zipfextR: Tres extensiones de la distribución Zipf

IX JORNADAS DE USUARIOS DE R.
Granada - Noviembre, 2017

A. Duarte-López [1,2] & M. Pérez-Casany[1,3]

[1]Data Management Group (DAMA - UPC)

[2]Dpt. de Arquitectura de Computadores, Universidad Politécnica de Cataluña

[3]Dpt. Estadística e Investigación Operativa, Universidad Politécnica de Cataluña

DAMA-UPC. DATA MANAGEMENT
UNIVERSITAT POLITÈCNICA DE CATALUNYA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Introducción

- moezipfR package. VIII Jornadas R-Users.
  - `https://cran.r-project.org/web/packages/moezipfR/index.html`.
- zipfextR package. IX Jornadas R-Users.
  - *Próximamente en CRAN*

Motivación

Una v.a. $X \sim Zipf(\alpha)$ con $\alpha > 1$ si y solo si:

$$P(Y = x) = \frac{x^{-\alpha}}{\zeta(\alpha)}, x = 1, 2, 3, ...$$

donde $\alpha > 1$ y $\zeta(\alpha)$ es la función Zeta de Riemann.

Principales carácteristicas:

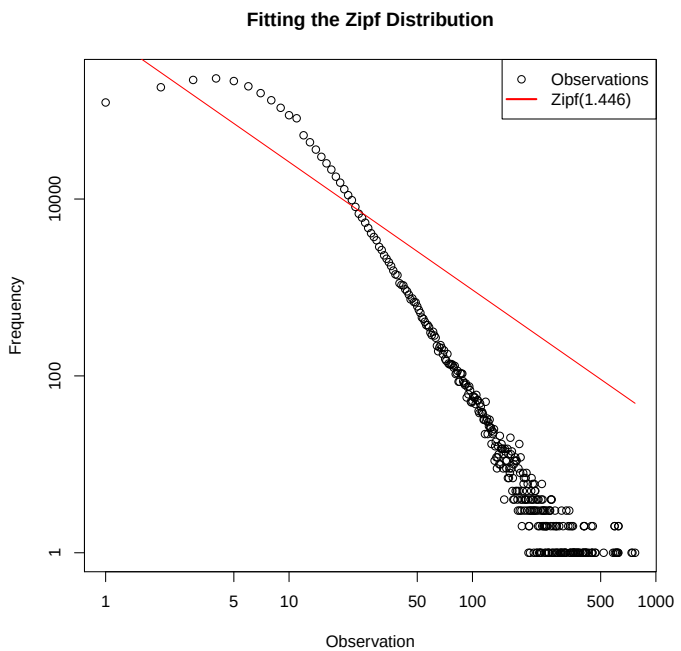
- La probabilidad más alta se obtiene cuando $x = 1$.
- Tiene un comportamiento linear en escala log-log.
- Distribución asimétrica sesgada a la derecha.

Motivación

Secuencia del out-degree de la red de referencias entre las patentes otorgadas en EEUU en el período 1975 - 1999. [1]

Total de nodos: 3774768. Total de aristas: 16518948

Grado de un nodo: Número de conexiones.



