▪ www.intersedes.ucr.ac.cr ▪ Revista InterSedes © Universidad de Costa Rica ▪ ISSN 1409-4746 ▪
Volumen V ▪ Número 8 ▪ 2004 ▪ Edición Digital: 08 / 04 / 2007 ▪ E-mail: intersed@cariari.ucr.ac.cr

DEFINIENDO LA FUNCIÓN DE DISTRIBUCIÓN DE UN PARÁMETRO PARA EL ANÁLISIS DE TRÁFICO

Susan Chen Mok[*]

Recepción: 16 de agosto de 2004 **Aprobación: 12 de noviembre de 2004**

RESUMEN

El siguiente artículo describe el procedimiento empleado para definir la función de distribución de los tiempos entre arribos de paquetes en una red de área local. El trabajo es importante puesto que es difícil encontrar alguna literatura que describa los pasos a seguir para definir este tipo de función. Además, el auge en el uso de la simulación por computadora ha incrementado el uso de softwares para el modelaje y simulación de redes, la mayoría de los cuales requieren que se les provea, para el proceso de modelaje del tráfico de la red, del parámetro de tiempo entre arribos de paquetes a ésta. Por tanto, el artículo provee una guía para los iniciadores en este campo acerca del análisis y definición de funciones de distribución.

Palabras clave: función de distribución, simulación por computadora, modelos de redes, simulación de redes, modelos de tráfico

ABSTRACT

This study describes the procedure to define a distribution function for the times between arrivals of packages on a local area network. This is an important study because it is difficult to find information describing the steps to define this type of function. Furthermore, the increasing use of computers simulation has increased the use of software for network modeling and simulation. The process of modeling network traffic for most of these software requires a time parameter between arrivals of packages to the network. Therefore, this study provides a guideline for the initiators in this field about the distribution functions analysis and definition.

Key words: distribution function, computers simulation, network models, network simulation, traffic models.

INTRODUCCIÓN

Para definir la función de distribución de cualquier parámetro para un análisis de tráfico, se requiere de información real de éste. Para el presente estudio se utilizó una red de área local, compuesta por: 1 conmutador, 2 concentradores de 16 puertos, 1 concentrador de 8 puertos, 4 computadores servidores y más de 30 microcomputadoras personales. La figura No.1 muestra el esquema de la red.

Se utilizó el programa “Landecoder” para capturar el tráfico de la red. Utilizando este “software” se extrajo datos del tráfico de la red, estos datos incluyen:

[*] Sede del Pacífico de la Universidad de Costa Rica [schen@cariari.ucr.ac.cr]

-Información de cada paquete: longitud, tiempo de llegada, tiempo transcurrido desde el último arribo.

-Información de protocolos de los diferentes niveles.

-Porcentaje de utilización de la red

Las aplicaciones que usan la red incluyen: transmisión de archivos a las impresoras, transferencias de archivos a los sistemas de almacenamiento local, programas especializados, correo electrónico y transferencia de archivos dentro de la red y hacia fuera de ella.

A continuación se explica detalladamente todo el proceso que se siguió para la definición de la función de distribución de la variable tiempo entre arribos.

