Modelos de sesiones en redes de datos

Luis López-Oliveros
ll278@cornell.edu

Junto con Sid Resnick - Cornell University

14 de marzo de 2014

Modelos de redes de datos

La conmutación de paquetes es un método de transmisión de datos que consiste en

1. Dividir un archivo en paquetes de tamaño acotado.
2. Etiquetar cada paquete con una cabecera: Masa, remitente y destinatario.
3. Guardar y retransmitir los paquetes por medio de enrutadores.
4. Reensamblar el archivo original cuando los paquetes llegan a su destino.

Nuestros datos consisten en cabeceras de paquetes. Nuestros objetivos son modelar la actividad de los usuarios finales y estimar características del tráfico de redes usando sesiones de datos.

Una sesión es un objeto que se obtiene al agrupar paquetes utilizando *cierta* regla – que puede no ser única, e.g.

- Sesión de punta a punta: (remitente, destinatario).
- Conexión: (remitente, destinatario, puerto de origen, puerto de destino, protocolo de red).

Problemas de interés en el estudio de redes

- Superposición de tráfico heterogéneo extendiendo la escala temporal.
  - Sobreponer sesiones,

    $A[0,t] = \text{tráfico de sesiones acumulado en [0,t] a través de un nodo crítico.}$
  - ¿Límite de $A[0, Tt]$ cuando $T \to \infty$? Gaussiano en la práctica, pero resultados teóricos dictan que podría ser Lévy estable.
  - Garantizar la calidad del servicio y ofrecer anchos de banda adecuados.

- Análisis estadístico de sesiones de redes.
  - Propiedades de la distribución de características clave de sesiones.
  - Simulación de sesiones y del comportamiento de usuarios finales.
  - Varios métodos de agrupación de sesiones:
    - Por velocidad máxima de transmisión.
    - Por aplicación.

Resumen de la plática

1. Breve repaso de teoría de valores extremos.

2. Tráfico recolectado en Cornell. ¿Es Gaussiano o estable?

3. Modelos para la superposición de tráfico de redes.
   - Tráfico homogéneo: Una sola corriente.
   - Tráfico heterogéneo: Multicorriente.

4. Teorema sobre tráfico heterogéneo.

5. Explicación de los resultados sobre los datos de Cornell.

Teoría de valores extremos

- Cola pesada: La cola de la distribución decae como una función de potencia. Eventos extremos tienen una probabilidad grande (relativamente a una distribución Gaussiana).

- La cola de la distribución de una variable aleatoria $X \geq 0$ es pesada si satisface
$$\bar{F}_X(x) := 1 - F(x) = x^{-\alpha} L(x),$$
donde L es tal que
$$\lim_{x \to \infty} L(tx)/L(x) = 1, \quad t > 0,$$
y $\alpha > 0$ se conoce como índice de cola.

- Mientras más chico es el índice de cola, más pesada es ésta.

- E.g. si $1 < \alpha < 2$, entonces $EX < \infty$ pero $Var(X) = \infty$.

Ejemplos de distribuciones con cola pesada

- t de Student: Índice de cola = grados de libertad.
- Cauchy: Índice de cola = 1.
- α-Lévy estables: Índice de cola = α.
- Pareto: $\bar{F}(x) = (k/x)^{\alpha}, x \geq k$. Índice de cola = α.

Tráfico de Cornell

- 55 días, 4 horas diarias, 1-5pm, sesiones de TCP y UDP.
- Corriente TCP = web, correo electrónico, ftp, ssh, punto-a-punto, etc. Velocidad de transmisión del remitente es restringida por medio de un sistema de acuse de recibos, donde si no se tiene acuse de recibo dentro de cierto tiempo, la velocidad se reduce a la mitad.
- Corriente UDP = tipo streaming, intercambio de archivos, VozIP. No existe ningún control; remitente sólo transmite. Sesiones llegan a ser muy largas.
- Tráfico de Cornell: TCP $\approx$ 90% de bytes y 80% de sesiones.
- $A_k^{(TCP)}$ y $A_k^{(UDP)}$ $:=$ total de bytes TCP y UDP capturados en la hora k-ésima, $k = 1, \ldots, 220$.
- Remoción de temporalidad diaria y de la tendencia, pero mismas conclusiones sin esta manipulación.

