Marginales non gaussiennes et longue mémoire : analyse et synthèse de trafic Internet

Antoine SCHERRER[1], Patrice ABRY[2]

[1] Laboratoire de l'Informatique du Parallélisme (UMR 7623, CNRS)
Ecole Normale Supérieure de Lyon, 46, Allée d'Italie 69364 LYON CEDEX 07 - France

[2]Laboratoire de Physique (UMR 5672, CNRS)
Ecole Normale Supérieure de Lyon, 46, Allée d'Italie 69364 LYON CEDEX 07 - France
antoine.scherrer@ens-lyon.fr,patrice.abry@ens-lyon.fr

Résumé – La modélisation statistique du trafic Internet constitue actuellement un outil incontournable pour le dimensionnement des réseaux, la prévision de leurs comportements et performances, permettant d'assurer la disponibilité, la qualité de service (QoS) ainsi que la sécurité des réseaux. Nous proposons de modéliser les séries temporelles de trafic par un processus non gaussiens à longue mémoire. Nous montrons que ce modèle reste pertinent pour un très large continuum de niveaux d'agrégation et pour une large variété de trafic. Nous décrivons les procédures permettant d'estimer les paramètres de ce modèle ainsi que celles permettant de synthétiser numériquement des réalisations de processus dont marginales et covariances sont prescrites.

Abstract – Internet traffic statistical modeling has now become a major tool used for network design, performance and behaviour prediction and hence to ensure disponibility, quality of service (QoS) as well as security to all its end-users. We propos to model Internet time series with a non Gaussian long range dependent process. We show that this model remains relevant over a large range of aggregation levels and for a wide variety of different traffics. We describe the procedures used to estimate the corresponding parameters as well as those enabling to numerically synthetize realisations of such processes with prescribed marginals and covariances.

1 Motivation

La modélisation statistique des séries de télétrafic informatique, du trafic Internet par exemple, constitue désormais un exercice obligatoire de la gestion des réseaux. Celle-ci, en effet, se révèle indispensable pour réaliser des prévisions de performances pertinentes, améliorer le fonctionnement du réseau, y assurer une certaine qualité de service (QoS), en optimiser la gestion ou décider de règles de tarifications. Le trafic Internet présente deux caractéristiques statistiques principales, unanimement reconnues, et qu'il est essentiel de prendre en compte pour réaliser des modélisations efficaces : il est **non gaussien**[1] et à **longue mémoire**[2, 3, 4, 5]. Souvent, les modélisations du trafic s'intéressent à l'une ou l'autre de ces caractéristiques mais n'essaient pas d'atteindre une description simultanément pertinente des deux. Souvent, les travaux qui visent ce double objectif reposent sur une superposition de processus de Poisson modulés par une chaîne de Markov [6], impliquant un grand nombre de paramètres à estimer afin d'ajuster et la distribution marginale et la structure de longue mémoire.

Dans ce travail, nous nous intéressons à une modélisation conjointe de la distribution marginale et de la structure de covariance de séries de télétrafic informatique, qui capture en peu de paramètres à la fois leur caractère non gaussien et leur longue mémoire. Nous proposons l'utilisation d'un processus stochastique non gaussien à longue mémoire, dont la distribution marginale est une loi Gamma, $\Gamma_{\alpha,\beta}$, et dont la structure de corrélation est celle d'un processus FARIMA(ϕ, d, θ). A partir de plusieurs jeux de traces *célèbres* de télétrafic informatique, nous montrons comment ce modèle à 5 paramètres, $\alpha, \beta, d, \phi, \theta$ capture efficacement leurs caractéristiques statistiques de premier et second ordres. Nous observons notamment que cette modélisation reste pertinente pour une très large gamme de niveaux d'agrégation Δ. Nous proposons ensuite une procédure analytique permettant de synthétiser numériquement des réalisations de processus possédant conjointement les marginale et covariance choisies a priori pour reproduire celles du trafic réel. Disposer de telles procédures constitue un enjeu essentiel. D'une part, elles permettent d'étudier par simulations numériques les performances de files d'attentes et de réseaux qui ne pourraient être atteintes par calcul analytique du fait des propriétés statistiques non standards du trafic correspondant. D'autre part, elles fournissent des simulateurs de trafic utilisés pour nourrir des maquettes de réseau afin d'étudier la qualité de leur fonctionnement et la QoS produite.