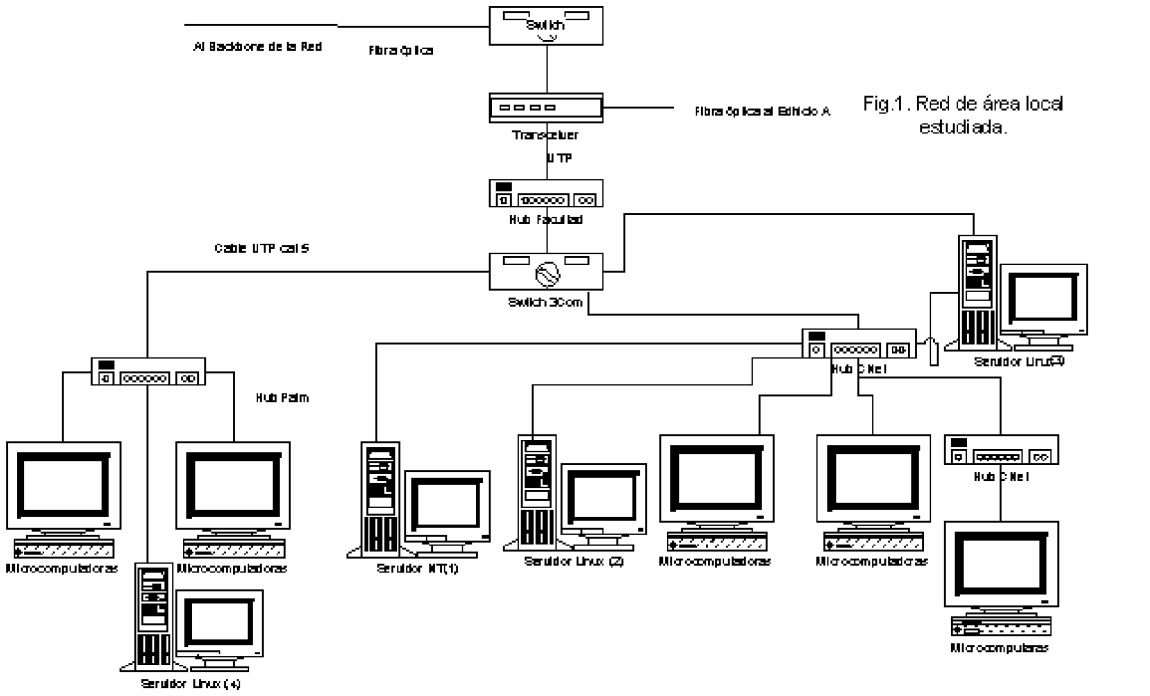


Fig.1. Red de área local estudiada.

DEFINICIÓN DE FUNCIONES DE DISTRIBUCIÓN DE PROBABILIDAD

De los datos reales extraídos se obtuvieron los siguientes valores medios:

Tiempo promedio entre arribos: 0,0293 seg.

Número promedio de arribos por unidad de tiempo: 25 paquetes/ segundo

De acuerdo con lo que se extrae de la teoría existente, tenemos que:

-el número de arribos en un intervalo de un segundo sigue una distribución Poisson, lo que implica que los tiempos entre arribos siguen una distribución exponencial negativa.

En el siguiente apartado se realiza la prueba de bondad de ajuste para verificar si la variable Número de Arribos que llegan en el intervalo de tiempo de un segundo sigue una distribución de Poisson.

Pruebas de bondad de ajuste

Variable: Número de arribos por segundo.

Promedio de la muestra: 25 paquetes por segundo. Eliminando los segundos en los que llegaron más de 80 arribos, para evitar que los valores atípicos afecten demasiado el promedio, se obtiene un promedio de 7 arribos por segundo.
Total de segundos observados: 1862.
Total de segundos eliminando valores atípicos: 1778

Prueba de bondad para ajustar a Poisson

Se llevó a cabo una prueba de bondad de ajuste de Chi-cuadrado con un nivel de significancia de 0,05 si los datos pueden considerarse como una variable aleatoria que tiene distribución de Poisson con λ=7 arribos en un segundo.
Solución:
1. Hipótesis nula: La variable aleatoria tiene una distribución de Poisson con λ=7.
Hipótesis alterna: La variable aleatoria no tiene distribución de Poisson con λ=7.
2. Nivel de significancia: α=0,05
3. Criterio: Se rechaza la hipótesis nula si la ChiCuadrado > 22,362, el valor de ChiCuadrado con significancia de 0,05 para k-m-1= 15-1-1=13 grados de libertad. El número de grados de libertad es 13, dado que sólo una cantidad, la frecuencia total de los 1.778, es necesaria en los datos observados para calcular las frecuencias esperadas.
K es el número de términos en la fórmula con que se calcula ChiCuadrado y m es el número de cantidades conseguidas de los datos observados, que no se necesitan para calcular las frecuencias observadas.
4. Cálculos: Estadístico para la prueba de bondad de ajuste:

$$X^2 = \sum_{i=1}^{k} \frac{(Obs_i - Esp_i)^2}{\quad} \qquad \text{(Ec. 5)}$$

Esp_i

Sustituyendo en la fórmula para X^2 obtenemos:

$X^2 = 20591,44$

5. Decisión: Dado que X^2=20591,44 sobrepasa demasiado el valor de 22,36, la hipótesis nula se rechaza; concluimos que con un nivel de significancia del 5%, los datos muestran evidencia para afirmar que la distribución de Poisson con λ=7 no proporciona un buen ajuste.
Puede consultar en los trabajos de Miller (1986), Ludwig y Reynolds (1988) y Mood (1974) sobre el procedimiento para el cálculo del Chi-cuadrado.

Considerando la relación

I = varianza/promedio

Si:

a) I > 1 la variable estudiada presenta indicio de conglomeración, esto implicaría que la distribución binomial negativa daría un mejor ajuste.

b) I ≈ 1 implica que es aleatorio y por lo tanto la distribución de Poisson daría un mejor ajuste.

c) I < 1 implica que la variable es uniforme y que la distribución binomial positiva tendría un mejor ajuste.

