Análisis del Comportamiento Autosimilar del tráfico Ethernet de las Redes de Datos

Santiago C. Pérez, Higinio A. Facchini, Luis Bisaro
Grupo de Investigación y Desarrollo en TICs (GRID TICs)
Universidad Tecnológica Nacional, Facultad Regional Mendoza
santiagocp@frm.utn.edu.ar

y

Jesús Rubén Azor
Cátedra de Estadística
Universidad de Mendoza - Mendoza
jesús.azor@um.edu.ar

Abstract

The queuing analysis has been and is of enormous usefulness for the designers of networks and analysts of traffic, to effects of planning the capacities of the elements of network and of predicting his performance. These analyses depend on the nature Poisson of the traffic of data. Nevertheless, many results predicted from the analysis of queuings differ significantly from the performance observed in the reality. Diverse studies have demonstrated that for some environments the pattern of traffic is self-similar, instead of Poisson. This concept is related to other acquaintances since it are the fractals and the theory of the chaos. From beginning of the 90s were begun to publish documents referred to the self-similarity of the traffic of Ethernet. The present work develops successively the following paragraphs: 1) Self-similarity, 2) Self-similar data traffic, 3) Ethernet data traffic, 4) Case of experimental Study of Ethernet traffic, 5) Analysis with distribution Pareto, and 6) Analysis of goodness of fit with Kolmogorov-Smirnov's test.

Key words: traffic, self-similarity, Ethernet, Pareto, goodness of fit

Resumen

El análisis de colas ha sido y es de enorme utilidad para los diseñadores de redes y analistas de tráfico, a efectos de planificar las capacidades de los elementos de red y predecir su rendimiento. Estos análisis dependen de la naturaleza Poisson del tráfico de datos. Sin embargo, muchos resultados predichos a partir del análisis de colas difieren significativamente del rendimiento observado en la realidad. Diversos estudios han demostrado que para algunos entornos el patrón de tráfico es autosimilar, en lugar de Poisson. Este concepto está relacionado con otros más conocidos como son los fractales y la teoría del caos. Desde principio de los años 90 se comenzaron a publicar documentos referidos a la autosimilitud del tráfico de Ethernet. El presente trabajo desarrolla sucesivamente los siguientes apartados: 1) Autosimilitud, 2) Tráfico de datos autosimilar, 3) Tráfico de datos Ethernet, 4) Caso de Estudio experimental de tráfico Ethernet, 5) Análisis con distribución Pareto, y 6) Análisis de bondad de ajuste con la prueba de Kolmogorov-Smirnov.

Palabras claves: tráfico, autosimilitud, Ethernet, Pareto, bondad de ajuste

1 AUTOSIMILITUD

La autosimilitud es un concepto muy importante, aunque sólo se aplicó al análisis del tráfico de comunicaciones de datos en los últimos tiempos [1], y se puso de manifiesto en una afirmación efectuada por Manfred Schroeder [2]. La autosimilitud, o invariancia frente a campos de escala o tamaño, es un atributo de muchas leyes de la naturaleza.

Los fenómenos autosimilares tienen el mismo aspecto o comportamiento cuando se visualizan con distintos grados de ampliación o a distintas escalas en una cierta dimensión. La dimensión puede ser el espacio (longitud, anchura) o el tiempo, aunque en este trabajo interesan las series temporales y los procesos que muestran una autosimilitud con respecto al tiempo.

Muchos ejemplos se han derivado del conjunto de Cantor, una famosa estructura que aparece en casi toda la bibliografía relativa a caos, fractales y dinámica no lineal. El conjunto de Cantor revela dos propiedades que se aprecian en todos los fenómenos autosimilares:

1. Posee estructura hasta escalas arbitrariamente pequeñas. Si se amplía una parte del conjunto repetidamente, se sigue observando un patrón complejo de puntos separados por huecos de distintos tamaños (figura 1-1).
2. Las estructuras se repiten. Las estructuras autosimilares contienen réplicas más pequeñas de sí mismas en todas las escalas.

Estas propiedades no siguen siendo ciertas indefinidamente para los fenómenos reales. La estructura y la similitud dejan de mantenerse en algún momento del proceso de ampliación. Pero también es cierto que en un notable rango de escalas hay muchos fenómenos que muestran autosimilitud.