2 Trafic Internet

Nous avons travaillé à partir d'une variété de traces de trafic informatique collectées entre 1989 et 2003, correspondant à des types de trafic et de réseau différents (Local, Metropolitan, Wide Area Network,..., périphérie ou cœur de réseau,...). Ces traces sont disponibles sur les sites des principaux groupes de recherches universitaires impliqués (WAND, Auckland, Nelle-Zélande, CAIDA, LBL s, UNC, Etats-Unis, ...), le tableau 1 les présente en détails.
A partir de ces traces, sont extraites des séries temporelles cor-

Trace	Date de départ	Durée (s)	Liaison	# Pkts (10^6)	IAT (ms)	Lien
PAUG	1989-08-29(11 :25)	2620	LAN(10BaseT)	1	2.6	ita.ee.lbl.gov/index.html
LBL-TCP-3	1994-01-20(14 :10)	7200	WAN(10BaseT)	1.7	4	ita.ee.lbl.gov/index.html
AuckIV	2001-04-02(13 :00)	10800	WAN(OC3)	9	1.2	wand.cs.waikato.ac.nz/wand/wits
CAIDA	2002-08-14(10 :00)	600	Backbone(OC48)	65	0.01	www.caida.org/analysis/workload/oc48/
UNC	2003-04-06(16 :00)	3600	WAN(xxx)	4.6	0.8	www-dirt.cs.unc.edu/ts/

TAB. 1 – Description des traces

respondant par exemple à la suite des nombres de paquets comptés (ou agrégés) dans des boîtes temporelles successives de durée Δ, notée $X_\Delta(k)$. Un travail équivalent peut être conduit pour la modélisation des séries de nombre d'octets agrégés ou d'inter-arrivée des paquets.
Une question centrale dans la modélisation du trafic Internet réside dans le choix d'un niveau d'agrégation Δ pertinent. La solution de cette délicate question dépend à la fois de l'utilisation qui sera faite de la modélisation, de la nature des données et de contraintes techniques. Il est donc essentiel de proposer des modèles qui incorporent naturellement et facilement la possibilité de travailler à différents niveaux d'agrégation.