Tráfico acumulado clasificado por protocolo de Internet

Fig. : *Izq*: Corriente TCP (p-valor Anderson-Darling: 0.1319). *Der*: Corriente UDP (p-valor Anderson-Darling: 9.8×10^{-16})

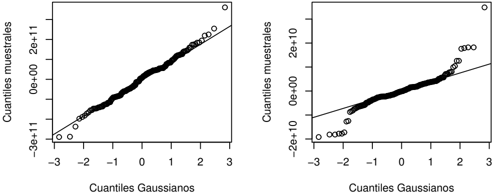
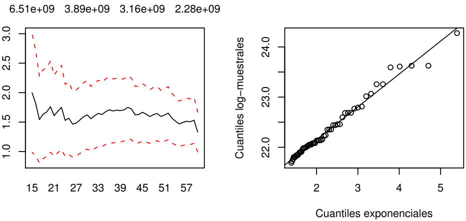
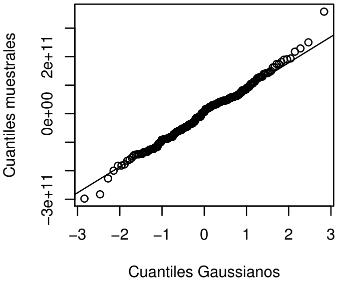


Fig. : Gráficas de corriente UDP. *Izq:* Gráfica de Hill del índice de cola. *Der:* Gráfica QQ exponencial de los log-datos.



Superposición del tráfico

Fig. : Gráfica QQ normal de $A^{(TCP)} + A^{(UDP)}$= tráfico acumulado multicorriente. (p-valor Anderson-Darling: 0.2117)



Fenómeno que requiere de explicación:

- $A^{(TCP)}$ se ve Gaussiano.
- $A^{(UDP)}$ se ve de cola pesada.
- $A^{(TCP)} + A^{(UDP)}$ se ve Gaussiano!

El resultado es anti-intuitivo debido a la naturaleza de la cola de las variables.

Deberían existir parámetros de localización y escala tales que:

- $\frac{A^{(TCP)} - \mu^{(TCP)}}{a^{(TCP)}}$ se ve Gaussiano.
- $\frac{A^{(UDP)} - \mu^{(UDP)}}{a^{(UDP)}}$ se ve de cola pesada.
- El tráfico multicorriente se ve Gaussiano bajo una normalización apropiada.

Descripción del problema y suposiciones del modelo

El tráfico de redes está constituído por p distintas *corrientes* (e.g. web, correo electrónico, tipo streaming, etc.)

Las sesiones de la corriente j-ésima son activadas por un proceso de Poisson homogéneo en $\mathbb{R}$ con puntos $\{\Gamma_k^{(j)}; -\infty < k < \infty\}$ e intensidad $\lambda^{(j)}$. Estos procesos son independientes.

Por lo tanto, las sesiones del tráfico multicorriente son activadas por un proceso de Poisson homogéneo con intensidad $\lambda = \sum_{j=1}^{p} \lambda^{(j)}$.

El tiempo $\Gamma_k^{(j)}$ de activación de la transmisión la sesión k-ésima de la corriente j-ésima tiene asociadas dos cantidades adicionales $(D_k^{(j)}, V_k^{(j)})$, donde

- $D_k^{(j)} \sim \mathrm{iid}\ F_D^{(j)}$, representa la duración de la sesión.
- $V_k^{(j)} \sim \mathrm{iid}\ F_V$, representa la velocidad de transmisión.
- $(D_k^{(j)}, V_k^{(j)})_{j,k}$ son independientes y $D_k^{(j)} \perp\!\!\!\perp V_k^{(j)}$.

