$ è la media sul secondo gruppo e così via. Si calcolano in seguito le varianze $Var[X^{(m)}] = \sigma_m^2$ di ciascuna serie e si costruisce un grafico in cui si riportano sugli assi i seguenti rapporti:

- $\log(m)/\log(\sigma^2)$ sulle ascisse;
- $\log(\sigma_m^2)/\log(\sigma^2)$ sulle ordinate;

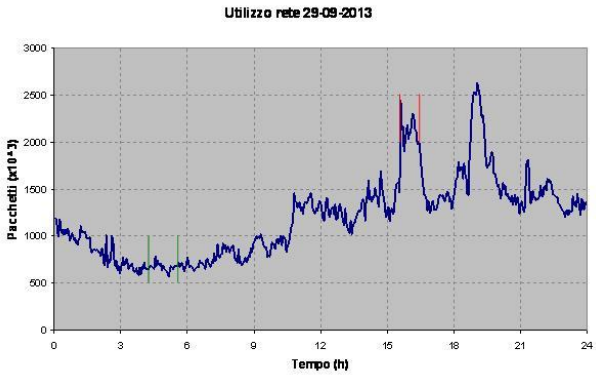
dove σ^2 è la varianza della serie originaria, con m=1. Queste quantità derivano dalla relazione 6, $\sigma_m^2 = \sigma^2 m^{-\beta}$ della sezione 4, linearizzando la quale si ottiene

$$\frac{\log(\sigma_m^2)}{\log(\sigma^2)} = 1 - \beta \frac{\log(m)}{\log(\sigma^2)} \qquad (12)$$

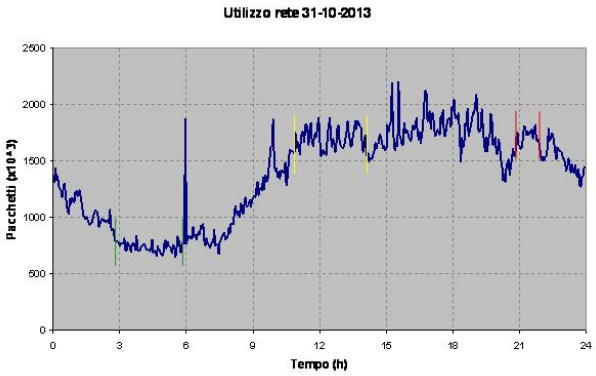
che nei rapporti indicati rappresenta una retta con pendenza pari a $-\beta$. Affinché il processo sia auto-similare deve essere $0 < \beta < 1$ cioè $-1 < -\beta < 0$; il tracciato sperimentale deve avere una pendenza esattamente o asintoticamente compresa tra -1 e 0. La Fig. 15 mostra il diagramma tempo-varianza relativo al periodo di alto utilizzo della rete del giorno 05-11-2013.

I dati sono stati inizialmente campionati con un intervallo di 5ms (verrà detto più avanti con quale criterio è stato scelto questo valore) e successivamente aggregati con i seguenti valori di m: 5, 10, 25, 50, 75, 100, 250, 500, 750, 1.000, 2.500, 5.000, 7.500. In Fig. 15 il punto di coordinate (0, 1) corrisponde all'elaborazione della serie originaria $X^{(1)}$, mentre il valore massimo di m (in questo caso 7500) è stato scelto in modo da ottenere una serie corrispondente con un numero di elementi non inferiore a 100. La "retta di riferimento" ha pendenza pari a -1; più i dati sperimentali si discostano da questo andamento verso pendenze maggiori, più il processo ha proprietà autosimilari. Nel caso riportato, dopo una fase iniziale di transizione che comprende i valori m=1 e m=5, il traffico mostra due gradi di autosimilarità rappresentati dai tratti lineari della curva: il primo intermedio (punti ♦), il secondo asintotico (punti ▲). Con un procedimento di regressione lineare sono stati calcolati tramite il programma Origin Pro 8.5 i coefficienti angolari corrispondenti, ricavando i seguenti valori di β: $\beta_1 = 0,673 \pm 0,009$ e $\beta_2 = 0,439 \pm 0,009$ (l'errore ed il numero di cifre dipendono dall'approssimazione con cui l'algoritmo esegue il fit lineare). Dato che il parametro di Hurst si esprime con la relazione $H = 1 - \beta/2$ (si veda la sezione 4) si ottiene: $H_1 = 0,663 \pm 0,005$ e $H_2 = 0,785 \pm 0,004$.

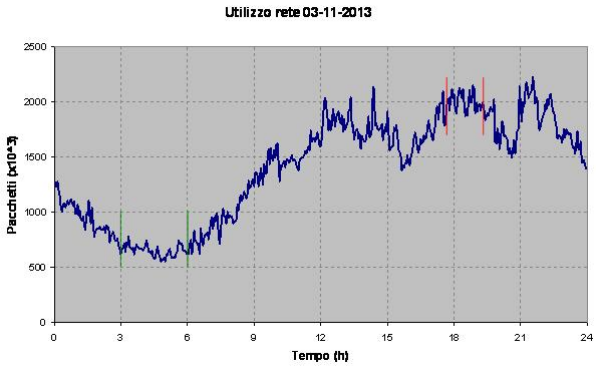
Riassumendo, nel periodo di osservazione considerato il carattere auto-similare del traffico emerge attraverso due fasi, la prima intermedia, non molto evidente, la seconda asintotica e con un valore maggiore del parametro



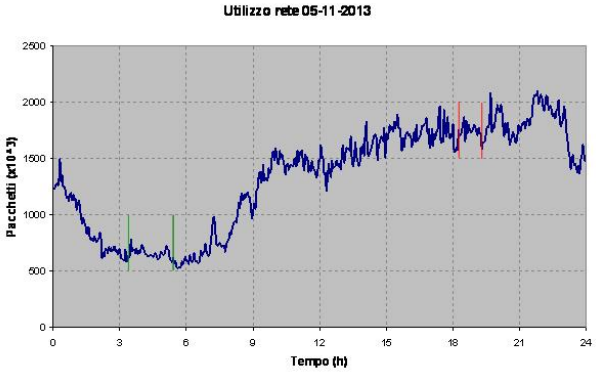
(a) Cronologia dei pacchetti 29-09-2013.



(b) Cronologia dei pacchetti il 31-10-2013.



(c) Cronologia dei pacchetti il 03-11-2013.



(d) Cronologia dei pacchetti il 05-11-2013.

Fig. 13

di Hurst.

Per completezza la Fig. 16 riporta il diagramma tempo-varianza relativo allo stesso giorno, allo stesso periodo e con i dati campionati con il medesimo intervallo, in cui però si è assunta come variabile casuale oggetto di studio la dimensione dei campioni. Il processo risulta ancora asintoticamente auto-similare, ma tale caratteristica si mostra attraverso una sola fase, i valori di β ed H ottenuti sono: $\beta = 0,309 \pm 0,005$ e $H = 0,846 \pm 0,003$.

6.7 Confronto con le distribuzioni di probabilità

E' lecito a questo punto prevedere che la statistica poissoniana non descriva efficacemente il comportamento dei dati fin qui considerati. Per verificarlo e per effettuare un confronto fra modelli è stata utilizzata come strumento di partenza la funzione *Statistica utilizzo* del programma per l'analisi preliminare. Come già accennato permette di ricavare la distribuzione delle frequenze dei valori assunti dalla variabile "popolazione dei campioni", per un intervallo di campionamento ed un ordine di aggregazione preventivamente impostati. I relativi dati sono stati salvati in file di testo, importati in una cartella di lavoro delsoftware Origin Pro 8.5 e normalizzati in modo da ottenere le corrispondenti densità di probabilità sperimentali. Con lo stesso software sono stati effettuati alcuni fit non lineari con distribuzioni notevoli e ricavati i valori di parametri statistici che descrivono l'efficacia di un determinato modello o permettono di valutare un confronto tra modelli. A tal proposito sono stati presi rispettivamente come riferimento il *Coefficiente di determinazione modificato* (*Adjusted coefficient of determination* o *Adjusted R square*) e l'*Akaike Information Criterion* (AIC). Il primo è costituito da un numero reale inferiore o uguale a 1 (può essere anche negativo); più è prossimo all'unità più il modello considerato descrive efficacemente la realtà sperimentale. Il secondo può assumere qualsiasi valore reale e viene calcolato per tutti i modelli messi a confronto, quello che offre l'AIC minore (più alto in modulo se negativo) si candida a rappresentare meglio i dati ottenuti. Il procedimento per il calcolo di questi coefficienti è riportato in Appendice A.