3 Modélisation : Processus non gaussiens à mémoire longue

Pour modéliser les séries $\{X_\Delta(k), k \in \mathbb{Z}\}$, nous proposons l'utilisation d'un processus stochastique stationnaire de marginale $\Gamma_{\alpha,\beta}$ et de covariance FARIMA(ϕ, d, θ), et ce pour chaque niveau d'agrégation indépendamment.
• **Marginale non gaussienne.** – La distribution *gamma*, $\Gamma_{\alpha,\beta}$, $\Gamma_{\alpha,\beta}(x) = \frac{1}{\beta\Gamma(\alpha)}(\frac{x}{\beta})^{\alpha-1}\exp(-\frac{x}{\beta})$, ($\Gamma$ étant la fonction Gamma standard[7]), est caractérisée par deux paramètres strictement positifs : la forme (α), et l'échelle (β). Elle fournit des variables aléatoires positives, de moyenne $\mu = \alpha\beta$ et de variance $\sigma^2 = \alpha\beta^2$ et possède la propriété intéressante d'être stable sous addition et par multiplication par une constante. Enfin l'inverse du paramètre de forme, $1/\alpha$ peut être envisagé comme un indicateur de distance à la loi Normale de même moyenne et variance.
• **Longue mémoire.** – Comme cela est désormais largement admis, le trafic Internet possède une propriété de longue mémoire, mais il possède aussi des dépendances à court terme, dont la structure dépend des méchanismes réseaux mis en œuvre. C'est pourquoi il est naturel d'utiliser un modèle de covariance pouvant rendre compte de ces deux structures : le modèle FARIMA (Fractionnaly Integrated Auto-regressive Moving Average).
Un processus FARIMA est défini par deux polynômes Φ et Θ d'ordre respectif P et Q et une intégration fractionnaire (d'ordre $d \in (-1/2, 1/2)$) :
$X_l = \sum_{p=1}^{P}\phi_p X_{l-p} + \triangle^{-d}(\epsilon_l - \sum_{q=1}^{Q}\theta_q\epsilon_{l-q})$, où ϵ_l sont des variables aléatoires indépendantes, identiquement distribuées de moyenne nulle et de variance σ_ϵ^2 et où $\triangle^{-d}$ est défini par son développement en séries entières :
$\triangle^{-d} = \sum_{i=0}^{\infty} b_i(-d)B^i$, B étant l'opérateur de retard $B\epsilon_i = \epsilon_{i-1}$, et $b_i(-d) = \Gamma(i+d)/\Gamma(d)\Gamma(i+1)$, $i = 1, 2, \ldots$, Γ étant la fonction Gamma. Le spectre de X prend la forme :

$$f_X(\nu) = \sigma_\epsilon^2 |1 - e^{-i2\pi\nu}|^{-2d} \frac{|1 - \sum_{q=1}^{Q}\theta_q e^{-iq2\pi\nu}|^2}{|1 - \sum_{p=1}^{P}\phi_p e^{-ip2\pi\nu}|^2}, \quad (1)$$

$-1/2 < \nu < 1/2$. Pour $d \in (0, 1/2)$, ce processus possède une propriété de longue mémoire [8]. Les polynômes P et Q, d'une part, l'ordre d'intégration d, d'autre part rendent respectivement compte des corrélations statistiques à court et long termes. Dans la plupart des cas, nous nous limiterons à des polynômes P et Q d'ordre 1, le nombre de paramètres pour la covariance est alors réduit à 3 : ϕ, d, θ.

4 Analyse

Nous détaillons maintenant les procédures d'analyse et d'estimation des paramètres du modèle proposé, utilisées pour l'étude des séries X_Δ, pour chaque Δ indépendamment.
• **Paramètres de la loi *Gamma*.** – Les paramètres α et β sont estimés par une procédure standard correspondant à un estimateur à maximum de vraisemblance pour des données i.i.d.. L'initialisation est réalisée à partir des estimateurs standards de moyenne et variance $\hat{\mu}$ et $\hat{\sigma^2}$, selon $\hat{\beta} = \hat{\sigma^2}/\hat{\mu}$ et $\hat{\alpha} = \hat{\mu}/\hat{\beta}$.
• **Paramètre de la covariance.** – L'estimation des paramètres de la covariance du FARIMA(1,d,1) est réalisée en deux étapes. Nous procédons en d'abord à une analyse en ondelettes de la série X_Δ, afin d'estimer le paramètre de longue mémoire d. Soient $\psi_{j,k}(t) = 2^{-j/2}\psi_0(2^{-j}t - k)$ les dilatées et translatées sur la grille dyadique d'une ondelette mère de référence ψ_0. On note $d_X(j,k) = \langle\psi_{j,k}, X_\Delta\rangle$ les coefficients d'ondelettes. Les coefficients d'ondelettes d'un processus stochastique stationnaire de spectre f_X satisfont :