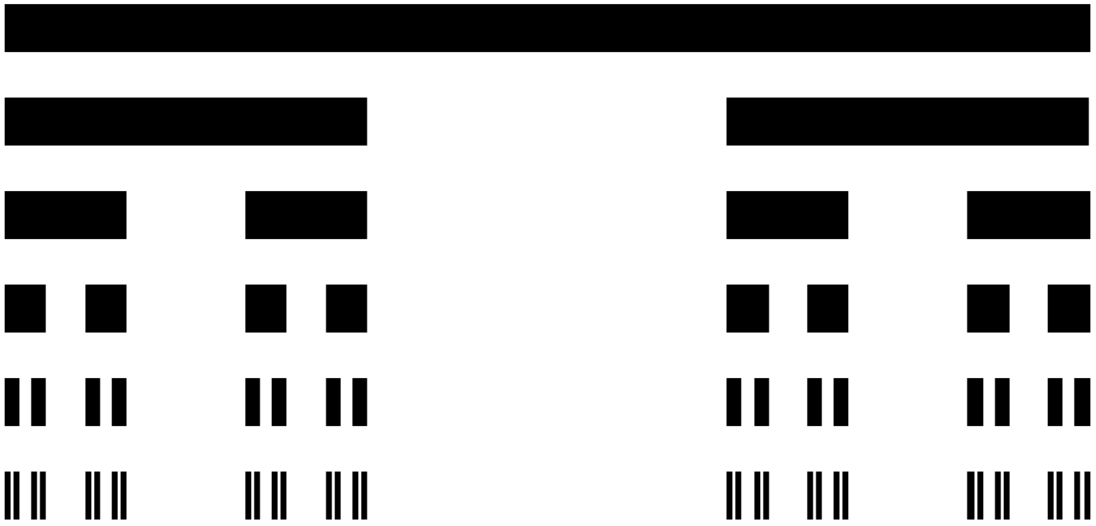

Figura 1-1

Aunque este es un sencillo ejemplo, su estudio puede aportar algunas ideas válidas para el tráfico de datos autosimilar. Es posible que la característica que más se destaque, desde el punto de vista del rendimiento de la red, es la persistencia de los agrupamientos. En un tráfico Poisson, los agrupamientos se producen a corto plazo (en una escala temporal pequeña), pero se van suavizando a largo plazo. Esto da lugar a la observación de que el análisis de colas tradicional, que supone un tráfico de Poisson, no prediga con precisión el rendimiento de un tráfico autosimilar.

2 TRÁFICO DE DATOS AUTOSIMILAR Y LA DISTRIBUCIÓN PARETO

Las señales autosimilares deterministas son invariantes frente a cambios de escala. Para un proceso estocástico, se puede decir que los estadísticos del proceso no cambian cuando cambia la escala

temporal. Tanto desde el punto de vista cualitativo como desde el cuantitativo, el proceso carece de una escala característica: el comportamiento medio del proceso a corto plazo es igual a su comportamiento medio a largo plazo.
La figura 2-1 muestra a la izquierda un ejemplo de proceso estocástico autosimilar. Obsérvese que la función temporal no se reproduce exactamente a distintas escalas temporales, pero las ondas a distintas escalas temporales se parecen entre sí. Compárese esto con el proceso estocástico estacionario típico que está a la derecha de la figura 2-1. En este caso, se observa que a niveles más elevados de ampliación la función posee características diferentes, volviéndose menos suave y más irregular. Desde el punto de vista opuesto, cuando se examinan escalas temporales más largas, la señal parece mostrar menos fluctuación y ser más regular.