Riprendendo quanto esposto nella sezione 3, per il counting process di una serie temporale i cui valori sono descritti dalla distribuzione di Poisson vale la relazione

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

dove $P\{n\}$ è la probabilità di ottenere n successi nell'intervallo t, ovvero di contare n pacchetti nel tempo di campionamento t, essendo λ la frequenza media dei successi. Proseguendo l'analisi del traffico del 05-11-2013 in condizioni di alto utilizzo, la Fig. 17 mostra il confronto tra la densità di probabilità sperimentale e la fun-

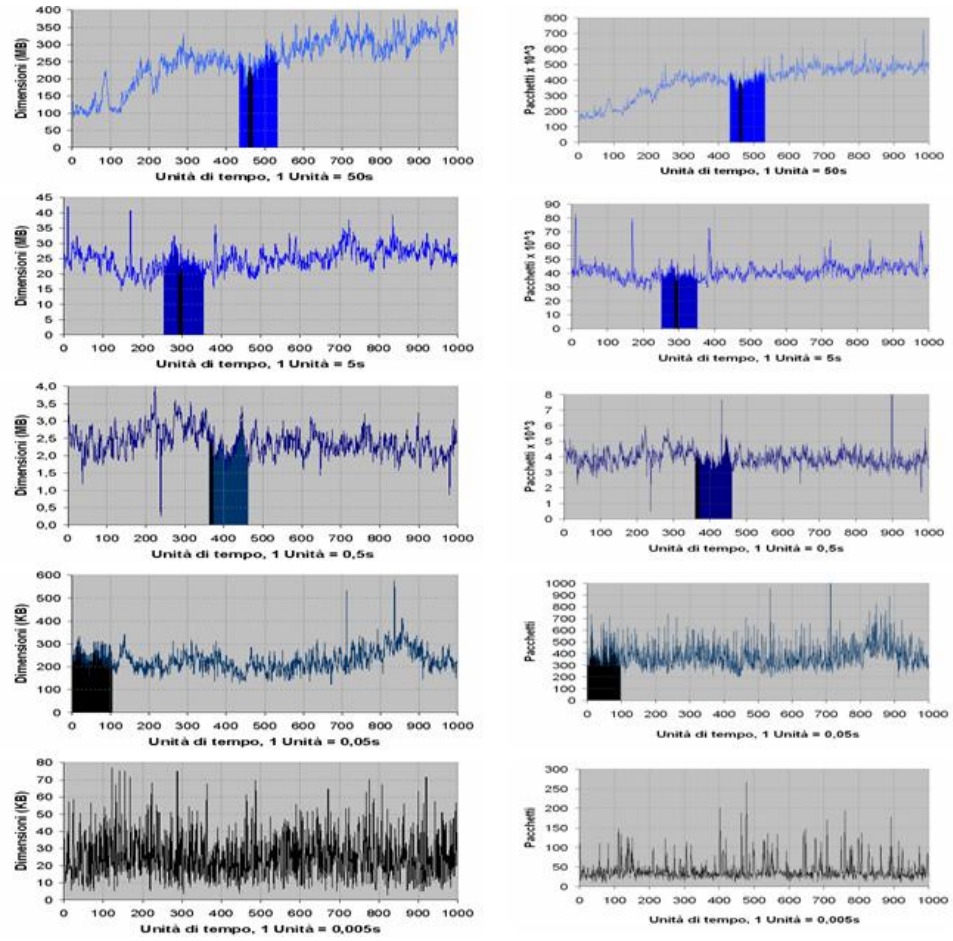


Fig. 14 Alta variabilità del traffico osservata su cinque scale temporali.

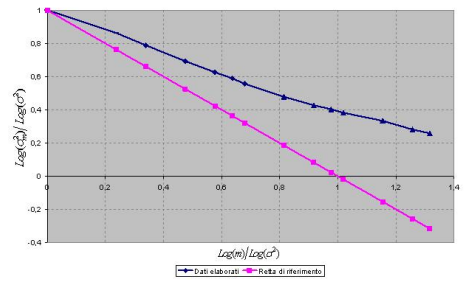


Fig. 15 Diagramma tempo-varianza del 05-11-2013 (alto utilizzo), costruito con la popolazione dei campioni.

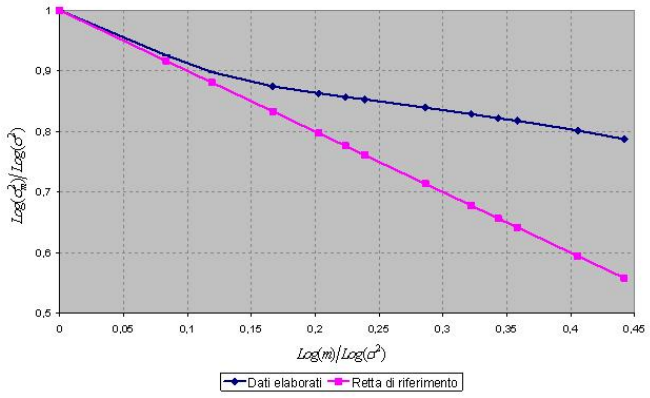


Fig. 16 Diagramma tempo-varianza del 05-11-2013 (alto utilizzo), costruito con la dimensione dei campioni.

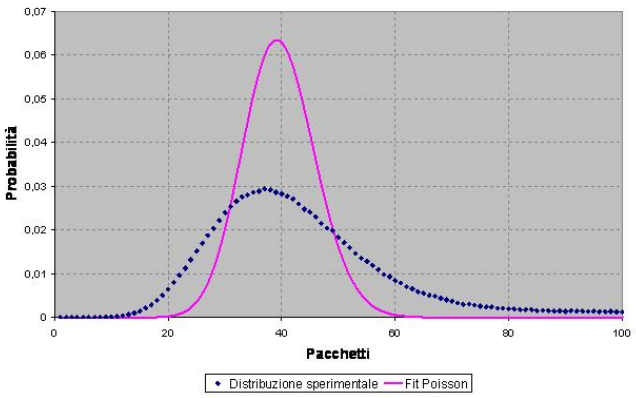


Fig. 17 Fit poissoniano dei dati del 05-11-2013 (alto utilizzo) campionati a 5ms

zione poissoniana che meglio vi si adatta al variare del parametro λ.

Per rendere l'immagine comprensibile è stata riportata solo una parte dell'asse orizzontale, inoltre la distribuzione di Poisson è stata rappresentata con una curva continua pur trattandosi di un insieme discreto di valori. Il tempo di campionamento t pari a 5ms, introdotto nel paragrafo precedente, rende visibile sia il ramo crescente che quello decrescente della densità sperimentale ed assicura in questo modo di non osservare il fenomeno

con un dettaglio temporale troppo elevato per le proprietà statistiche che si desidera esaminare. Se fosse stato scelto, ad esempio, un intervallo di 1ms sarebbe stata visibile solo la coda della distribuzione. Con il fit realizzato è stato determinato un valore del parametro λ tale che $\lambda t = 39,6 \pm 0,5$ (numero medio dei pacchetti contati nel tempo t) ed un coefficiente di determinazione modificato $Adj.R^2 = -0,2419$, quindi sia i valori numerici che la Fig. 17 confermano le previsioni sull'inadeguatezza del modello di Poisson.

All'inizio di questa sezione sono state introdotte altre due funzioni per il confronto con la statistica autosimilare: la distribuzione log-normale e la distribuzione di Weibull. La prima ha la seguente forma analitica:

$$y = \frac{A}{\sqrt{2\pi}x\sigma} \exp\left[\frac{-\left(\ln\frac{x}{x_c}\right)^2}{2\sigma^2}\right] \qquad (13)$$

Se una variabile casuale x ha una distribuzione log-normale, il suo logaritmo ln(x) è descritto dalla statistica di Gauss. Nella (13) σ rappresenta la deviazione standard di $\ln(x)$ mentre $x_c = e^{\mu}$ dove μ è il valor medio di $\ln(x)$, infine A è l'area sottesa al grafico della funzione (i valori che si otterranno saranno prossimi a 1 avendo normalizzato i dati). I. Antoniou, V.V. Ivanov, Valery V. Ivanov e P.V. Zrelov[5] suggeriscono di confrontare il traffico di una rete geografica anche con la distribuzione di Weibull che si presenta nel seguente modo

$$y = \frac{b}{a}\left(\frac{x-c}{a}\right)^{b-1} \exp\left[-\left(\frac{x-c}{a}\right)^{b}\right] \qquad (14)$$

dove a, b, c sono parametri che determinano l'ampiezza, l'altezza e la posizione del massimo della curva. Le Fig. 18 mostrano graficamente i risultati dei fit realizzati con le densità di probabilità log-normale (a), b)) e di Weibull (c), d)), sempre sui dati di alto utilizzo del 05-11-2013, per valori dell'ordine di aggregazione m=100 e m=1000 che cadono rispettivamente nel primo e nel secondo intervallo di auto-similarità messi in evidenza nella Fig. 16.

La Tabella 2 riporta i valori calcolati dei parametri delle funzioni con i relativi errori, i coefficienti di determinazione modificati (a), b)) e i risultati del confronto (AIC) fra i due modelli (c)), dall'ordine m=10 a partire dal quale il processo diventa auto-similare (viene indicato con un bordo tratteggiato il passaggio tra i due gradi di autosimilarità).

In base a quanto detto a proposito dei parametri statistici $Adj.R2$ e AIC, è evidente come la distribuzione log-normale sia più adatta a rappresentare le caratteristiche del traffico analizzato. Inoltre la funzione di Weibull ha dei coefficienti che variano in modo alquanto irregolare all'aumentare di m, ciò non induce a sceglierla per descrivere dati che dovrebbero conservare le loro proprietà analitiche su diverse scale temporali.

6.8 Analisi in condizioni di basso utilizzo del giorno 05-11-2013

Viene ora completata l'esposizione dei risultati ottenuti dall'analisi dei dati raccolti il 05-11-2013, relativamente al periodo di bassa attività della rete corrispondente all'intervallo compreso fra le 03.18 e le 05.21 del mattino (si vedano la Tabella 1 e la Fig. 13d)). Il tempo di campionamento è sempre pari a 5ms, la variabile casuale è ancora rappresentata dalla popolazione dei campioni; tali parametri dovranno intendersi adottati anche nelle considerazioni successive, salvo ove diversamente specificato. La Fig. 19 mostra il confronto con la distribuzione di Poisson. Il fit non lineare ha fornito un valore di λ tale che $\lambda t = 11,8 \pm 0,2$ con un coefficiente di determinazione modificato $Adj.R^2 = 0,5798$, migliore di uello ottenuto in precedenza ma non ancora sufficiente per ritenere valida la descrizione del traffico attraverso questo modello, come si può dedurre anche graficamente. Il passo successivo consiste nel verificare l'eventuale carattere auto-similare dei dati; la Fig. 20 riporta il diagramma tempo-varianza con valori di m compresi fra 1 e 10.000.

Anche in questo caso si possono individuare due gradi di autosimilarità dopo una fase iniziale transitoria. I valori del parametro di Hurst ottenuti sono rispettivamente $H_1 = 0,651 \pm 0,002$ e $H_2 = 0,779 \pm 0,014$, prossimi a quelli ricavati in condizioni di alto utilizzo.