$$\mathbb{E}d_X(j,k)^2 = \int f_X(\nu)2^j|\Psi_0(2^j\nu)|^2 d\nu, \quad (2)$$

où Ψ_0 rend compte de la transformée de Fourier de ψ_0 et $\mathbb{E}$ l'espérance mathématique.
Le spectre d'un processus à longue mémoire se comporte en une loi de puissance à l'origine : $S_X(\nu) \simeq_{|\nu|\to 0} C|\nu|^{-2d}$. On peut alors montrer que ces coefficients d'ondelettes se comportent en :

$$\mathbb{E}d_X(j,k)^2 = C2^{j(2d)}, \quad (3)$$

La moyenne temporelle $S_j = 1/n_j \sum_{k=1}^{n_j}|d_X(j,k)|^2$ estime la moyenne d'ensemble $\mathbb{E}d_X(j,k)^2$. On trace ensuite le diagramme log-échelle, $\log_2 S_j$ en fonction de $\log_2 2^j = j$, dans lequel la longue mémoire se matérialise par l'apparition d'un segment de droite dans la limite des grandes échelles (j grand). Une régression linéaire pondérée permet d'en estimer la pente donc d. Cette estimation ondelette du paramètre de longue mémoire est robuste et efficace [8]. Les données sont ensuite *quasiment blanchies* par intégration fractionnaire d'ordre $-\hat{d}$, cela a pour effet de gommer la longue mémoire. On peut ensuite utiliser une procédure classique d'estimation ARMA (procédure itérative reposant sur l'algorithme de Gauss-Newton) pour mesurer θ et ϕ.
• **Validation et application aux données.** – Appliquées à

des réalisations synthétiques de processus $\Gamma_{\alpha,\beta}$ farima(ϕ, d, θ), produits par la procédure décrite à la section suivante, ces procédures d'estimation se révèlent présenter des performances statistiques très satisfaisantes.
Ces procédures sont mises en œuvre sur des portions de données réelles jugées stationnaires. La stationnarité est ici vérifiée expérimentalement en vérifiant la consistance des estimations des paramètres obtenues sur des blocs adjacents sans chevauchement.

5 Synthèse

Nous présentons maintenant une procédure de synthèse permettant de produire des réalisations d'un processus X possédant une marginale $\Gamma_{\alpha,\beta}$ et une covariance de FARIMA(1,d,1), prescrites a priori pour correspondre à celles des données.

• **Principe.** – Le principe général de la synthèse de processus non gaussiens de covariance prescrite est décrit dans [9]. Nous l'adaptons au cas de la synthèse d'un processus X possédant une marginale $\Gamma_{\alpha,\beta}$ et une covariance de farima(1,d,1), prescrites a priori pour correspondre à celles des données analysées et en détaillons les point-clés.

– On obtient une variable $\Gamma_{\alpha,\beta}$ en sommant 2α variables aléatoires gaussiennes indépendantes de moyenne nulle et de variance $\beta/2$ (en effet, en sommant le carré de deux variables aléatoires gaussiennes i.i.d, on obtient une variable exponentielle ; en sommant α exponentielles, on obtient une gamma) :

$$X = \sum_{i=1}^{i=\alpha} Y_{2i}^2 + Y_{2i+1}^2. \qquad (4)$$

– On peut montrer (cf. ci-dessous) que la covariance γ_X et la corrélation $\rho_X = \gamma_X/\sigma_X^2$ du processus $X(k)$ sont liées à ρ_Y et γ_Y, identiques pour les 2α processus $Y_i(k)$, par :

$$\rho_Y = \sqrt{\rho_X},\ \gamma_Y = \sqrt{\gamma_X/(4\alpha)}. \qquad (5)$$

– On synthétise les 2α processus gaussiens Y_i de covariance γ_Y par la a méthode dite "*circulant embedded matrix*" (voir, par exemple, une présentation pédagogique et comparée dans [10]), puis on utilise la relation 4.