Marshall-Olkin extended Zipf

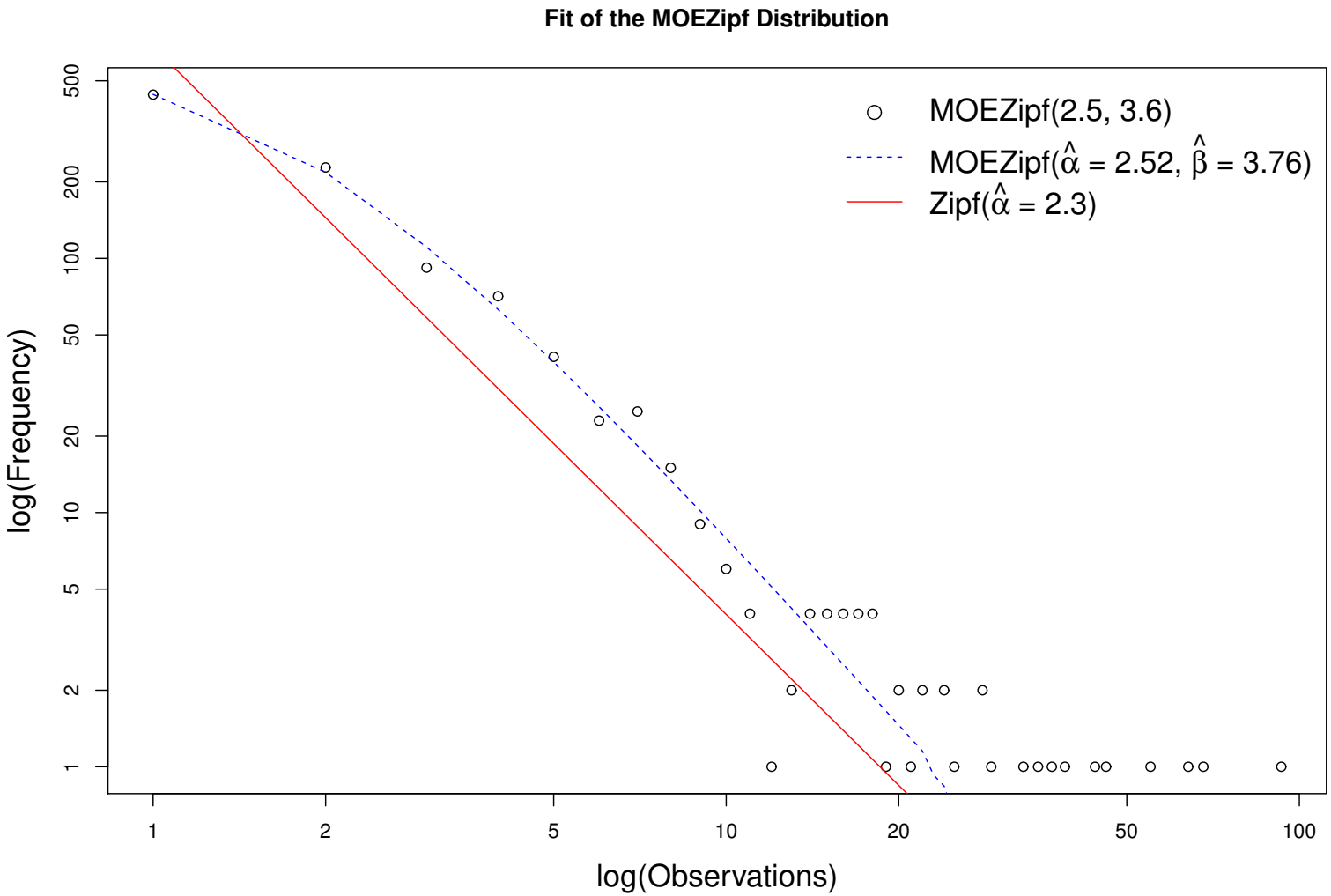
Sea una v.a $Y \sim MOEZipf(\alpha, \beta)$ con $\alpha > 1$ y $\beta > 0$. Si $X_1, X_2, ..., X_N$ son copias i.i.d de una v.a. $X \sim Zipf(\alpha)$ entonces:

- Si $\beta \in (0, 1)$ donde $Y = max(X_1, X_2, ..., X_N)$ donde $N \sim Geom(\beta)$.
- Si $\beta > 1$ donde $Y = min(X_1, X_2, ..., X_N)$ donde $N \sim Geom(\frac{1}{\beta})$.
- Si $\beta = 1$ entonces $Y \sim Zipf(\alpha)$