Per il confronto tra le distribuzioni log-normale e di Weibull, la Fig. 21 mostra i fit non lineari realizzati con ordini di aggregazione pari a 100, 1.000 e 10.000. Il passaggio dal primo al secondo grado di autosimilarità si ha per valori di m compresi fra 500 e 750. Analogamente a quanto visto nel paragrafo precedente, la Tabella 3 elenca i parametri analitici delle funzioni e quelli statistici per la valutazione dei modelli; vengono riportati solo gli ordini di aggregazione corrispondenti alle potenze di 10.

Anche ora i risultati numerici e grafici suggeriscono che la distribuzione log-normale si presta meglio ad interpretare teoricamente le proprietà del traffico. Se si considerano anche i dati raccolti nel periodo di intensa attività, i valori del parametro di Hurst ottenuti non sembrano dipendere dal carico a cui la rete è sottoposta.

6.9 Risultati ottenuti dalle altre acquisizioni

Dopo aver illustrato dettagliatamente il procedimento con il quale sono stati elaborati i dati raccolti martedì 05-11-2013, vengono ora presentati i risultati forniti dal traffico raccolto durante gli altri tre giorni ovvero: domenica 29-09-2013, giovedì 31-10-2013, domenica 3-11-2013. Ovviamente i metodi adottati per ricavare le informazioni statistiche di rilievo sono quelli già introdotti e sono

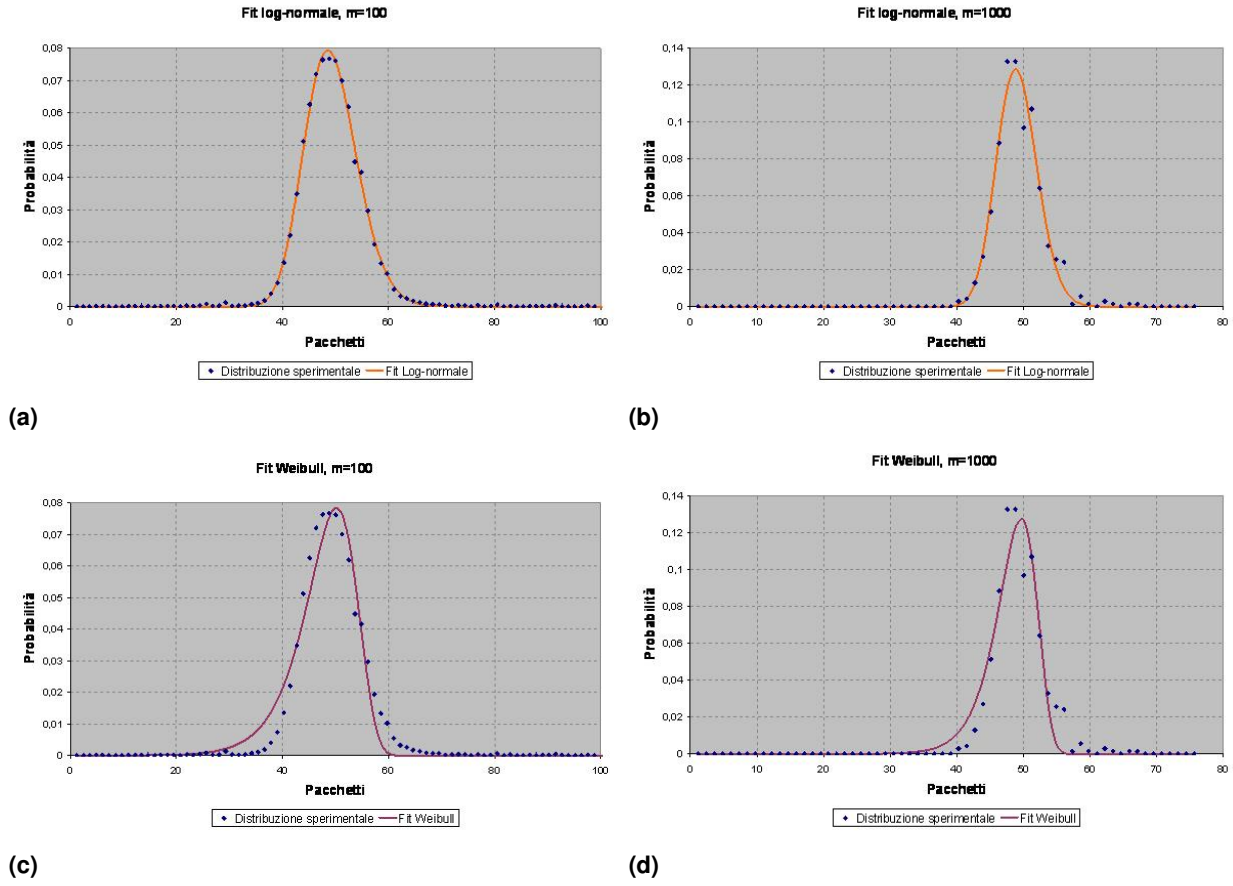


Fig. 18 Fit log-normale e di Weibull dei dati del 05-11-2013 (alto utilizzo), per m=100 e m=1000

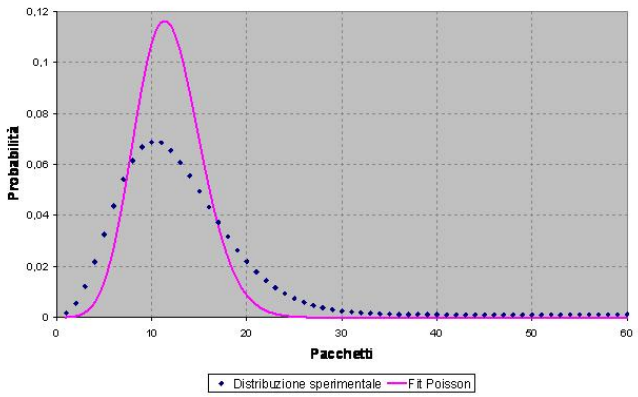


Fig. 19 Fit poissoniano dei dati del 05-11-2013 (basso utilizzo) campionati a 5ms.

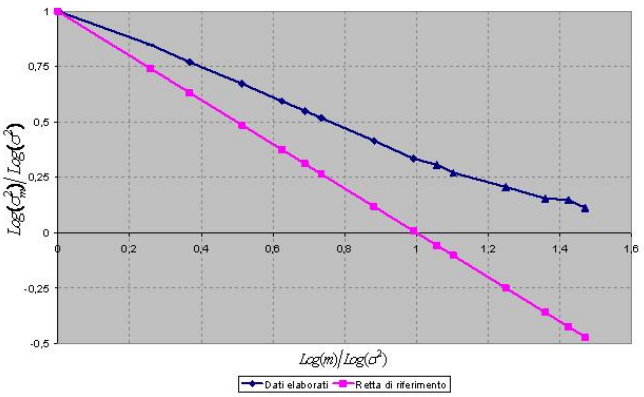


Fig. 20 Diagramma tempo-varianza del 05-11-2013 (basso utilizzo).

stati applicati ai dati acquisiti nei periodi di alta e bassa attività, i passaggi seguiti consistono quindi nel:

1. confrontare le densità di probabilità sperimentali con la distribuzione di Poisson
2. costruire i diagrammi tempo-varianza
3. riportare i valori dei parametri statistici relativi alle distribuzioni log-normale e di Weibull.

In merito all'ultimo punto, non verranno elencate le proprietà analitiche delle singole funzioni (x_c , σ, A, a, b, c), ma solo i valori dei coefficienti $Adj.R^2$ e AIC. Inoltre per motivi di sintesi, i fit non lineari rappresentati graficamente si riferiranno solo alla distribuzione di Poisson; si veda la Fig. 22 relativa a tutti e tre i giorni.

Come atteso il modello non sembra adattarsi ai dati elaborati, ma le discrepanze sono meno evidenti nei periodi di bassa attività. In effetti il traffico si presenta più stazionario in tali intervalli, rispetto ai periodi di alto utilizzo (si vedano anche le Fig. 13 a), b), c), d)) ed è probabilmente questa la causa del fenomeno, peraltro già rilevato con i dati del 05-11-2013. In Tabella 4 i valori dei coefficienti $Adj.R^2$ in alta (H) e bassa (L) intensità del

Tabella 2 Valori numerici dei fit per il giorno 05-11-2013 (alto utilizzo) e relativo confronto fra modelli

a) Risultati Fit Log-Normale				
m	x_c	σ	A	$Adj.R^2$
10	$45,81 \pm 0,09$	$0,203 \pm 0,002$	$0,958 \pm 0,008$	0,9710
100	$49,06 \pm 0,02$	$0,101 \pm 0,001$	$0,986 \pm 0,004$	0,9983
1.000	$49,00 \pm 0,07$	$0,062 \pm 0,001$	$0,976 \pm 0,020$	0,9798
7.500	$49,01 \pm 0,06$	$0,042 \pm 0,001$	$1,000 \pm 0,025$	0,9679

b) Risultati Fit Weibull				
m	x_c	σ	A	$Adj.R^2$
10	47 ± 6	$5,0 \pm 0,9$	$1,99 \pm 0,06$	0,9082
100	258 ± 4	55 ± 2	-208 ± 4	0,9643
1.000	122 ± 3	42 ± 4	-72 ± 3	0,9387
7.500	170 ± 1	93 ± 4	-121 ± 1	0,9584

c) Confronto fra i modelli		
m	AIC log-normale	AIC Weibull
10	-7843	-7173
100	-2216	-1750
1.000	-662	-593
7.500	-662	-6444

Tabella 3 Valori numerici dei fit per il giorno 05-11-2013 (basso utilizzo) e relativo confronto fra modelli.