• **Calcul de la covariance.** – Pour obtenir l'un et l'autre des résultats de la relation 5, il faut faut décomposer $Y_i(t+k)$ en une prédiction et une innovation [11] : $Y(t+k) = \rho_Y(k)Y(t) + Z(t,k)$. Il est alors aisé de vérifier que $\mathbb{E}Y(t)Z(t,k) = 0$. En reportant cette décomposition dans 4, un calcul lourd mais non difficile permet d'obtenir le résultat souhaité.

• **Limitations.** – Cette procédure de synthèse ne fonctionne que dans le cas où α est pas entier.

L'équation 5 implique aussi que $\gamma(k)$ soit positive, ce qui impose les restrictions suivante sur ϕ et θ : $\phi > 0$ et $\phi > \theta$.

6 Résultats et Discussion.

Les figures 1 et 2 illustrent les résultats obtenus. La figure 1 superposent les marginales empiriques aux lois théoriques $\Gamma_{\hat{\alpha},\hat{\beta}}$ pour les données réelles et synthétiques (colonnes de gauche et droite, respectivement). La figure 2 superposent les diagrammes log-échelle expérimentaux à ceux calculés analytiquement en combinant les relations 1 et 2 pour les données réelles et synthétiques (colonnes de gauche et droite, respectivement). Par

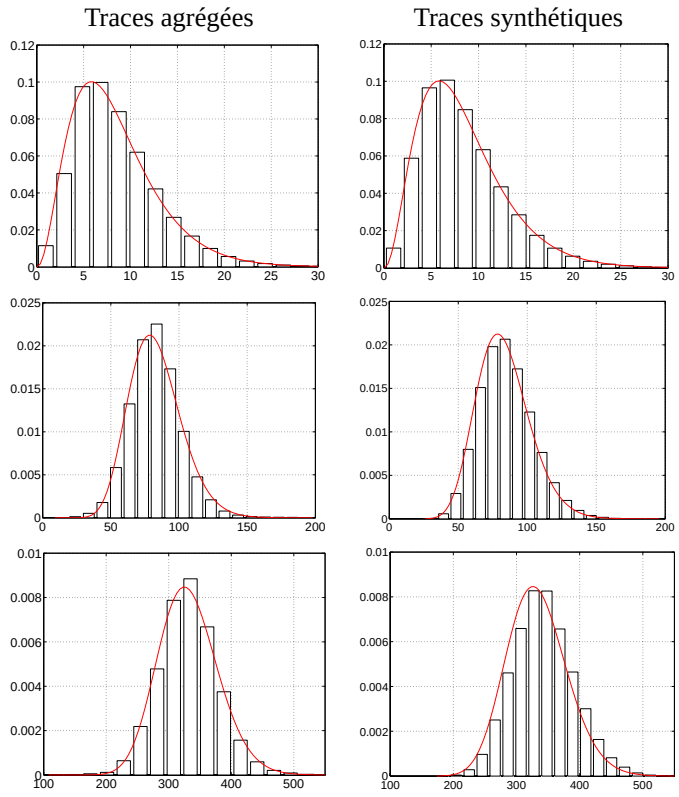

FIG. 1 – **Analyse et synthèse des marginales.** Colonne de gauche : ajustements par des lois $\Gamma_{\alpha,\beta}$ des marginales des séries X_Δ agrégées à $\Delta = 10, 100, 400$ ms (de haut en bas). Colonne de droite : les mêmes ajustements pour des processus synthétiques dont les paramètres ont été choisis a priori pour correspondre à ceux de la trace réelle. Données AUCKIV.