Por lo tanto y teniendo los siguientes datos de la variable arribos por segundo:
Promedio: 7 (eliminando las colas)
Varianza: 94
Se obtiene que I = 13 que es mucho mayor que 1

Utilizando el promedio de la muestra completa tenemos:
Promedio=25
Varianza=10673
Se obtiene que I=427

En ambos casos, se obtiene un I mucho mayor que 1, por lo tanto y de acuerdo con lo expuesto, tenemos que la distribución binomial negativa daría un mejor ajuste a esta variable. Un estudio más detallado de esta teoría puede ser encontrado en el trabajo de (Mood, 1974).

A continuación se realizan los cálculos para conocer si la distribución binomial negativa da un mejor ajuste a la variable estudiada.

Prueba de bondad para ajustar a la Binomial Negativa

Se llevó a cabo una prueba de bondad de ajuste de Chi-cuadrado con un nivel de significancia de 0,05 si los datos pueden considerarse como una variable aleatoria que tiene distribución Binomial Negativa.

Solución:

1. Hipótesis nula: La variable aleatoria tiene una distribución Binomial Negativa.
   Hipótesis alterna: La variable aleatoria no tiene distribución Binomial Negativa.
2. Nivel de significancia: α=0,05
3. Criterio: Se rechaza la hipótesis nula si ChiCuadrado > 37,652, el valor de ChiCuadrado con significancia de 0,05 para 25 grados de libertad.
4. Cálculos: Estadístico para la prueba de bondad de ajuste:

$$X^2 = \sum_{i=1}^{k} \frac{(Obs_i - Esp_i)^2}{Esp_i} \qquad \text{(Ec. 5)}$$

Sustituyendo en la fórmula para X^2 obtenemos:

$X^2 = 207,96$

5. Decisión: Dado que X^2=207,96 sobrepasa demasiado el valor de 37,65, la hipótesis nula se rechaza. Concluimos que con un nivel de significancia del 5% los datos no muestran evidencia para afirmar que la distribución Binomial Negativa proporciona un buen ajuste. Sin embargo su ajuste es mucho mejor que la de la distribución de Poisson, debido a que el valor de la Chi-cuadrado es mucho menor que la Chi-cuadrado para Poisson.

Para el cálculo de Chi-cuadrado se utilizaron programas de Ludwig y Reynolds (1988).

Conclusión sobre las distribuciones de probabilidad para la variable número de arribos por segundo.

Utilizando el estadístico de Chi-cuadrado para la prueba de bondad de ajuste, se llegó a probar que la variable de número de arribos que llegan en un intervalo de tiempo de un segundo no sigue una distribución de Poisson, como se supone en la mayoría de los análisis teóricos. Además, de acuerdo con el índice definido I, se obtiene que la variable estudiada es mejor ajustada por la binomial negativa, y la prueba de bondad para dicha distribución demuestra que efectivamente da un mejor ajuste que la Poisson, debido a que el valor de la Chi-cuadrado para la Binomial Negativa es mucho menor que la Chi-cuadrado para Poisson.

Haciendo una revisión de la teoría de procesos Poisson y del tipo de datos se obtiene lo siguiente:

1) Se está trabajando con arribos de paquetes, NO de mensajes.
2) Los arribos de paquetes no son independientes, puesto que los arribos de los paquetes que forman el mismo mensajes no son independientes entre sí.
3) Para Poisson está definido que el promedio y la varianza son iguales y los datos de la variable estudiada da un promedio muy diferente a la varianza.

Por lo tanto, de acuerdo con esto y a las pruebas de bondad de ajuste realizadas, se concluye que la variable número de arribos que llegan en el intervalo de un segundo no sigue una distribución de Poisson, y por lo tanto, no se puede decir nada sobre la distribución de los tiempos entre arribos.

Esto coincide con lo que se indica en el trabajo de Law y McComas (1994). Ellos mencionan que en estudios recientes de medidas de tráfico de alta resolución y de alta calidad, que incluye un análisis de cientos de millones de paquetes observados en una red de área local "Ethernet", el tráfico de paquetes no sigue una distribución de Poisson.

Si se llegaba a concluir que los arribos siguen una distribución Poisson, fácilmente se podía determinar que los tiempos entre arribos siguen un distribución Exponencial Negativa; sin embargo no ocurrió así, por lo tanto, a continuación se presenta el análisis para encontrar la función de distribución que mejor se ajusta a la variable tiempo entre arribos de paquetes.