a) Risultati Fit Log-Normale				
m	x_c	σ	A	$Adj.R^2$
10	$13,22 \pm 0,03$	$0,231 \pm 0,003$	$0,743 \pm 0,007$	0,9045
100	$17,98 \pm 0,01$	$0,180 \pm 0,001$	$0,995 \pm 0,002$	0,9986
1.000	$18,17 \pm 0,02$	$0,088 \pm 0,001$	$0,981 \pm 0,008$	0,9943
7.500	$18,25 \pm 0,04$	$0,062 \pm 0,002$	$0,985 \pm 0,028$	0,9655

b) Risultati Fit Weibull				
m	x_c	σ	A	$Adj.R^2$
10	14 ± 1	$2,9 \pm 0,1$	$1,49 \pm 0,01$	0,77822
100	50 ± 3	16 ± 1	-31 ± 3	0,9598
1.000	36 ± 1	24 ± 2	-18 ± 1	0,9741
7.500	145 ± 1	136 ± 8	-126 ± 1	0,9198

c) Confronto fra i modelli		
m	AIC log-normale	AIC Weibull
10	-20327	-18884
100	-6346	-4879
1.000	-1150	-991
7.500	-486	-419

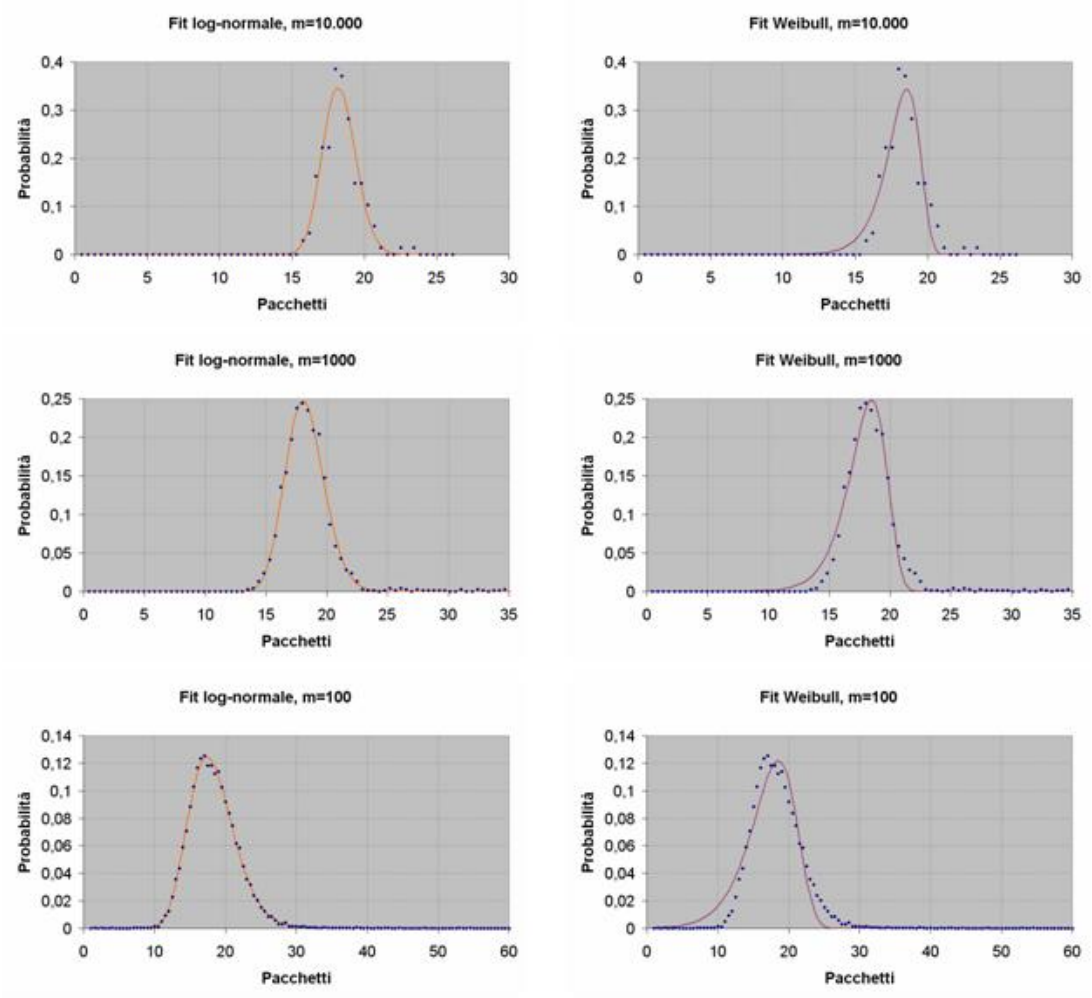


Fig. 21 Fit log-normale e di Weibull dei dati del 05-11-2013 (basso utilizzo), per m pari a 100, 1.000, 10.000.

Tabella 4 Coefficienti $Adj.R^2$ dei fit poissoniani

Risultati fit poisson		
Giorni	$Adj.R^2(H)$	$Adj.R^2(L)$
29/09/2013	-1,7288	0,7371
31/10/2013	-0,1295	0,6966
03/11/2013	-0,5466	0,6779

flusso confermano i risultati grafici. Le proprietà autosimilari del traffico possono essere dedotte dai diagrammi tempo-varianza mostrati in Fig. 23.

Nel periodo di alto utilizzo del 29-09-2013 (immagine in alto a sinistra) non è stato possibile determinare i valori corrispondenti di β e H perché il traffico, pur tendendo all'autosimilarità, non stabilizza le sue proprietà entro gli ordini con i quali è stato aggregato ($1 \leq m \leq 5.000$). Ciò è dovuto probabilmente al debole carattere stazionario del flusso rispetto all'ampiezza dell'intervallo considerato, come si può dedurre anche dalla Fig. 13 a). Nello stesso giorno ma in condizioni di bassa attività (immagine in alto a destra), la disposizione dei punti individua un solo grado di autosimilarità; il grafico cambia pendenza in due occasioni tornando però a seguire quella d'origine. Ciò non avviene nelle altre quattro figure, benché non sia sempre evidente salvo che per i simboli adottati, nelle quali il comportamento dei dati è simile a quello riscontrato il 05-11-2013. La Tabella 5 riporta i risultati numerici dei fit realizzati e i valori ottenuti del parametro di Hurst.

Come già dedotto graficamente, nei periodi di alto e basso utilizzo del 03-11-2013 le differenze fra le rispettive coppie di valori di H sono piccole. I risultati di tutte le tabelle confermano quanto ottenuto con i dati del 05-11-2013:

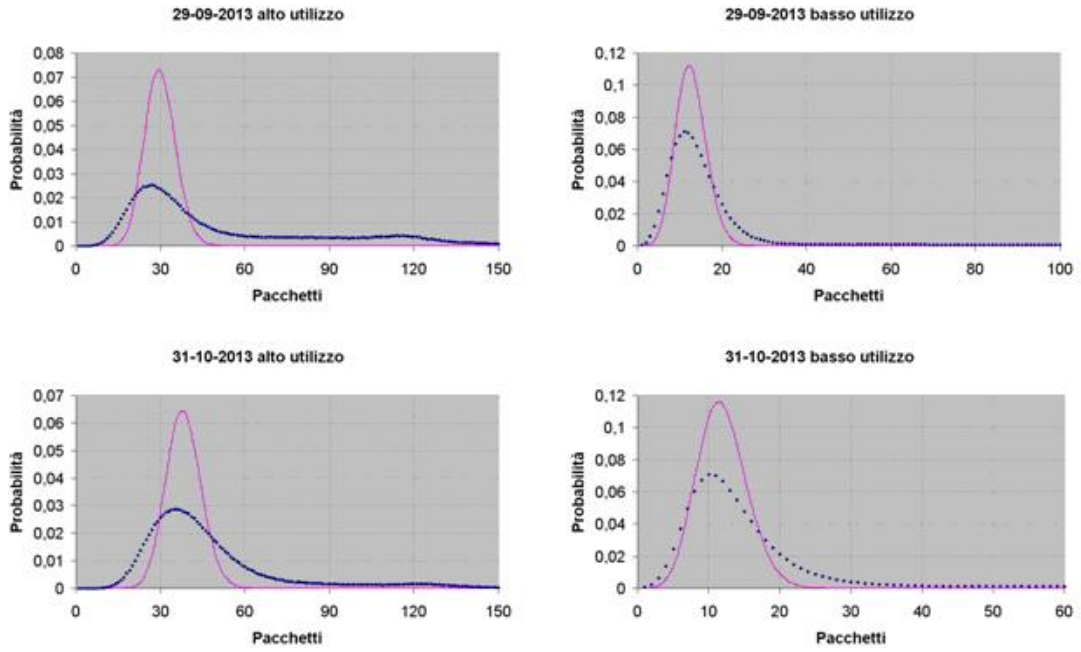
- la statistica di Poisson non è adatta a descrivere il comportamento dei dati sperimentali;
- l'intero traffico di questa rete geografica presenta proprietà asintoticamente autosimilari che, da un confronto fra la distribuzione log-normale e quella di Weibull, risultano ben interpretabili attraverso la prima.

6.10 Elaborazione statistica del traffico di broadcast

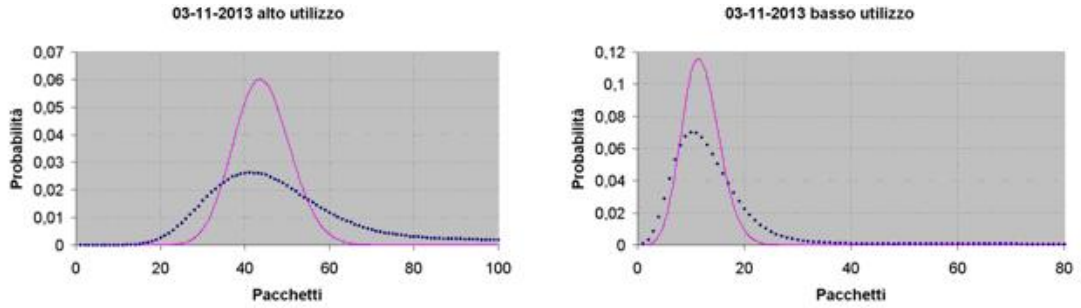
Contestualmente alla traduzione degli archivi binari .pcap generati dal packet sniffer in documenti di testo relativi all'intero traffico della rete, attraverso opportu-

Tabella 5 Valori dei parametri statistici relativi al traffico dei tre giorni indicati.