soucis de place, ceux-ci ne seront présentés que sur la trace AUCKIV. Des conclusions identiques restent néanmoins valables pour toutes les traces analysées. Ces figures montrent que le modèle $\Gamma_{\alpha,\beta}$ FARIMA(ϕ, d, θ) décrit de façon très satisfaisante les marginales et covariance des données de trafic X_Δ et ce pour une très large gamme de paramètres Δ (pour la trace considérée, $10ms \leq \Delta \leq 10s$). Ce modèle offre donc une modélisation souple et valide pour un très large continuum de Δ. Cette adéquation des lois $\Gamma_{\alpha,\beta}$ aux marginales de X_Δ résulte essentiellement du fait que les lois $\Gamma_{\alpha,\beta}$ constituent une famille stable sous addition (donc par agrégation). Le paramètre α augmente avec Δ, indiquant que les marginales de X_Δ évoluent vers une gaussienne pour les grands Δ ; ainsi la loi $\Gamma_{\alpha,\beta}$ fournit qualitativement une version adaptée aux traces agrégées X_Δ du théorème de la limite centrale.

La forme des diagrammes log-échelle de la figure 2 rend compte de l'existence de longue mémoire (une droite croissante dans la limite des grandes échelles) ainsi que de celle de corrélation à court terme (décrochement de la droite matérialisant la longue mémoire pour les petites échelles). La pertinence du modèle FARIMA(ϕ, d, θ) se matérialise par le fait que $\hat{d}$ ne varie pas quand Δ évolue, indiquant qu'il rend bien compte d'une propriété de longue mémoire présente dans les données et persistante sous-agrégation. Au contraire, les estimés $\hat{\phi}$ et $\hat{\theta}$ décroissent quand Δ augmente, matérialisant le

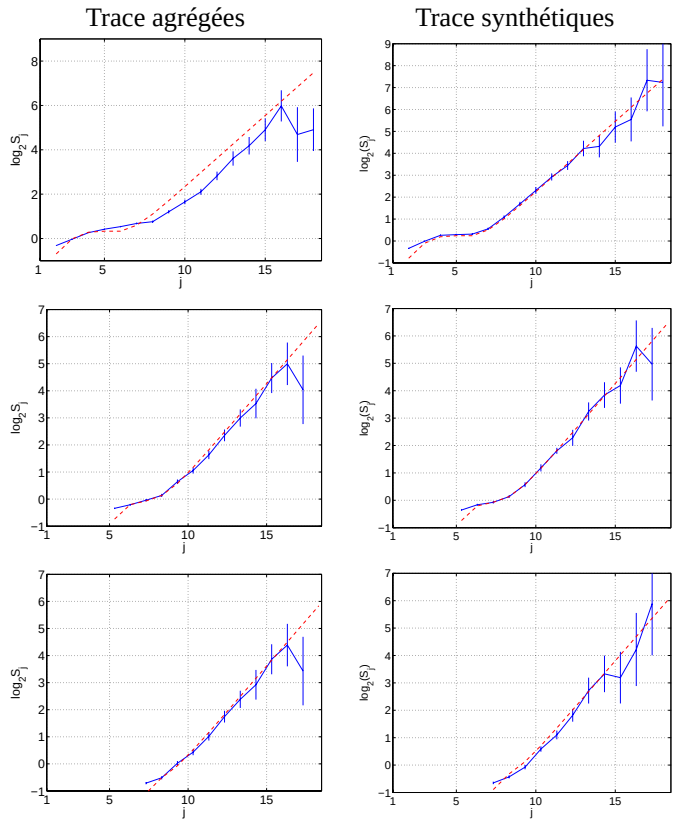

FIG. 2 – **Analyse et synthèse des covariances.** Colonne de gauche : ajustements des diagrammes log-échelle des séries X_Δ agrégées à $\Delta = 10, 100, 400$ ms (de haut en bas) par ceux, calculés numériquement, de processus FARIMA($\hat{\phi}, \hat{d}, \hat{\theta}$). Colonne de droite : les mêmes ajustements pour des processus synthétiques dont les paramètres ont été choisis a priori pour correspondre à ceux de la trace réelle. Données AUCKIV.

fait que les corrélations à court termes sont, elles, peu à peu gommées par le niveau croissant d'agrégation. La structure de covariance de X_Δ tend alors vers celle d'un processus asymptotiquement autosimilaire, bien approximée par celle d'un FARIMA(0,d,0).
Les colonnes de gauche des figures 1 et 2 illustrent les mêmes analyses réalisées sur des traces synthétiques simulées à partir de la procédure décrite plus haut pour des paramètres correspondants à ceux estimés sur la trace réelle de la même ligne. On note d'une part que la procédure de synthèse fournit des traces qui possèdent parfaitement les statistiques prescrites et d'autre part, que la ressemblance entre séries expérimentales et simulées est très satisfaisante.