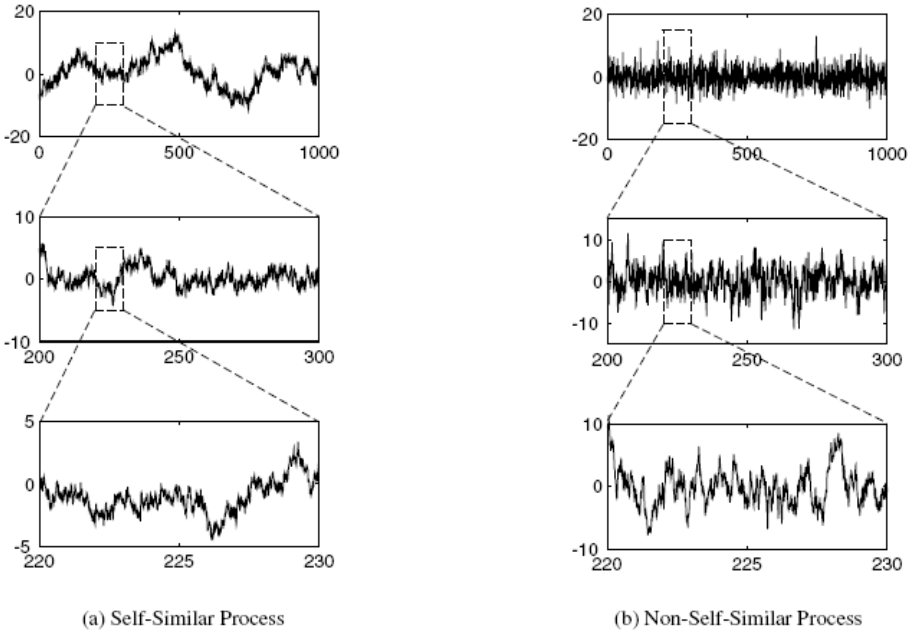

Figura 2-1

El parámetro H, que se denomina parámetro de Hurst, o parámetro de autosimilitud, es una medida clave de la autosimilitud. Más exactamente, H es una medida de la persistencia de un fenómeno estadístico y es un indicador de la longitud de la dependencia a largo plazo de un proceso estocástico. Un valor de H=0,5 indica la ausencia de dependencia a largo plazo. Cuanto más próximo esté H a 1, mayor será el grado de persistencia o de dependencia a largo plazo.
Se pueden definir procesos estocásticos autosimilares con distribuciones de cola hiperbólica. Uno de los atractivos de las distribuciones hiperbólicas es que dan lugar a modelos de simulación manejables.
Las distribuciones de cola periódica sirven para caracterizar densidades de probabilidad que describen procesos de tráfico como los tiempos entre llegadas de paquetes y las longitudes de ráfaga. En general, una variable aleatoria con distribución de cola hiperbólica muestra una varianza infinita y posiblemente una media infinita. Las variables aleatorias con distribución de cola hiperbólcia contienen valores muy grandes con una probabilidad no despreciable. Generalemente, si se muestra una de estas variables, el resultado incluirá muchos valores relativamente pequeños, pero también unos pocos valores relativamente grandes.
La distribución de cola hiperbólica más sencilla es la distribución de Pareto con parámetros a y b, cuyas funciones de densidad y de distribución son

$$f(x) = \frac{ab^a}{x^{a+1}}$$

$$F(x) = 1 - \left(\frac{b}{x}\right)^a$$

para valores de $x \geq b$

El parámetro ***b*** especifica el valor mínimo que puede tomar la variable aleatoria. El parámetro ***a*** determina la media y la varianza de la variable aleatoria. Cuando se comparan las funciones de densidad de Pareto y exponencial en una escala semilogarítmica, se observa que en esta escala la función densidad exponencial es una recta, reflejando el decrecimiento exponencial de la distribución. El final de la distribución de Pareto decrece mucho más lentamente que la exponencial; de aquí viene el nombre de cola hiperbólica.
La distribución de Pareto se ha observado en una amplia gama de fenómenos procedentes de las ciencias sociales y físicas, y del mundo de las comunicaciones [3]. La cola hiperbólica de ciertas variables de las redes (por ejemplo, los tamaños de los archivos y las duraciones de las conexiones) son la causa seminal subyacente de la dependencia de largo alcance y de la autosimilitud que se observa en el tráfico de red [4].