29-09-2013 BASSO UTILIZZO ($H = 0,682 \pm 0,009$)				
m	Log-normal		Weibull	
	$Adj.R^2$	AIC	$Adj.R^2$	AIC
10	0,8997	-17850	0,7719	-16603
100	0,9994	-6120	0,9556	-4436
1000	0,9930	-872	0,9846	-811
10000	0,9809	-331	0,9710	-315
31-10-2013 ALTO UTILIZZO ($H_1 = 0,665 \pm 0,005$ $H_2 = 0,829 \pm 0,005$)				
m	Log-normal		Weibull	
	$Adj.R^2$	AIC	$Adj.R^2$	AIC
10	0,9862	-9185	0,9349	-8206
100	0,9988	-4626	0,9787	-3763
1000	0,9958	-738	0,9748	-631
31-10-2013 BASSO UTILIZZO ($H_1 = 0,670 \pm 0,004$ $H_2 = 0,757 \pm 0,009$)				
m	Log-normal		Weibull	
	$Adj.R^2$	AIC	$Adj.R^2$	AIC
10	0,87582	-12356	0,7919	-11810
100	0,9990	-7473	0,9814	-6048
1000	0,9983	-1339	0,9719	-1047
10000	0,9780	-330	0,9087	-278
03-11-2013 ALTO UTILIZZO ($H_1 = 0,719 \pm 0,002$ $H_2 = 0,755 \pm 0,003$)				
m	Log-normal		Weibull	
	$Adj.R^2$	AIC	$Adj.R^2$	AIC
10	0,9914	-8520	0,9461	-7493
100	0,9990	-5526	0,9699	-4346
1000	0,9957	-1115	0,9488	-901
10000	0,9894	-348	0,9783	-325
03-11-2013 BASSO UTILIZZO ($H_1 = 0,659 \pm 0,005$ $H_2 = 0,691 \pm 0,0049$)				
m	Log-normal		Weibull	
	$Adj.R^2$	AIC	$Adj.R^2$	AIC
10	0,9340	-24075	0,8267	-22205
100	0,9994	-8230	0,9814	-6447
1000	0,9942	-942	0,9755	-823
10000	0,9908	-271	0,9625	-231



(a) Confronto tra le distribuzioni sperimentali e quella di Poisson nei 3 giorni di raccolta indicati



(b) Confronto tra le distribuzioni sperimentali e quella di Poisson nei 3 giorni di raccolta indicati

Fig. 22

ni comandi di TCPdump sono stati creati altrettanti file .txt contenenti informazioni sui soli pacchetti di broadcast. Notizie sulla definizione, sui meccanismi di propagazione e sull'importanza del monitoraggio di questa componente del flusso sono state date nella sezione 5 e nell'introduzione alla presente.

Verrà ora illustrata l'elaborazione statistica condotta su questo sottoinsieme adottando gli stessi metodi e lo stesso procedimento seguito per l'analisi dell'intero traffico. Come primo passo sono quindi stati creati 480 file di testo per ogni giorno di acquisizione ed in ciascun documento sono stati memorizzati i dati relativi ad un intervallo di 3 minuti. Attraverso il programma per l'analisi preliminare è stata esaminata la natura del traffico costruendo gli istogrammi di distribuzione dei protocolli di trasmissione, tramite un opportuno tempo di campionamento è stata ricavata la cronologia del flusso assumendo ancora come variabile aleatoria notevole la popolazione dei campioni. A questo punto i due periodi di alta e bassa attività non sono stati isolati poiché, come si vedrà, il traffico di broadcast costituisce meno di qualche unità per mille del flusso totale, quindi il metodo ottimale per ottenere una popolazione statisticamente significativa è stato quello di condurre l'analisi su un intero giorno di raccolta. I risultati sono stati ottenuti attraverso fit non lineari e diagrammi tempo-varianza secondo quanto già collaudato.

La Tabella 6 riporta i dati riassuntivi dei quattro giorni di acquisizione.

Confrontando questi valori con quelli elencati nella Tabella 1 ci si rende conto dell'esiguità di questa componente, soprattutto se si considera che i presenti si riferiscono ad intervalli di 24 ore, i precedenti a periodi non più ampi di 3 ore.

Allo stato attuale, osservando in particolare il throughput medio, si può quindi affermare che l'entità del traffico di broadcast non rappresenta un fattore di degrado delle prestazioni della rete.

6.11 Informazioni generali sul traffico

In modo analogo a quanto già visto con l'esame dell'intero traffico, le Figure 24a-d mostrano le distribuzioni dei

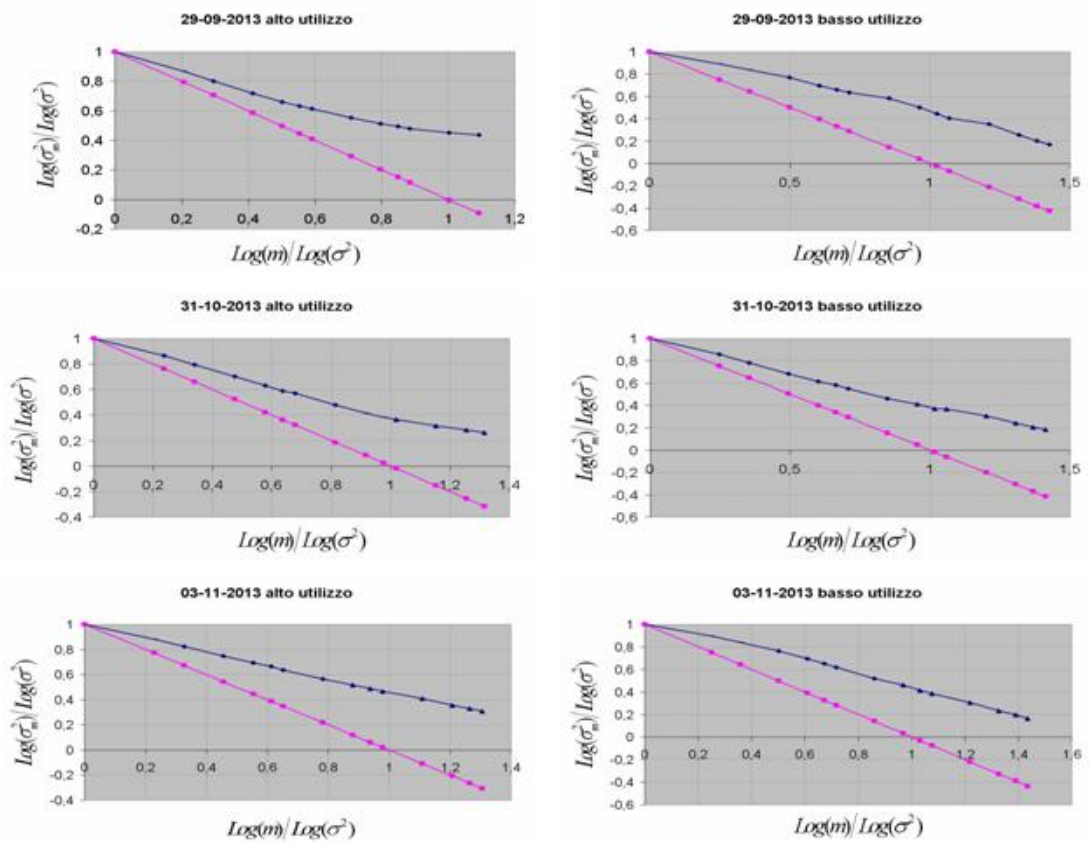


Fig. 23 Diagrammi tempo-varianza relativi ai tre giorni di raccolta indicati.

Tabella 6 Dati riassuntivi del traffico di broadcast.

ELENCO ACQUISIZIONI TRAFFICO DI BROADCAST			
Giorni di raccolta	Numero di pacchetti	Dimensioni del traffico	Throughput medio
29-09-2013 (domenica)	308.017	23,7 MB	2 kbps
31-10-2013 (giovedì)	347.856	26,2 MB	2 kbps
03-11-2013 (domenica)	323.345	24,4 MB	2 kbps
05-11-2013 (martedì)	433.015	32,7 MB	3 kbps

protocolli di trasmissione nei quattro giorni di raccolta.

E' evidente come il protocollo ARP (*Address Resolution Protocol*) sia quello più diffuso. Appartiene al livello datalink (il secondo) del modello ISO/OSI e viene utilizzato per determinare l'indirizzo fisico (MAC address) di un host della rete a partire dal corrispondente indirizzo IP. Una macchina che vuole conoscere il MAC address di un'altra con indirizzo IP noto, invia in broadcast una richiesta ARP che viene quindi ricevuta da tutti i nodi della rete. Ciascun host confronta l'indirizzo IP di tale richiesta con il proprio, solo quello che trova una corrispondenza risponde al mittente comunicando il MAC address cercato. Dato che i dispositivi della WAN sono impostati per operare al livello due (si veda la sezione 5), non stupisce che i pacchetti di broadcast di tipo ARP siano i più frequenti. Il secondo in ordine di diffusione è il protocollo iSNS (*Internet Storage Name Service*) associato a servizi di individuazione, gestione e configurazione di dispositivi di memorizzazione, mentre il terzo è DB-LSP-DISC (*Drop Box LAN Sync Discovery*) che il software di archiviazione Dropbox utilizza per mantenere la sincronizzazione tra i dati depositati su server remoti e i computer di una rete.