Definición de la distribución de probabilidad para la variable Tiempo entre arribos

Se hizo una serie de pruebas antes de llegar a escoger la función de distribución de probabilidad Gamma para la variable tiempo entre arribos. De acuerdo con Grodzinsky (1999), las distribuciones ampliamente usadas para el modelaje de tráfico incluyen las distribuciones de la exponencial, gamma, normal, log-normal, Weibull, beta, Person V y Person VI.

Para la búsqueda de la distribución se utilizó el "software" Systat, un programa estadístico (Wilkinson, 1990). Utilizando el Systat, se realizó una serie de ploteos de probabilidad, en los cuales se compara una variable cuantitativa con una distribución de probabilidad conocida, graficando sus valores contra los puntos de porcentaje correspondientes de esa distribución. Si los datos siguen esa distribución, los puntos siguen una línea recta diagonal.

En la figura 2 se hace un ploteo de probabilidad para comparar la variable tiempo entre arribos con la distribución de probabilidad Gamma con r = 0,03 y 1/λ = 1, r y 1/λ son los parámetros de esta distribución.

En la figura 3 se compara la variable tiempo entre arribos con la distribución de probabilidad Exponencial negativa.

Véase en las figuras 2 y 3 que la distribución de probabilidad Gamma se ajusta mejor a la variable tiempo entre arribos que la Exponencial negativa, puesto que la mayoría de los puntos del gráfico de la figura 7 cae muy cerca de la línea recta diagonal; en cambio, en la Exponencial negativa, los puntos caen muy distantes de la línea recta.

Figura 2. Tiempo entre arribos vs distribución Gamma (1, 0,03).

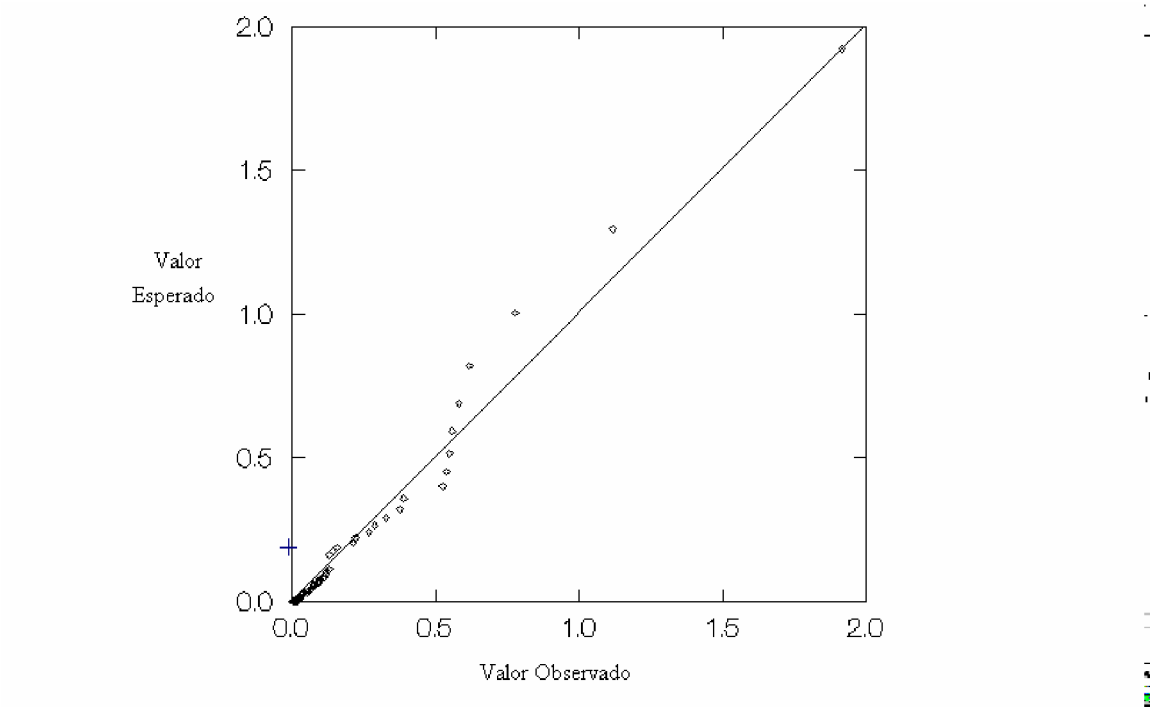
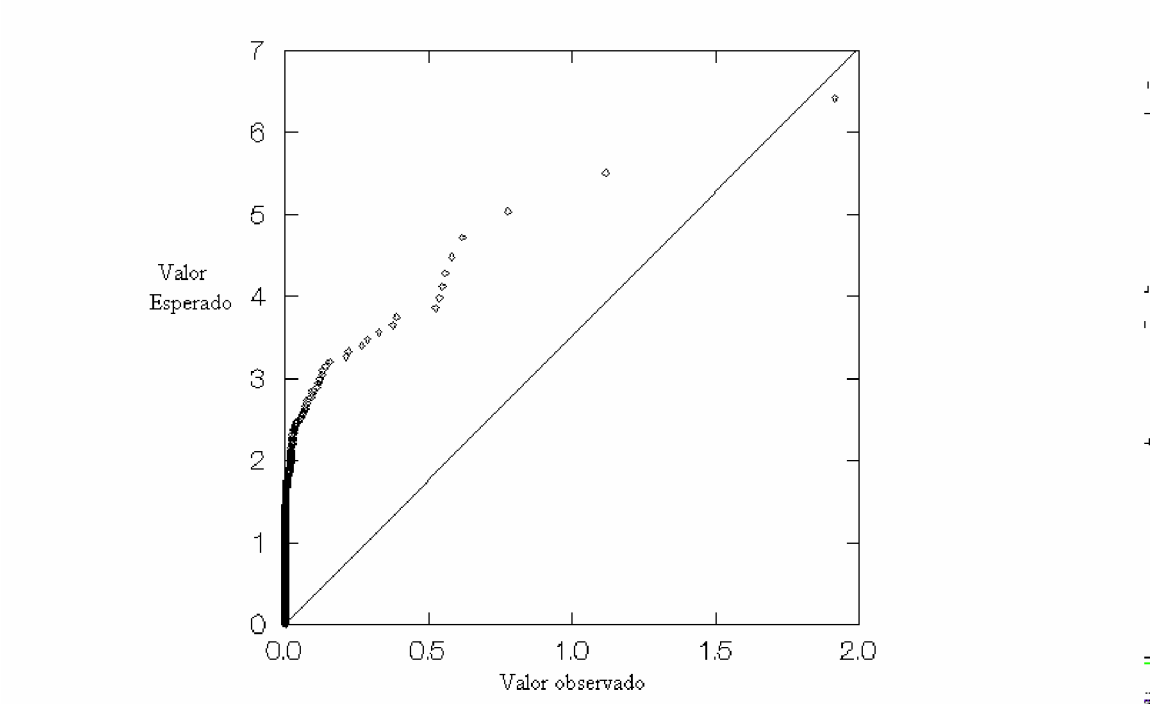