3 TRÁFICO DE DATOS ETHERNET

El artículo fundamental del estudio de los datos de tráfico autosimilar es «On the Self-Similar Nature of Ethernet Traffic» (La naturaleza autosimilar del tráfico de Ethernet), que posteriormente sería corregido y aumentado en [5]. Este documento contradijo la idea de que un simple análisis de colas basado en la suposición de que el tráfico fuera de Poisson pudiera modelar adecuadamente todo tráfico de red. Empleando una masiva cantidad de datos y un cuidadoso análisis estadístico, el artículo manifiesta que, para el tráfico de Ethernet, se requiere un nuevo planteamiento de modelado y de analísis.
La columna izquierda de la figura 3-1 muestra gráficos del número de paquetes por unidad de tiempo correspondientes a un conjunto de medidas de 1989, formando por más de 27 horas de monitorización continua del tráfico de Ethernet. El primer gráfico muestra todo el experimento de 27 horas, con unidades de tiempo de 100 segundos, dando lugar a un gráfico de 1.000 puntos. Los gráficos subsiguientes se obtienen a partir del anterior incrementando la resolución temporal por un factor 10 y visualizando un subintervalo seleccionado aleatoriamente (que se indica mediante un tono más oscuro). Por tanto, el segundo gráfico abarca un período aproximado de 2,7 horas, el tercero abarca 0,27 horas y así sucesivamente. Si se considera en dirección opuesta, a medida que avanzamos hacia arriba en la primera columna de gráficos, un punto del gráfico se forma promediando los diez puntos de datos correspondientes en el gráfico inmediatamente inferior.
Se pueden hacer varias observaciones interesantes con respecto de estos datos. Con la posible excepción del primer gráfico, todos los gráficos se parecen entre sí desde el punto de vista de la distruibución.
Esto es, todos los gráficos muestran cierta cantidad de picos o ráfagas. De este modo, el tráfico de Ethernet posee un aspecto similar para escalas grandes (horas y minutos) y para escalas pequeñas (segundos y milisegundos). Este tráfico autosimilar es muy diferente de lo que se observa en el tráfico telefónico y en los modelos estocásticos usando Poisson de los análisis y diseño de redes de datos.

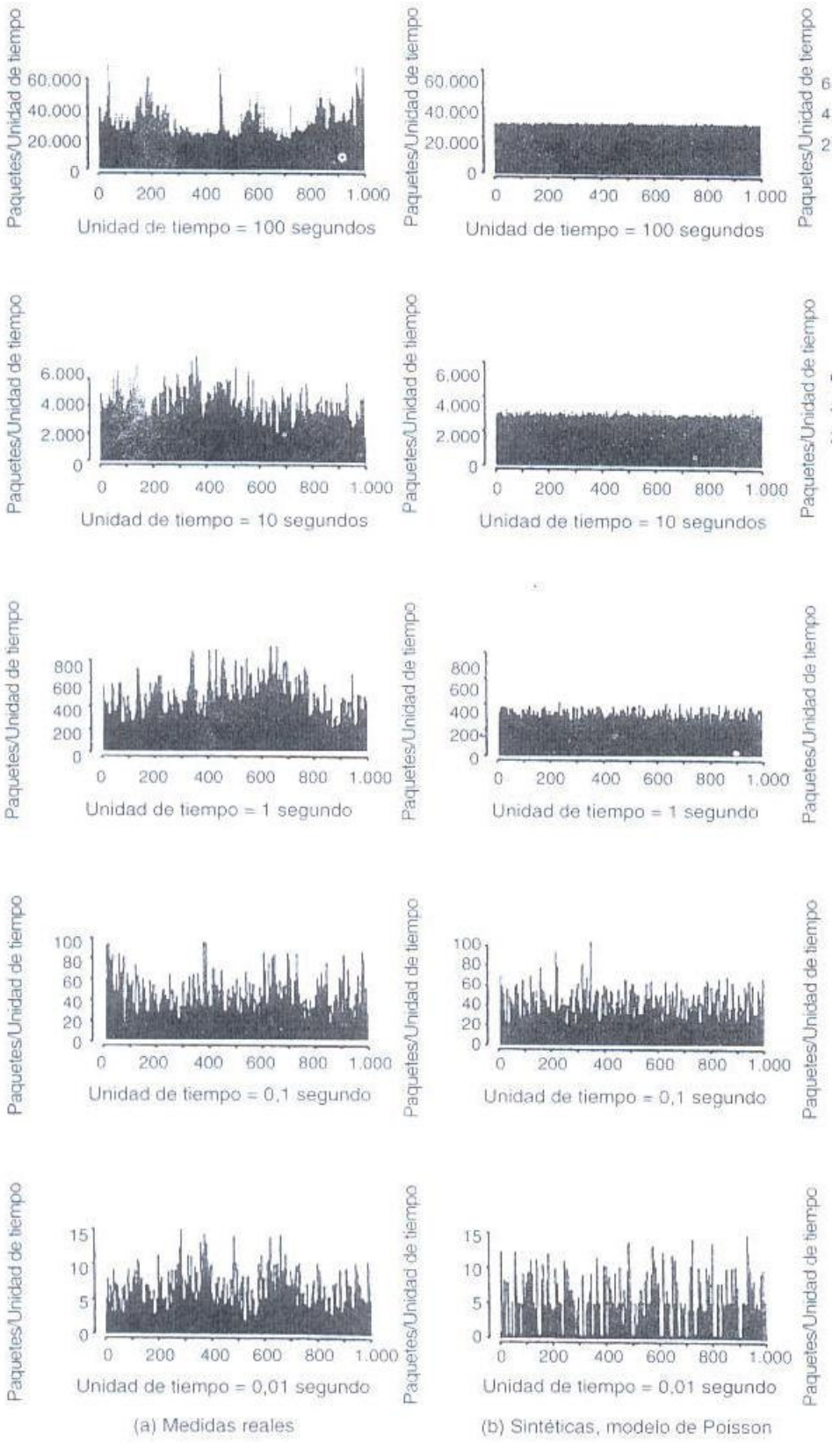
Paquetes/Unidad de tiempo
60.000
40.000
20.000
0
0 200 400 600 800 1.000
Unidad de tiempo = 100 segundos
6.000
4.000
2.000
Unidad de tiempo = 10 segundos
800
600
400
200
Unidad de tiempo = 1 segundo
100
80
60
40
20
Unidad de tiempo = 0,1 segundo
15
10
5
Unidad de tiempo = 0,01 segundo
(a) Medidas reales
(b) Sintéticas, modelo de Poisson