La cronologia dei pacchetti di broadcast è rappresentata nelle Fig. 25 a), b), c), d) dove, analogamente a quanto visto in precedenza, il traffico è stato campionato con un intervallo di 3 minuti.Confrontando queste immagini con le Fig. 13 si nota come il comportamento della componente di broadcast sia sensibilmente diverso dall'andamento dell'intero traffico. Le figure a) e c) si riferiscono a giorni festivi, mentre in b) e d) sono rappresentati giorni feriali. In quest'ultimo caso il flusso si intensifica quasi in corrispondenza delle ore lavorative e dato che l'apertura

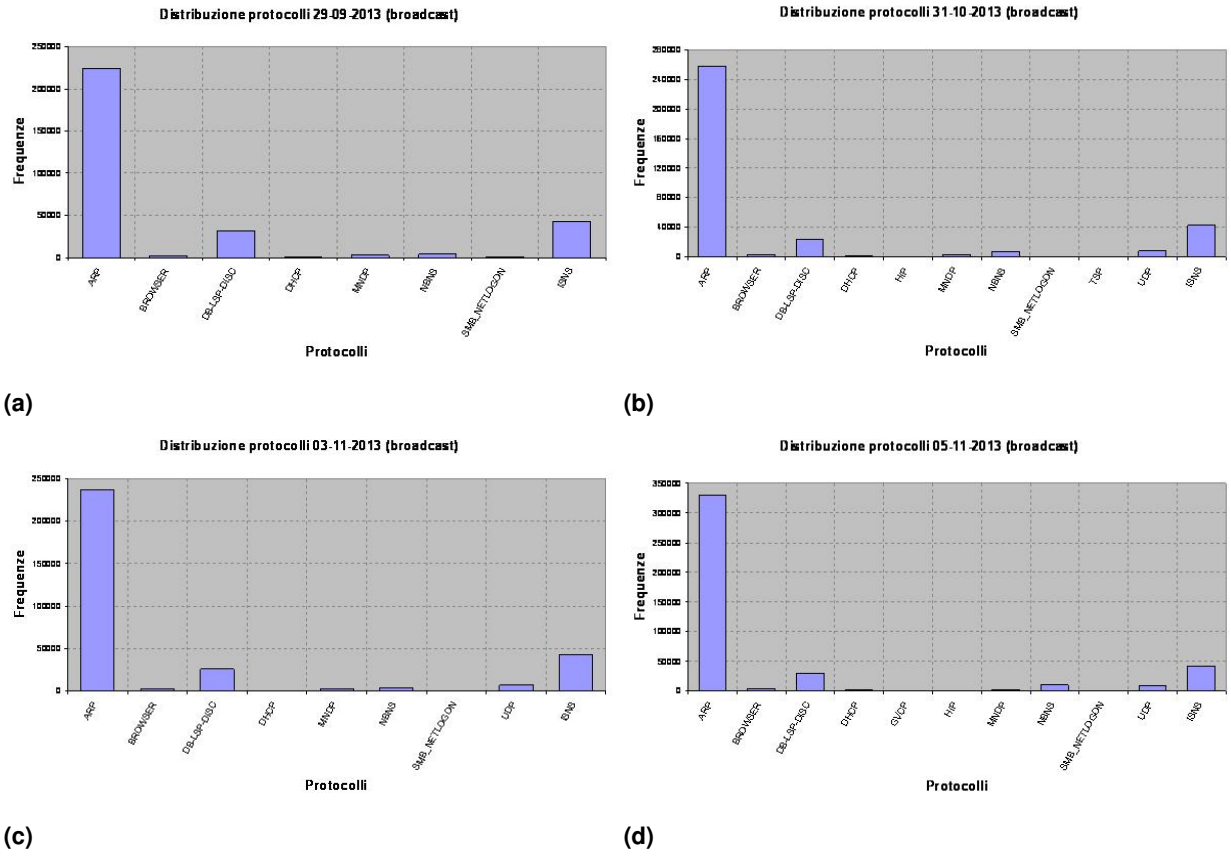


Fig. 24 Distribuzioni dei protocolli nel traffico di broadcast

degli uffici pubblici costituisce la sostanziale differenza rispetto alle attività svolte durante le due domeniche, è ragionevole supporre che sia questa la causa del maggiore utilizzo. Gli eventuali picchi del traffico possono essere dovuti all'invio di un numero consistente di pacchetti da parte degli apparati della WAN, con lo scopo di sincronizzarsi e controllare la reciproca presenza in rete che può venir meno a causa di malfunzionamenti. Per verificare le precedenti ipotesi sarebbe comunque necessario un esame più approfondito che andrebbe oltre lo scopo di questa analisi. Un altro particolare evidente, prevedibile dai dati della Tabella 6, è la differenza tra i valori riportati sulle ordinate di questi grafici e quelli presenti nella Fig. 13, che conferma la netta inferiorità della componente di broadcast rispetto al traffico complessivo.

Per le finalità dell'indagine statistica è interessante infine osservare la cronologia dei giorni festivi (a), c)) e delle prime ore del mattino di quelli feriali (b), d)). Il flusso campionato con un intervallo di 180 secondi tende a regolarizzarsi, ad assumere la tipica forma di un rumore di fondo e questo comportamento, come introdotto nella sezione 4, è tipico dei processi poissoniani. Prove effettuate considerando le prime 6-8 ore dei giorni lavorativi hanno condotto a popolazioni statistiche non significative, quindi l'analisi è stata realizzata su un periodo di 24 ore dei soli giorni festivi (29-09-2013 e 03-11-2013) anche per rispettare la condizione di stazionarietà del processo.

6.12 Diagrammi tempo-varianza e fit non lineari

Dato il ragionevole sospetto di dover esaminare un traffico descritto dalla statistica di Poisson, in questa fase sono stati prima costruiti i diagrammi tempo-varianza per il consueto test di autosimilarità, in modo da scartare subito tale ipotesi qualora non fosse verificata. Le Fig. 26 a) e b) mostrano i risultati ottenuti.

E' stato più volte precisato che la componente di broadcast costituisce una piccola frazione del traffico complessivo, ciò ha reso necessario l'utilizzo di un intervallo di campionamento maggiore di quello adottato nella precedente analisi, ma scelto con criteri simili.

I dati appena rappresentati sono stati campionati con un tempo pari a 0,5 secondi, che ha generato un insieme di 172.800 elementi (2 x 86.400 secondi in 24 ore). E' chiaro come il processo non possa essere considerato auto-similare dato che il tracciato elaborato tende a seguire in entrambi i casi la retta di riferimento, è stata quindi considerata priva di senso la realizzazione di fit non lineari con le funzioni log-normale e di Weibull. Questo controllo è stato dunque eseguito con la distribuzione di Poisson e le Fig. 27 a), b) mostrano l'aderenza del modello.

L'intervallo di campionamento già menzionato è stato scelto in modo che fosse appena visibile il massimo della

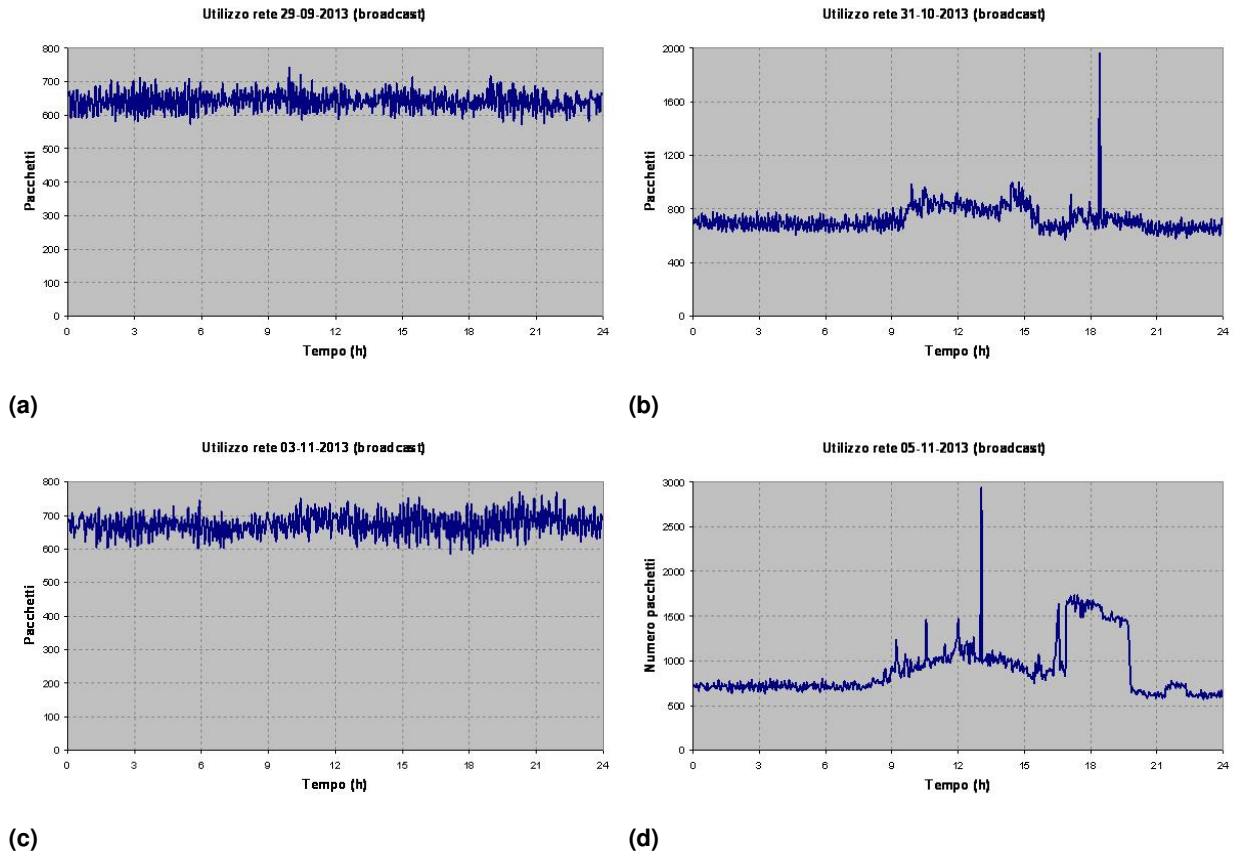


Fig. 25 Cronologia dei pacchetti di broadcast

curva, non l'intero ramo crescente come stabilito nell'analisi dell'intero traffico. Questo criterio è stato adottato per aumentare il numero dei campioni generati e di conseguenza il loro significato statistico.