Figura 3. Tiempo entre arribos vs distribución Exponencial.



Se estudiaron varias otras distribuciones de probabilidad, aquí solo se muestran los gráficos correspondientes a la Gamma y la Exponencial. De todos los gráficos que se hicieron, la distribución Gamma con parámetros r=0,03 y $1/\lambda=1$ fue la que mejor ajustó la variable de tiempo entre arribos.

Conclusión

El análisis realizado se hizo con base a los datos colectados, no se descarta que con otros datos pueda dar resultados diferentes, esto debido a que el tráfico de una red es altamente variable lo que hace muy difícil hacer una caracterización del mismo. Por lo tanto, la información acerca de la función de distribución de la variable de tiempo entre arribos de paquetes obtenida debe ser tomada solo como una referencia para cualquier otro estudio. Lo importante de este trabajo radica en la descripción detallada que se hace del proceso que se siguió para la obtención de dicha función de distribución, y que puede ser de gran ayuda y guía para los iniciadores en el estudio y análisis del tráfico de una red de área local.

Bibliografía

Grodzinsky, Frances S. **Networking and Data Communications Laboratory Manual.** Prentice Hall: USA. 1999.

Law, Averill; McComas, Michael. **Simulation "software" for communications, Networks: The state of the art**. IEEE Communications Magazine. March 1994.

Ludwig, John A.; Reynolds, James F. **Statistical Ecology: a Primer on Methods and Computing.** John Wiley & Sons: USA. 1988.

Miller, Irwin; Freund, John E**. Probabilidad y Estadística para Ingenieros.** 3 ed. Prentice Hall Hispanoamericana.. México. 1986.

Mood, Alexander; Graybill, Franklin; Boes, Duane. **Introduction to the Theory of Statistics**. 3 ed. McGraw Hill: USA. 1974

Triticom. **LANdecoder Protocol Analyzers**. Triticom. USA. 1997.

Wilkinson, L. **Systat: The System for Statistic.** Evanston, Il, Systat Inc. 1990.