7 Conclusion et perspectives

Nous avons ici proposé l'utilisation d'un processus stochastique non gaussien à longue mémoire comme modèle parcimonieux pour les séries temporelles de trafic informatique. Nous avons mis en évidence le fait que ce modèle était pertinent pour une très large gamme de niveau d'agrégation. Nous avons présenté des procédures d'estimation des paramètres du modèle ainsi que de synthèse numérique pour produire des réalisations de ces processus. Des routines MATLAB, développées par nos soins implantent la totalité de ces procédures.
La procédure de synthèse de processus dont les statistiques d'ordre 1 et 2 sont prescrites peut être étendue à d'autres types de marginales (log-normal, exponentiel, chi-2, Pareto, ...) et covariances (celles d'un bruit gaussien fractionnaire, d'un FARIMA(p,d,q),...). Ces calculs et développements algorithmiques sont en cours. Par ailleurs, il sera intéressant d'utiliser la possibilité de synthétiser des traces ressemblant à celles observées sur Internet pour nourrir des maquettes de réseaux étudiant l'évolution de la qualité de service selon des scénarios donnés. On pourra notamment simuler des traces dont les paramètres s'écartent de façon contrôlée de ceux des traces réelles pour évaluer l'impact induit sur la qualité de service.

• **Remerciements.** – Nous remercions nos collègues, S. Marron, F. Hernandez-Campos et C. Park d'UNC, USA, et D. Veitch et N. Hohn de CubinLab, Université de Melbourne, qui nous ont fourni les données et en ont assuré la mise en forme.

Références

[1] Benjamin Melamed, "An overview of tes processes and modeling methodology," in *Performance/SIGMETRICS Tutorials*, 1993, pp. 359–393.

[2] K. Park and W. Willinger, "Self-similar network traffic : An overview," in *Self-Similar Network Traffic and Performance Evaluation*, Kihong Park and Walter Willinger, Eds., pp. 1–38. Wiley (Interscience Division), 2000.

[3] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *ACM/IEEE transactions on Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[4] J. Beran, *Statistics for Long-memory processes*, Chapman & Hall, 1994.

[5] G. Samorodnitsky and M. Taqqu, *Stable Non-Gaussian Random Processes*, Chapman & Hall, 1994.

[6] S. Paulo, V. Rui, and P. António, "Multiscale fitting procedure using markov modulated poisson processes," *Telecommunication Systems*, vol. 23 (1/2), pp. 123–148, June 2003.

[7] M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*, Wiley (Interscience Division), June 2000.

[8] P. Abry, P. Flandrin, M.S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation and synthesis of scaling data," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. 2000, WILEY.

[9] M. Crouse and R. Baraniuk, "Fast, exact synthesis of gaussian and nongaussian long-range-dependent processes," *IEEE Trans. on Info. Theory*, 1999.

[10] J.M. Bardet, G. Lang, G. Oppenheim, A. Philippe, and M.S. Taqqu, *Long-Range Dependence : Theory and Applications*, chapter Generators of long-range dependent processes : a survey, pp. 579–623, Birkhäuser, 2003.

[11] S.B. Lowen, S.S. Cash, M. Poo, and M.C. Teich, "Quantal neurotransmitter secretion rate exhibits fractal behavior," *The journal of Neuroscience*, vol. 17, no. 15, pp. 5666–5677, Aug. 1997.