Da un punto di vista grafico il modello di Poisson appare un buon descrittore dei dati sperimentali, impressione confermata secondo le previsioni dai coefficienti di determinazione modificati di seguito riportati:

$$\text{Fit Poisson del 29-09-2013} \rightarrow Adj.R^2 = 0,9873$$

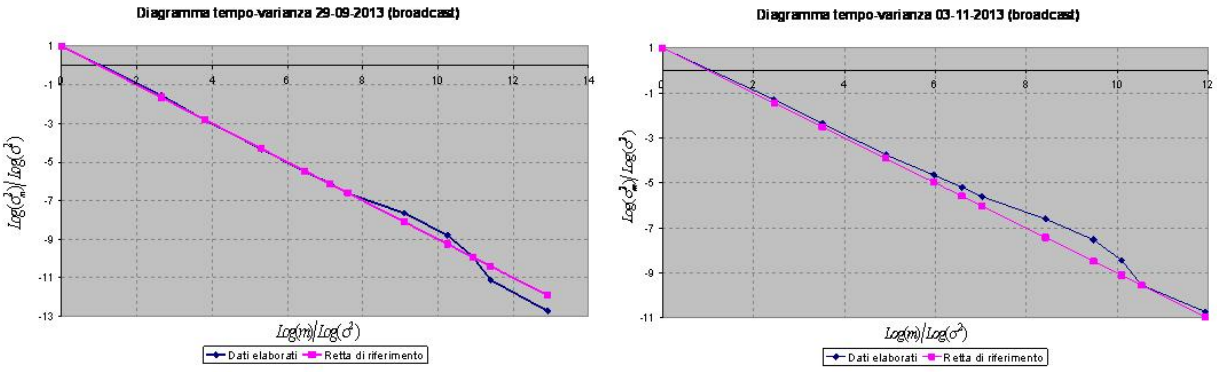
$$\text{Fit Poisson del 03-11-2013} \rightarrow Adj.R^2 = 0,9898$$

7 CONCLUSIONI

L'analisi statistica del traffico della rete geografica wireless ad accesso pubblico gestita dall'Istituto di Cristallografia presso l'Area della ricerca Roma 1, ha confermato come i dati che attraversano un network informatico abbiano una natura auto-similare, ovvero conservino un andamento altamente variabile e la forma analitica della densità di probabilità che li descrive se osservati attraverso diverse scale temporali. Ciò a causa dei meccanismi che regolano la diffusione delle informazioni in una simile struttura, primo fra tutti la trasmissione a commutazione di pacchetto, che rende altresì inadeguata l'interpretazione teorica condotta con la statistica di Poisson, più idonea alla descrizione di quei processi che tendono a regolarizzarsi all'aumentare della scala dei tempi di osservazione.

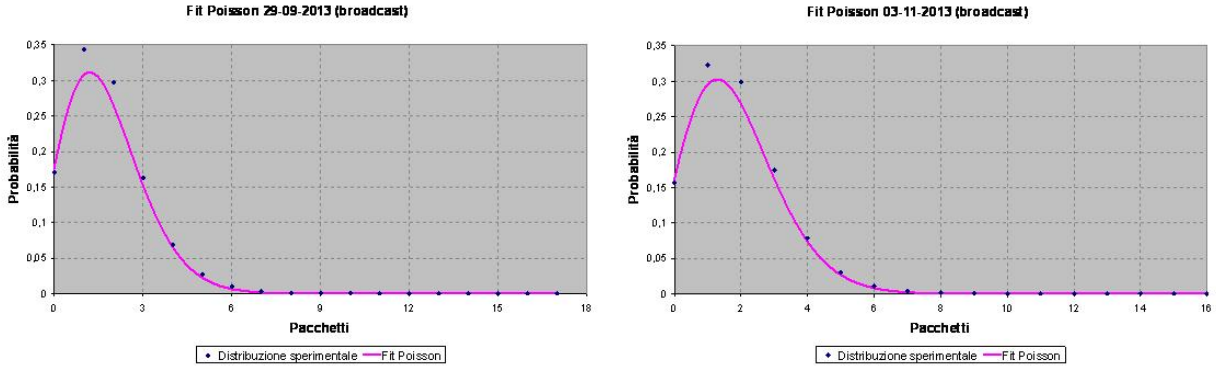
Il campionamento, la successiva aggregazione dei dati e la costruzione di diagrammi tempo-varianza, aventi come variabile casuale oggetto di studio il numero dei pacchetti contenuti nei campioni generati, hanno messo in evidenza il comportamento asintoticamente autosimilare del traffico. I fit non lineari elaborati hanno mostrato come la distribuzione log-normale sia quella che abbia maggiore aderenza con le densità sperimentali ottenute, escludendo la funzione di Weibull come possibile alternativa.

Avendo effettuato acquisizioni su una rete che non adotta sistemi per limitare o circoscrivere i pacchetti di broadcast, è stato possibile isolare il relativo traffico dal flusso complessivo per condurre un'analisi separata. Data l'esigua presenza, questa componente non rappresenta un fattore di degrado delle prestazioni della WAN; dall'indagine statistica sono emersi risultati opposti a quelli ricavati dall'intero traffico. La propagazione dei pacchetti di broadcast non segue il modello auto-similare bensì quello poissoniano, ovvero regolarizza il suo andamento al crescere della dimensione dei campioni. Ciò non deve sorprendere poiché è connesso con il meccanismo di diffusione di questo tipo di dati: un pacchetto di broadcast emesso da una sorgente arriva a tutte le destinazioni da



(a) **(b)**

Fig. 26 Diagrammi tempo-varianza relativi al traffico di broadcast dei giorni festivi.



(a) **(b)**

Fig. 27 Fit poissoniani realizzati con il traffico di broadcast dei giorni festivi

essa raggiungibili. Pur adottando il metodo del packet switching, i dispositivi instradatori replicano il traffico su tutti i rami a cui sono collegati. A meno di improvvisi picchi di attività, ovvero in condizioni stazionarie i dati di broadcast tendono quindi a diffondersi uniformemente sul dominio a loro disposizione, in questo caso l'intera rete, proprio questa caratteristica li rende adatti ad essere interpretati attraverso il modello di Poisson.

Riferimenti

1 M. Becchi, From Poisson Processes to Self-Similarity: a Survey of Network Traffic Models, consultato il 31 ottobre 2012. Reperibile all'indirizzo: http://www.cse.wustl.edu/~jain/cse567-06/ftp/traffic_models1/index.html.

2 W. Willinger, V. Paxson, Where Mathematics Meets the Internet, consultato il 3 dicembre 2012. Reperibile all'indirizzo: http://www.ams.org/notices/199808/paxson.pdf.

3 C. M. Kozierok, The TCP/IP guide – Electronic Book, consultato il 13 febbraio 2013. Reperibile all'indirizzo: http://eeweb.poly.edu/el933/papers/Willinger.pdf.

4 W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, On the self-similar nature of ethernet traffic (extended version), consultato il 13 febbraio 2013. Reperibile all'indirizzo: http://eeweb.poly.edu/el933/papers/Willinger.pdf.

5 I. Antoniou, V. V. Ivanov, V. V. Ivanov, P. V. Zrelov, On the log-normal distribution of network traffic, Physica D 167 (2002) 72–85. doi:10.1016/S0167-2789(02)00431-1.

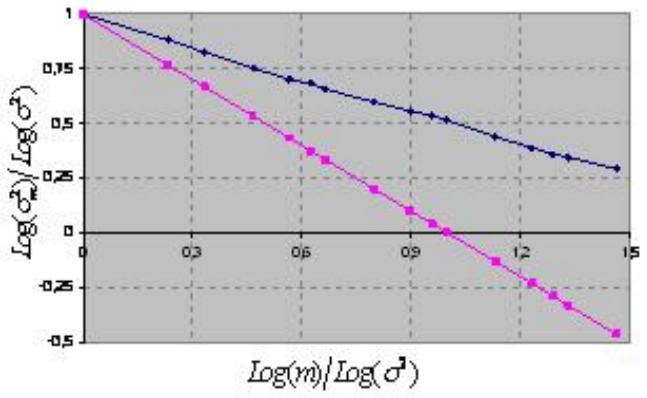


Fig. 28 Diagramma tempo-varianza

8 APPENDICE A - SUCCESSIVE ANALISI DEL 31-10-2013

Come accennato nella sezione 6, nella cronologia del traffico di giovedì 31-10-2013 definita come numero di pacchetti transitati al nodo di raccolta nel tempo di campionamento di 180s, è stato isolato un intervallo di dati (linee gialle nella Fig. 13b)) la cui analisi è stata condotta separatamente poiché rispetto agli insiemi elencati nella Tabella 1, in particolare ai periodi di alto utilizzo, mostra una più evidente variabilità conservando per un tempo maggiore l'andamento stazionario richiesto. Lo scopo di tale ulteriore esame è stato verificare l'ipotesi di ottenere una diversa forma della densità di probabilità per la variabile casuale "popolazione dei campioni". Non una formulazione analitica nuova ma una distribuzione con una coda pronunciata, rispetto a quelle quasi simmetriche mostrate nelle Fig. 18 e 21, ottenute da intervalli di dati poco variabili intorno al rispettivo valore medio oppure di breve durata. Ovviamente è stato anche valutato l'effetto del particolare andamento dei dati selezionati sulle proprietà autosimilari del sistema, attraverso la costruzione di un diagramma tempo-varianza. Vengono di seguito riportate alcune informazioni di carattere generale sull'intervallo di traffico analizzato, coerentemente con quanto descritto nella Tabella 1.