Figura 3-1

Este contraste se nota con la columna de la derecha, generada de igual modo que los gráficos de la columna de la izquierda, pero empleando un modelo Poisson. Con alta resolución (unidad de tiempo 0,1 seg), el tráfico tienen bastantes picos. A medida que se van agregando los datos mediante escalas de tiempo cada vez mayores, el patrón de tráfico se suaviza. Por lo tanto, es de esperar que la varianza de los datos se reduzca por un factor de 10 por cada nivel, a diferencia de lo que sucede en un tráfico autosimilar.
Por ello, en las simulaciones se prefiere modelar los periodos de tiempo de tráfico, con distribuciones de varianza infinita, utilizando en particular la distribución de Pareto. Esto da como resultado una distribuciçon de elevada varianza, con muchas ráfagas muy cortas, muchas ráfagas largas y algunas ráfagas muy largas. Esto ha permitido determinar el origen de las descrepancias, por ejemplo, entre el tiempo real de espera y el tiempo estimado de espera obtenidos mediante el uso de la teorçia de colas convencional usando Poisson.

4 CASO DE ESTUDIO EXPERIMENTAL DE TRÁFICO ETHERNET

Los métodos de colección de tramas de red Ethernet, son el punto de partida para el entendimiento del comportamiento del tráfico y de los nodos de red [6]. Pueden ser clasificadas en tres categorías: 1) métodos basados en polling, los cuales registran las asociaciones de los nodos de red en intervalos de tiempo periódicos, usando el protocolo snmp, o algún software de tracking de asociación sobre los nodos, 2) métodos basados en programas que registran eventos online/offline de usuarios de red usando un servidor de logeo (syslog) como archivos de sesión, de dhcp, de traps, etc. y 3) métodos basados en programas sniffers que colectan el tráfico de la red en la medida que se produce.
A los fines del trabajo, en el que se pretende analizar algún patrón característico del tráfico de red Ethernet, es importante incluir en la colección de trazas, las tramas de tráfico con la mayor cantidad de información posible usando sniffers. Existe una extensa variedad de estos programas para analizar tramas y trafico de red, como TCPdump [7], IPtraf [8], Wireshark [9] (ex Ethereal), NTOP (Network TOP) [10], entre otros.