Medida de conteo de las sesiones

- La medida de conteo del proceso de Poisson aumentado $(\Gamma_k^{(j)}, D_k^{(j)}, V_k^{(j)})$ por los descriptores de las sesiones de la corriente j-ésima en $\mathbb{R} \times [0,\infty)^2$ es

$$N^{(j)} := \sum_k \epsilon_{(\Gamma_k^{(j)}, D_k^{(j)}, V_k^{(j)})} = \mathbb{POISSON}(\lambda^{(j)} ds F_D^{(j)}(dd) F_V(dv)), \quad j = 1, \ldots, p.$$

- Por independencia, la medida de conteo de los descriptores de las sesiones del tráfico agregado es

$$N := \sum_{j=1}^{p} N^{(j)} = \mathbb{POISSON}\left(\lambda ds \sum_{j=1}^{p} (\lambda^{(j)}/\lambda) F_D^{(j)}(dd) F_V(dv)\right)$$
$$= \mathbb{POISSON}(\lambda ds F_D(dd) F_V(dv)).$$

Tráfico acumulado

Notación: (s, d, v) es una sesión que comienza en tiempo s, con duración igual a d y con velocidad v de transmisión.

$$\begin{aligned} A^{(j)}(t) &:= \text{tráfico la } j\text{-ésima corriente acumulado en } [0, t] \\ &= \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{\infty} v|[0, t] \cap [s, s+d]| N^{(j)}(ds, dd, dv), \end{aligned}$$

donde

$$N^{(j)} = \text{medida de conteo de las sesiones de la } j\text{-ésima corriente.}$$

$$\begin{aligned} A(t) &:= \text{tráfico multicorriente acumulado en } [0, t] \\ &= \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{\infty} v|[0, t] \cap [s, s+d]| N(ds, dd, dv), \end{aligned}$$

donde

$$N = \text{medida de conteo de las sesiones del tráfico multicorriente.}$$

Regímenes de crecimiento de la intensidad temporal

- Supongamos que $F_D^{(j)}$ es de cola pesada, $1 < \alpha_D^{(j)} < 2$, y que $E(V_1^{(1)})^2 < \infty$.

- Normalicemos el tráfico acumulado por T y si es necesario permitamos que $\lambda^{(j)} := \lambda^{(j)}(T) \to \infty$ (¿bajo que condiciones?) y consideremos el T-ésimo modelo:
$$A_T(t) = A(Tt) = A[0, Tt].$$

- Familia de modelos indexados por T. Al movernos a través de la familia cuando T a ∞, ¿existe un límite informativo?

- Regímenes de crecimiento de la intensidad temporal:
  - Crecimiento rápido: $\lambda^{(j)} T \bar{F}_D^{(j)}(T) \to \infty$.
  - Crecimiento lento: $\lambda^{(j)} T \bar{F}_D^{(j)}(T) \to 0$.

Resultados previos para una sola corriente

Para la corriente $j-$ésima, $j = 1, \ldots, p$:

Teorema (Mikosch et al. 2002; Kaj y Taqqu, 2008)

El proceso $(A^{(j)}(Tt), t \geq 0)$ *que describe el tráfico acumulado en* $[0, Tt]$ *satisface los siguientes límites*

$$\frac{A^{(j)}(T\cdot) - \lambda^{(j)}\mu_D^{(j)}T\cdot}{a^{(j)}_{rápido}(T)} \xrightarrow{\text{difi}} B_{H^{(j)}}(\cdot), \qquad \textit{crecimiento rápido,}$$

$$\frac{A^{(j)}(T\cdot) - \lambda^{(j)}\mu_D^{(j)}T\cdot}{a^{(j)}_{lento}(T)} \xrightarrow{\text{difi}} \Lambda_{\alpha_D^{(j)}}(\cdot), \qquad \textit{crecimiento lento,}$$

donde

- $B_{H^{(j)}}$ *es un movimiento Browniano fraccional estándar y* $H^{(j)} = (3 - \alpha_D^{(j)})/2 \in (1/2, 1)$.
- $\Lambda_{\alpha_D^{(j)}}$ *es un movimiento* $\alpha_D^{(j)}-$*Lévy estable.*

¿Tráfico de redes en su hábitat natural?

- Pregunta: ¿Alguien ha avistado tráfico del tipo del movimiento Lévy estable?
- Answer: No. (Guerin et al. 2003).
- Sid Resnick: “Nunca he conocido a un especialista de redes que crea que el tráfico pudiera ser estable y en conferencias de redes me he topado con muchas reacciones hostiles a la más mínima insinuación”.
- Teóricamente el tráfico acumulado podría ser estable. Entonces, ¿cómo se explica la falta de avistamientos?
- ¿Qué pasaría si combinamos diferentes corrientes, algunas de las cuales pudieran ser Gaussianas, y otras Lévy estables?