Obs: Soporte de la distributición: $x = 1, 2, 3, ...$

Marshall-Olkin extended Zipf

1000 valores de una $MOEZipf(\alpha = 2.5, \beta = 3.6)$



Zipf Poisson Extreme

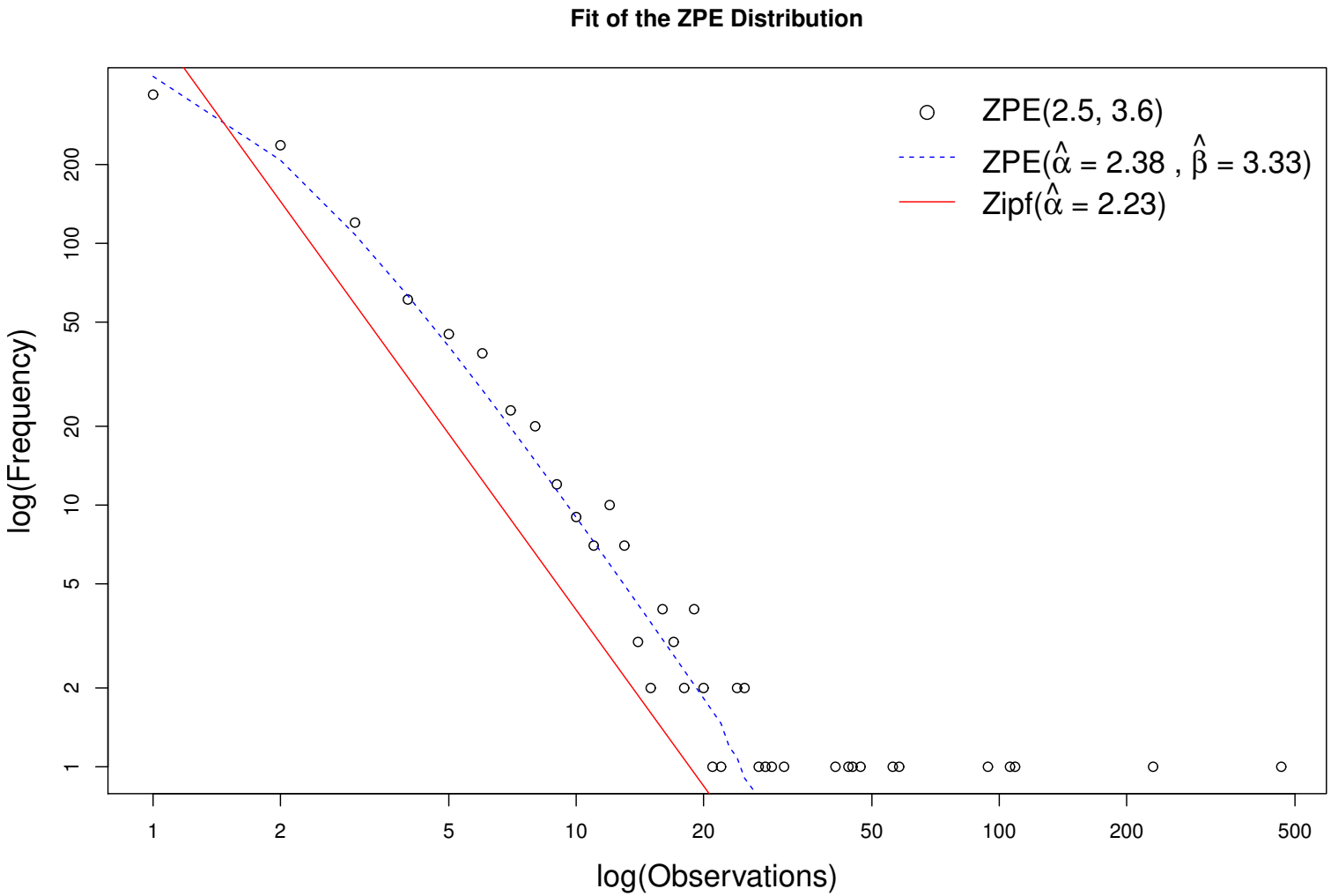
Sea una v.a $Y \sim ZPE(\alpha, \beta)$ con $\alpha > 1$ y $\beta \in \mathcal{R}$. Si $X_1, X_2, ..., X_N$ son copias i.i.d de una v.a. $X \sim Zipf(\alpha)$ entonces:

- Si $\beta > 0$ donde $Y = max(X_1, X_2, ..., X_N)$ donde $N \sim PosPoiss(\beta)$.
- Si $\beta < 0$ donde $Y = min(X_1, X_2, ..., X_N)$ donde $N \sim PosPoiss(-\beta)$.
- Si $\beta = 0$ entonces $Y \sim Zipf(\alpha)$

Obs: Soporte de la distributición: $x = 1, 2, 3, ...$

Zipf Poisson Extreme

1000 valores de una $ZPE(\alpha = 2.5, \beta = 3.6)$



Zipf Poisson Stopped Sum

Sea una v.a $Y \sim ZPSS(\alpha, \lambda)$ con $\alpha > 1$ y $\lambda > 0$. Si $X_1, X_2, ..., X_N$ son copias i.i.d de una v.a. $X \sim Zipf(\alpha)$ entonces:

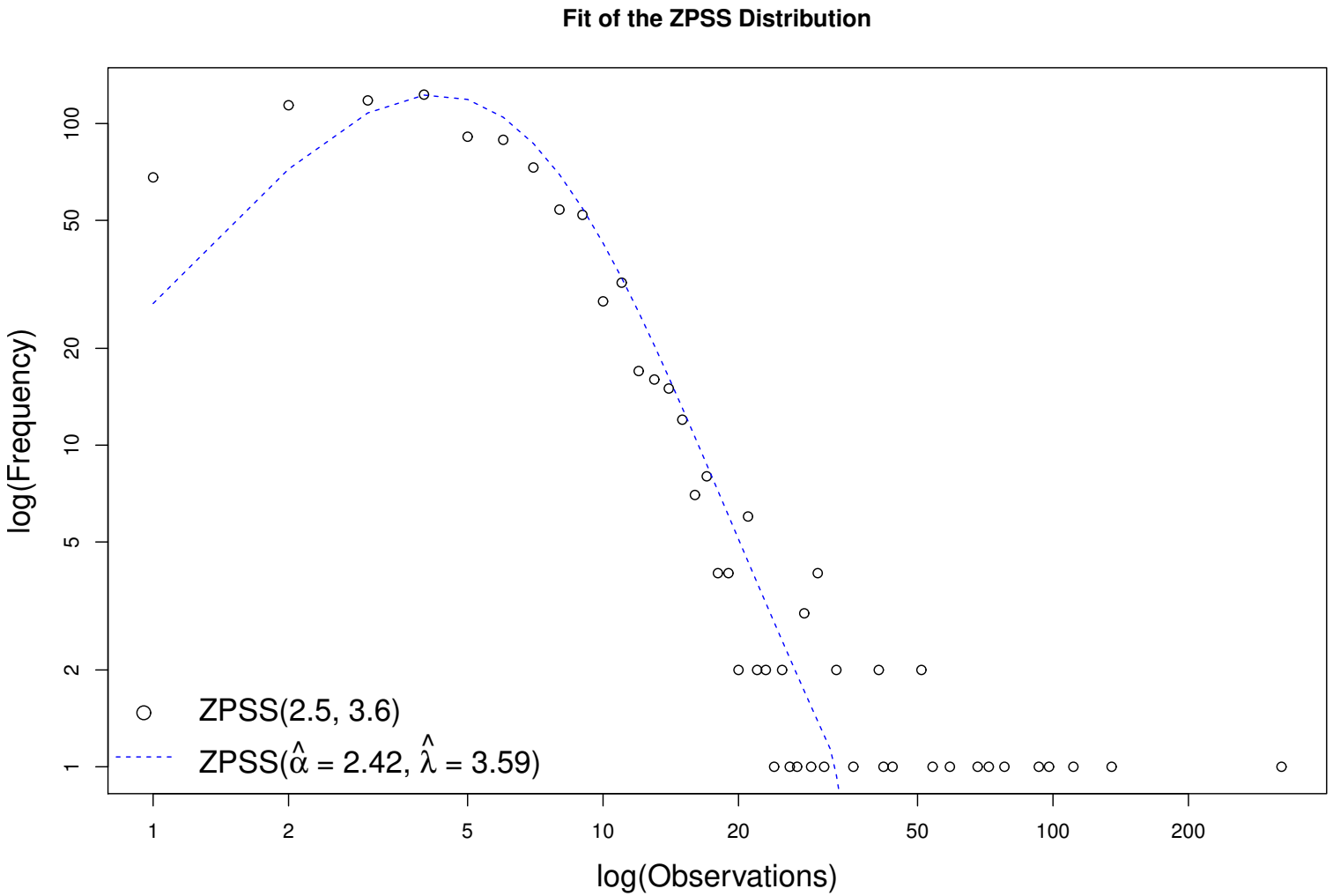
- Si $\beta > 0$ donde $Y = X_1 + X_2 + ... + X_N$ donde $N \sim Poiss(\lambda)$.
- Si $\mathcal{L}(Y) \xrightarrow[\lambda \to 0]{} Zipf(\alpha)$