Para el estudio experimental, se tomó una muestra de N=19300 tramas Ethernet, durante 550 segundos, utilizando el sniffer EtherPeek [11]. Una vez colectadas las tramas, se comenzó su análisis en la misma herramienta, dado que posee una gran flexibilidad en su interfaz, identificando en cada una el protocolo de capa superior TP/IP involucrado, su longitud en bytes, etc y especialmente el instante de tiempo de muestreo. Además, da la posibilidad de exportar los datos a una planilla de cálculo para facilitar su manipulación y representación.
La figura 4.1 muestra la representación de la cantidad de tramas en función del tiempo, en los primeros 60 segundos, con una resolución de 1 segundo. Puede observarse la similitud a las gráficas de la figura 3.1, que se mantiene para distintas resoluciones, ratificando que el tráfico es del tipo autosimilar, como se ha mencionado previamente.
Luego, las tramas identificadas cronológicamente y sin los campos innecesarios, se exportaron como un vector al programa Matchcad [12], para proceder a su análisis estadístico. Las tramas se agruparon en un vector $\boldsymbol{A}$ de 55 elementos, en intervalos de 10 segundos, y se ordenaron en forma decreciente según la cantidad de tramas, como se indica a continuación:

$A^T =$

	0	1	2	3	4	5
0	$8.763 \cdot 10^3$	$3.677 \cdot 10^3$	$1.992 \cdot 10^3$	840	735	...

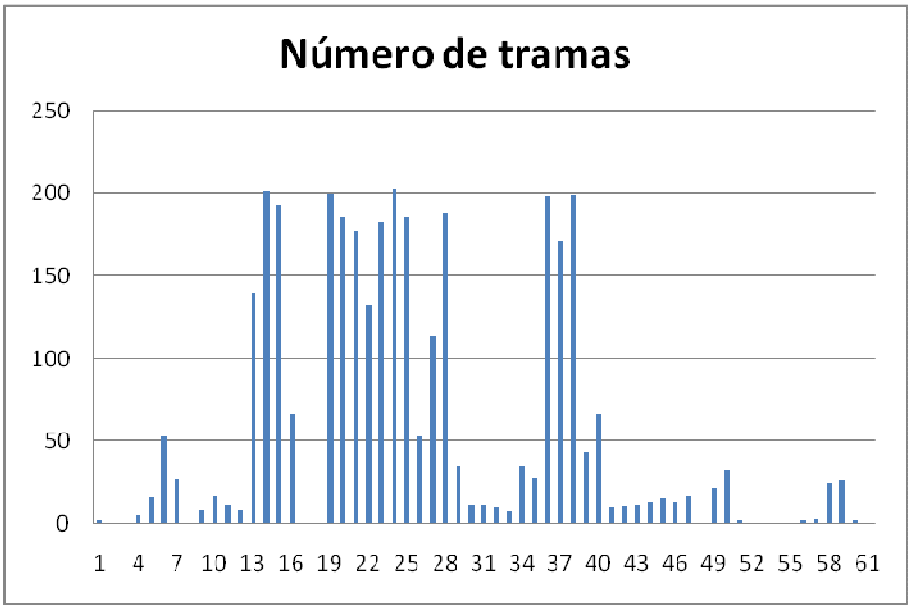

figura 4.1

Posteriormente, estos elementos se normalizaron dividiendo el vector A por N, dando origen a un nuevo vector que se llamará O (vector de los valores observados), a los fines del resto del artículo:

$O^T =$

	0	1	2	3	4
0	0.454	0.19	0.103	0.044	...

5 ANÁLISIS CON DISTRIBUCIÓN PARETO

En estadística, la distribución Pareto, formulada por el sociólogo Vilfredo Pareto, es una distribución de probabilidad continua con dos parámetros a y b cuya función de densidad para valores $x \geq b$ es:

$$f(x) = \frac{ab^a}{x^{a+1}}$$

Y su función de distribución es:

$$F(x) = 1 - \left(\frac{b}{x}\right)^a$$

El valor esperado y la varianza de una variable aleatoria X de distribución Pareto son

$$E[X] = \frac{ab}{a-1}$$

$$V[X] = \frac{ab^2}{(a-1)^2(a-2)}$$

La distribución de Pareto, puede expresarse como una función f(x,a,b), de la siguiente forma:

$$f(x,a,b) := \frac{a \cdot b^{a}}{x^{a+1}}$$

Asignando a los parámetros los valores a=0,9 y b=1, se puede generar el vector E (vector de valores esperados) para la distribución de Pareto, con x variando entre 1 y 10.

$$i := 1 .. 10$$

$$E_{i-1} := 1 \cdot \int_{i}^{i+1} f(x,0.9,1)\, dx$$

La figura 5-1 muestra la representación del vector observado O y del vector esperado E, en función de i, variando entre 0 y 10, construida usando la herramienta Mathcad.