Regímenes de escalas temporales para el tráfico agregado

Medida de conteo de sesiones:

- $N^{(j)}(ds, dd, dv) = \mathbb{POISSON}(\lambda^{(j)} ds F_D^{(j)}(dd) F_V(dv))$.
- $N(ds, dd, dv) = \mathbb{POISSON}(\lambda ds F_D(dd) F_V(dv))$.
- $\lambda^{(j)} := \lambda^{(j)}(T) \to \infty$.

Tenemos que

- Intensidad de llegada de la multicorriente: $\lambda = \sum_{j=1}^{p} \lambda^{(j)}$.
- $F_D := \sum_{j=1}^{p} (\lambda^{(j)}/\lambda) F_D^{(j)}$, $\bar{F}_D$ tiene índice de cola $\alpha_D := \min_{1 \le j \le p} \alpha_D^{(j)}$.
- Proporción de sesiones que pertenecen a la corriente j-ésima: $\lambda^{(j)}/\lambda$.

Regímenes de crecimiento de la intensidad temporal cuando $T \to \infty$:

- Escenario $\mathcal{R}$: Al menos una corriente satisface la condición de crecimiento rápido, por lo que

$$\lambda T \bar{F}_D(T) = \sum_{j=1}^{p} \lambda^{(j)} T \bar{F}_D^{(j)}(T) \to \infty.$$

- Escenario $\mathcal{L}$: Todas las corrientes satisfacen la condición de crecimiento lento, por lo que

$$\lambda T \bar{F}_D(T) \to 0.$$

Teorema (López-Oliveros & Resnick, 2010)

Si suponemos que

$$\liminf_{T\to\infty} \max_{j:\alpha_D^{(j)}=\alpha_D} \lambda^{(j)}/\lambda > 0,$$

el proceso $(A(Tt), t \geq 0)$ *que describe el tráfico multicorriente acumulado en* $[0, Tt]$ *satisface*

$$\frac{A(T\cdot) - \lambda\mu_D T\cdot}{a_{\mathcal{R}}(T)} \xrightarrow{difi} B_H(\cdot), \qquad \text{Escenario } \mathcal{R},$$

$$\frac{A(T\cdot) - \lambda\mu_D T\cdot}{a_{\mathcal{L}}(T)} \xrightarrow{difi} \Lambda_{\alpha_D}(\cdot), \qquad \text{Escenario } \mathcal{L}.$$

Observaciones:

- En la práctica, el tráfico web es una corriente omnipresente con tasas de iniciación rápida, i.e. régimen rápido y límite de movimiento Browniano fraccional.
- van De Meent et al. (2006): ¿Tráfico Gaussiano omnipresente? Cuando juntamos varias corrientes, algunas de las cuales son estables y otras movimiento Browniano fraccional, el tráfico multicorriente sigue un movimiento Browniano fraccional.

Idea de la demostración

- El tráfico acumulado normalizado es una integral con respecto a la medida de Poisson aleatoria normalizada, para la cual podemos calcular la función característica de los procesos de dimension finita (*fc difi*).

- Demostramos la convergencia de fc difi . El límite es la fc difi de un movimiento Browniano fraccional en el escenario $\mathcal{R}$ y la fc difi de un proceso estable en el escenario $\mathcal{L}$.

- La medida de conteo de las sesiones de la corriente j-ésima tiene la forma $N^{(j)}(ds, dl, dr) = \mathbb{POISSON}(\lambda^{(j)} ds F_D^{(j)}(dd) F_V(dv))$, donde
  - $\lim_{T\to\infty} \bar{F}_D^{(j)}(Td)/\bar{F}_D^{(j)}(d) = d^{-\alpha_D^{(j)}}, \quad d > 0.$ (1)
  - Cotas de Potter para $\bar{F}_D^{(j)}, \quad j = 1, \dots, p.$ (2)

- La medida de conteo de las sesiones del tráfico multicorriente es $N(ds, dd, dv) = \mathbb{POISSON}(\lambda ds F_D(dd) F_V(dv))$, y podemos demostrar propiedades simulares a (1) y (2) para

$$F_D := \sum_{j=1}^{p} (\lambda^{(j)}/\lambda) F_D^{(j)}.$$

Gracias

¿Preguntas?