- Durata: 3h 12m
- Pacchetti raccolti: 109.343.417
- Dimensione dati: 70,6GB
- Throughput medio: 49Mbps

Il tempo di campionamento per l'analisi statistica è stato di 5ms, scelto seguendo i criteri già esposti nella sezione 6; i campioni sono stati successivamente aggregati con valori di "m" compresi fra 1 e 25.000. In Fig. 28 viene mostrato il diagramma tempo-varianza ottenuto.

Da un confronto con le precedenti elaborazioni, appare ora evidente la natura esattamente autosimilare dei dati: l'andamento rettilineo del tracciato sperimentale si manifesta anche per bassi valori di "m". Il grado di autosimilarità non è molto diverso da quelli già riscontrati, avendo rilevato un valore del parametro di Hurst $H = 0,759 \pm 0,003$; la differenza sostanziale risiede nella rapidità con cui tale condizione viene raggiunta ed in seguito mantenuta al variare della scala dei tempi. L'impossibilità di ottenere un soddisfacente adattamento tra la densità di probabilità sperimentale ed il modello poissoniano è coerente con questo risultato. Il software Origin Pro infatti non è stato in grado di realizzare un fit adeguato con la distribuzione di Poisson.

La funzione log-normale si è dimostrata ancora una volta più idonea, rispetto a quella di Weibull, nell'interpretare le proprietà statistiche del traffico, tuttavia solo le curve corrispondenti a valori di "m" non superiori a 10 mostrano code moderatamente pronunciate. Al crescere di "m" la definizione del processo di aggregazione riduce la varianza delle distribuzioni rendendole analoghe a quelle già viste, come mostrato nella Fig. 29.

Questa analisi ha messo dunque in evidenza come, mantenendo le condizioni di stazionarietà, una crescente variabilità del traffico possa segnare il passaggio dall'autosimilarità asintotica a quella esatta. Ciò non implica particolari cambiamenti nei valori assunti dal parametro di Hurst o nella forma delle densità di probabilità sperimentali, ma una maggiore rapidità e stabilità con cui il sistema manifesta le sue proprietà frattali su diverse scale temporali ed un distacco definitivo dalle caratteristiche del modello di Poisson.

9 APPENDICE B – COEFFICIENTI STATISTICI

9.1 Coefficiente di determinazione

Nelle analisi statistiche il *coefficiente di determinazione* (*coefficient of determination* o *R square*) indicato con il simbolo R^2 , fornisce una misura del grado di aderenza di una curva ad un insieme discreto di punti. Viene utilizzato per la valutazione di modelli statistici nell'ambito delle previsioni o dei test di ipotesi e indica l'efficienza con cui detti modelli riescono a descrivere i risultati sperimentali.

Esistono diverse definizioni del coefficiente R^2 , che solo in alcuni casi possono essere considerate tra di loro equivalenti. In genere nelle regressioni lineari R^2 assume valori compresi fra 0 e 1, soprattutto se tra i parametri da stimare viene inclusa anche l'intercetta; risultati negativi nel calcolo di R^2 possono invece essere ottenuti nei fit non lineari.

Si supponga che $y_i (i = 1, \ldots, n)$ sia uno dei risultati di una serie di n osservazioni condotte su una ipotetica grandezza e che f_i rappresenti il corrispondente valore teorico previsto. In quanto segue viene indicato con y_m il valor medio delle misure effettuate:

$$y_m = \frac{1}{n} \sum_{i=1}^{n} y_i$$

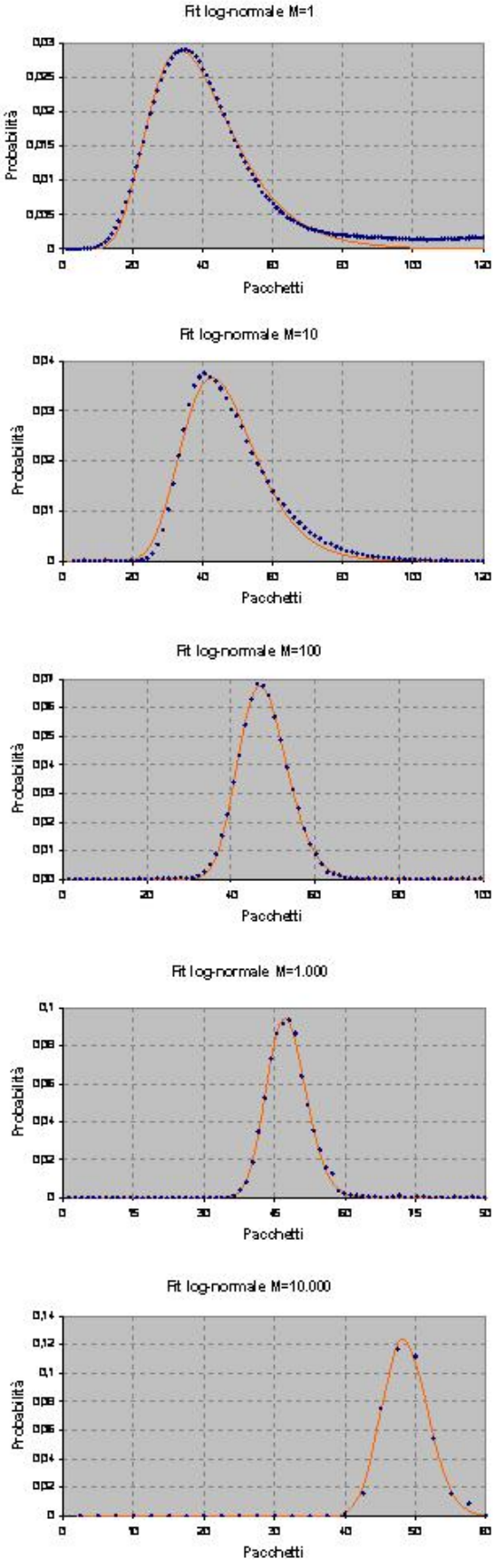


Fig. 29 Fit realizzati con la distribuzione log-normale.

La variabilità dell'insieme di dati viene espressa attraverso differenti somme quadratiche.

- Somma quadratica totale (*total sum of squares*)
  $SS_{tot} = \sum_i (y_i - y_m)^2$
- Somma quadratica di regressione (*regression sum of squares*)
  $SS_{reg} = \sum_i (f_i - y_m)^2$
- Somma quadratica dei residui (*residual sum of squares*)
  $SS_{res} = \sum_i (y_i - f_i)^2$

La definizione più generale del coefficiente di determinazione è data da:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Questo parametro dà un'indicazione sulla bontà del fit tra le osservazioni sperimentali ed il modello teorizzato; un valore di R^2 esattamente pari a 1 rappresenta il perfetto accordo tra gli insiemi messi a confronto. Tuttavia in linea di principio, il coefficiente di determinazione non stabilisce se la funzione con cui si tenta di descrivere il processo statistico sia corretta o meno.

9.2 Coefficiente di determinazione modificato

L'uso del *coefficiente di determinazione modificato* (*adjusted R square*) spesso indicato con il simbolo R^2, rappresenta un metodo per controllare l'aumento indesiderato di R^2 dovuto all'eventuale aggiunta, nel modello da valutare, di parametri da determinare statisticamente sulla base dei dati sperimentali di cui si dispone. Da un punto di vista analitico prende in considerazione il rapporto tra il numero p di quantità da ricavare attraverso il procedimento di regressione e il numero n di osservazioni effettuate. Il coefficiente R^2 può essere negativo e il suo valore è sempre minore o uguale a quello di $\underline{R}^2$; la sua definizione è la seguente:

$$\underline{R}^2 = R^2 - \left(1 - R^2\right) \frac{p}{n - p - 1}$$

Questo parametro è un indice della bontà di un fit proprio come R^2, ma fornisce anche una misura comparativa dell'adeguatezza di un modello al variare del numero di incognite da determinare.

9.3 Akaike Information Criterion

Il test *Akaike Information Criterion* (AIC) si esegue 1 quando si desidera confrontare più modelli statistici per stabilire quale si adatta meglio ad un fissato insieme di dati sperimentali. Per effettuarlo occorre prima aver realizzato i fit tra i valori osservati e i modelli teorici ed aver quindi determinato i parametri delle corrispondenti funzioni.

Detto n il numero di osservazioni, p quello delle incognite di un modello e SS_{res} la somma quadratica dei

residui già introdotta, la definizione di AIC dipende dal rapporto fra n e p secondo quanto segue:

$$AIC = \begin{cases} nln\left(\frac{SS_{res}}{n}\right) + 2p, & \frac{n}{p} \geq 40 \\ nln\left(\frac{SS_{res}}{n}\right) + 2p + \frac{2p(p+1)}{n-p-1}, & \frac{n}{p} \leq 40 \end{cases}$$

In genere la quantità che ha maggior peso nel calcolo di AIC è il rapporto SS_{res}/n quindi a parità di popolazione statistica, quanto più piccola è la somma quadratica dei residui tanto minore è l'argomento del logaritmo. Sono dunque preferibili valori negativi di AIC e fra i modelli messi a confronto, quello che offre il minor valore di tale parametro si presta meglio a rappresentare i dati sperimentali.