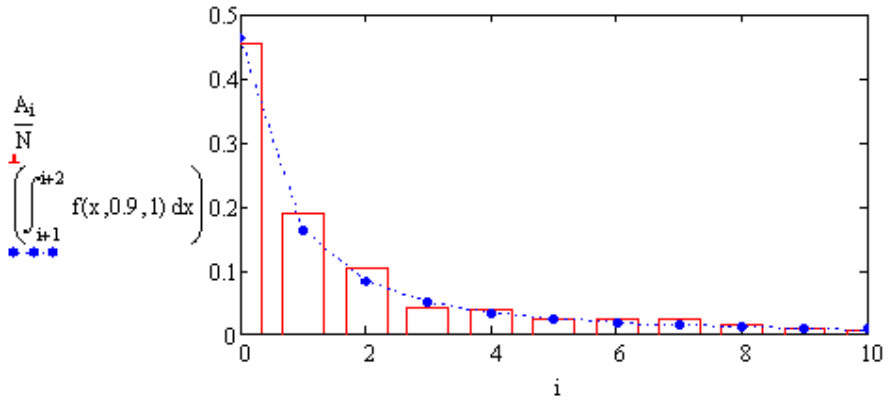

Figura 5-1

6 ANÁLISIS DE BONDAD DE AJUSTE CON LA PRUEBA DE KOLMOGOROV-SMIRNOV.

El uso de la Estadística es de gran importancia en la investigación científica. Casi todas las investigaciones aplicadas requieren algún tipo de análisis estadístico para que sea posible evaluar sus resultados. Por ejemplo, los tests o dócimas paramétricos y no paramétricos.
Dentro de las pruebas no paramétricas, se destacan las pruebas de Kolmogorov-Smirnov para una y dos muestras. Se han propuesto diferentes métricas para describir y comparar utilizando diferencias entre distribuciones acumuladas. La prueba unimuestral de Kolmogorov-Smirnov es una prueba de Bondad de Ajuste apropiada para este caso en que se está usando la distribución Pareto [13]. Es más eficiente que la prueba χ^2 en muestras pequeñas, y no se aplica a distribuciones discretas.
La prueba unimuestral se funda en la diferencia absoluta máxima D entre los valores de la distribución acumulada de una muestra aleatoria de tamaño n, y una distribución teórica determinada. Para decidir si esta diferencia es mayor de la razonablemente esperada con un nivel de significación α, se buscan los valores críticos de ***D*** en Tablas apropiadas.
En el caso en cuestión, se comprobó con un nivel de significancia de α=0.05, que los valores del vector ***O_ac*** (observados acumulados) y del vector ***E_ac*** (esperados acumulados) tuviesen un ***Dmax*** menor a 0.457.

$O\_ac =$

	0
0	0.454
1	0.644
2	0.748
3	0.791
4	0.829
5	0.853
6	0.877
7	0.9
8	0.915
9	0.925

$E\_ac =$

	0
0	0.464
1	0.628
2	0.713
3	0.765
4	0.801
5	0.826
6	0.846
7	0.862
8	0.874
9	0.884

Se probó esta hipótesis nula con un nivel de significación de 0.05

1. Ho Hipótesis nula: están uniformemente distribuidos
   Ha Hipótesis alterna: no están uniformemente distribuidos
2. Nivel de significación 0.05
3. Criterio: Se rechaza Ho si $\boldsymbol{D} > 0.457$, donde $\boldsymbol{D}$ es la diferencia máxima entre la distribución acumulada observada y la supuesta bajo la Hipótesis Nula.
4. Cálculos: Se construyó el siguiente gráfico (figura 6-1) que muestra la distribución acumulada observada y la supuesta.

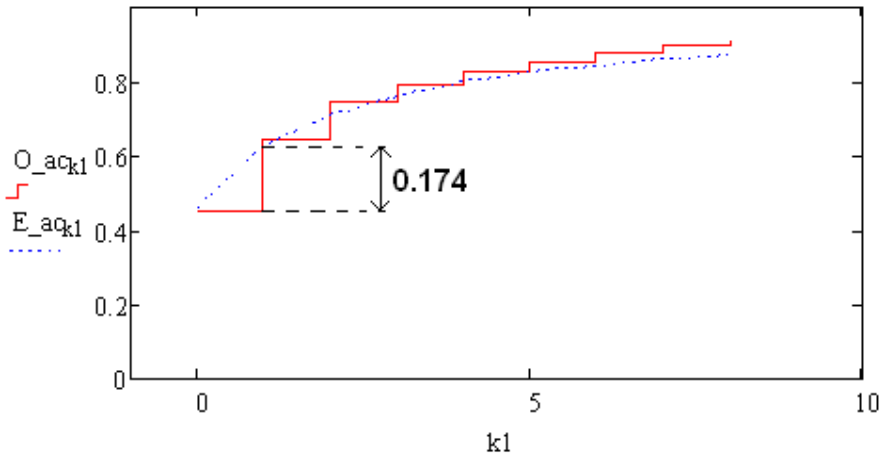

figura 6-1

Dado que:

$$E\_ac_1 = 0.628$$
$$O\_ac_0 = 0.454$$

Se deduce que:

$$D_{max} = E\_ac_1 - O\_ac_0 = 0.174$$

Por otro lado, y de acuerdo a los datos y con un nivel de significación de 0.05, el D crítico de la Tabla 1 es 0.457.

5. Decisión: Ya que $0.174 < 0.457$ (valor de Tabla) *se acepta* la Hipótesis Nula.
   Luego, los valores acumulados están significativamente distribuidos conforme a una Distribución de Pareto con coeficientes a=0.9 y b=1, y por lo tanto, el tráfico de datos Ethernet responden a la citada distribución con ese nivel de significancia.

TAMAÑO DE LA MUESTRA (N)	NIVEL DE SIGNIFICANCIA PARA D = MAX [$F_0(X)$ - $S_n(X)$]				
	.20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490

Tabla 1

7 CONCLUSION

En este documento, se han relacionado los temas de autosimilitud, con el tráfico Ethernet, la distribución de Pareto y la prueba de Kolmogorov-Smirnov. El volumen de los trabajos y literatura sobre tráfico de datos es creciente, y el tema de la autosimilitud ha significado el principio de un nuevo examen del rendimiento del tráfico de datos, las técnicas de modelado, y control de tráfico, entre otros. En el trabajo se ha verificado a través de un estudio experimental y usando la prueba de Kolmogorov-Smirnov, que el tráfico de datos Ethernet responde efectivamente a la distribución Pareto.

8 REFERENCIAS

[1] Stalling, W.; "Redes e Internet de Alta Velocidad. Rendimiento y QoS", Pearson, 2003.

[2] Schroeder, M.; "Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise", Nueva York, Freemam, 1991.

[3] Paxson,V. y Floyd, S.; "Wide Area Traffic: The Failure of Poisson Modeling", IEEE ACM Transacctions on Networking, 2000.

[4] Park, K. y Williams, W.; "Self-Similar Network Traffic and Performance Evaluation", Nueva York, Wiley, 2000.

[5] Leland, W; Taqqu, M; Willinger, W. y Wilson, D; "On the Self-Similar Nature of Ethernet Traffic", IEEE/ACM Transaccions on Networking, Febrero 1994

[6] Perez, S; Mercado, G. y Facchini, H.; "Análisis y Determinación de Patrones de Tráfico de Protocolos en redes LAN", WICC 2007, 2007.

[7] http://www.tcpdump.org

[8] http://iptraf.seul.org

[9] http://www.wireshark.org

[10] http://www.ntop.org/

[11] http://ether-peek.softonic.com/mac

[12] http://www.ptc.com/products/mathcad/

[13] http://www.um.edu.ar/math/estadis/programa.htm