La distribución ZPSS utiliza la Recursión de Panjer [3, 5] para calcular las probabilidades.

Obs: Soporte de la distributición: $x = 0, 1, 2, 3, ...$
$P(Y = 0) = e^{-\lambda}$

Zipf Poisson Stopped Sum

1000 valores de una $ZPSS(\alpha = 2.5, \lambda = 3.6)$



Funciones

MOEZipf	ZPE	Z-PSS	Descripción
dmoezipf	dzpe	dzpss	Función de probabilidad.
pmoezipf	pzpe	pzpss	Función de probabilidad acumulada.
qmoezipf	qzpe	qzpss	Función de cuantiles.
rmoezipf	rzpe	rzpss	Generador de números aleatorios.
moezipfMoments	zpeMoments	zpssMoments	$k-$ésimo momento de la distribución.
moezipfMean	zpeMean	zpssMean	Esperanza de la distribución.
moezipfVariance	zpeVariance	zpssVariance	Varianza de la distribución.
moezipfFit	zpeFit	zpss	Estimador máximo verosímil.

Caso de Uso: Redes reales

La red utilizada en este anánalis contine información sobre la interaci ón de proteínas en la *Arabidopsis thaliana* (planta herbácea). Los nodos representan genes, micro RNA, metabolitos, y proteíanas virales.

Principales características:

Nodos	20012
Aristas	70091
Nodos con grado cero	0

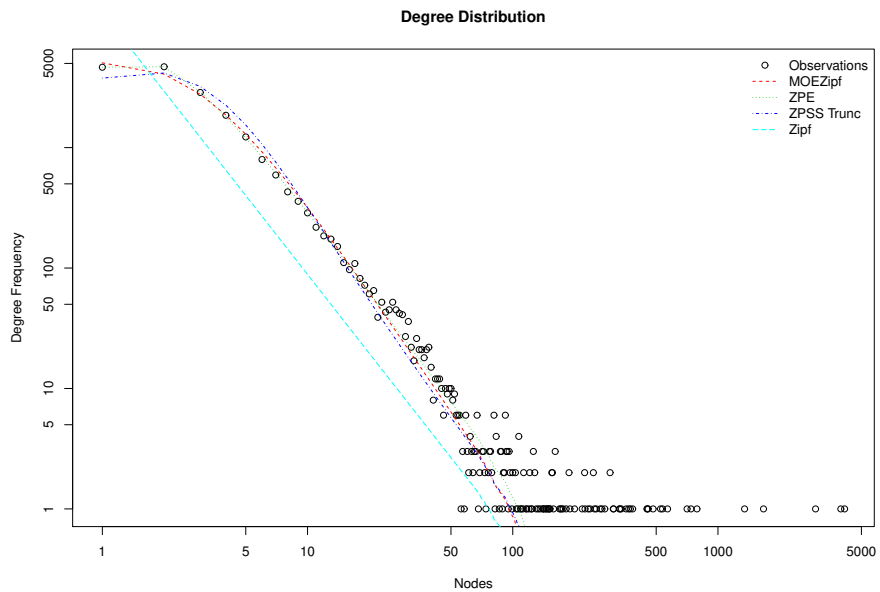
Caso de Uso: Redes reales

Parameters

Distribution	*Parameters*
MOEZipf	$\hat{\alpha} = 2.5476$ $\hat{\beta} = 9.1472$
ZPE	$\hat{\alpha} = 2.3198$ $\hat{\beta} = 4.8279$
Z-PSS_trunc	$\hat{\alpha} = 2.165$ $\hat{\lambda} = 1.6676$
Zipf	$\hat{\alpha} = 2.1763$

Goodness of the fit

Distribution	*AIC*	ΔAIC
MOEZipf	−99312.39	52.8544
ZPE	−99110.88	254.3682
Z-PSS_trunc	−99365.25	0
Zipf	−54680.79	44684.4552



Caso de Uso: Redes reales

Red dirigida donde los nodos representan páginas web de la Universidad de Notre Dame (domain: nd.edu) y las aristas representan los links entre ellas. Los datos han sido recogidos en 1999 por Albert, Joeng y Barabasi.

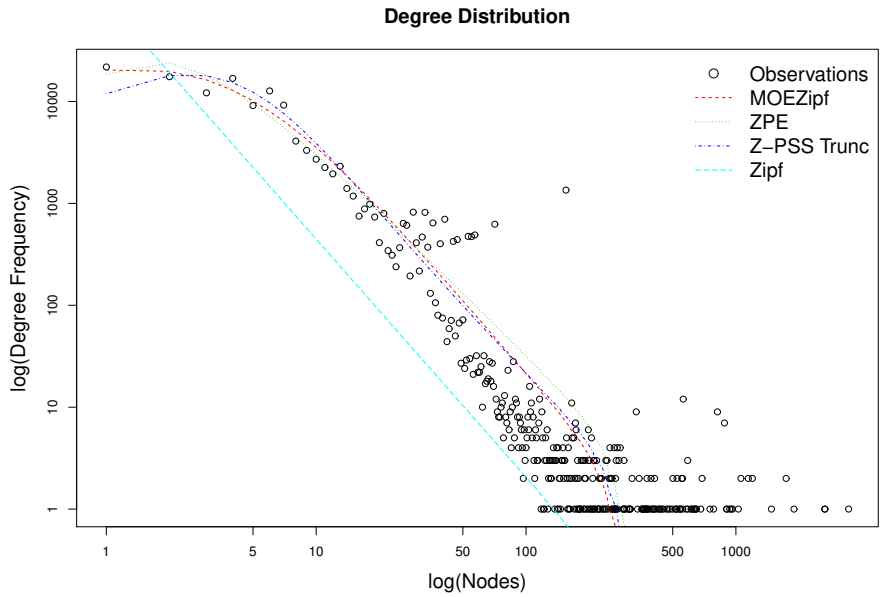
Principales características:

Nodos	325729
Aristas	1497134
Nodos con grado cero	188795

Caso de Uso: Redes reales

Parameters

Distribution	*Parameters*
MOEZipf	$\hat{\alpha} = 2.4223$ $\hat{\beta} = 15.3069$
ZPE	$\hat{\alpha} = 2.0825$ $\hat{\beta} = 5.4104$
Z-PSS_trunc	$\hat{\alpha} = 2.025$ $\hat{\lambda} = 2.4407$
Zipf	$\hat{\alpha} = 2.3367$



Goodness of the fit

Distribution	*AIC*	ΔAIC
MOEZipf	−838185.1	4402.696
ZPE	−842587.8	0
Z-PSS_trunc	−837962.3	4625.484
Zipf	−321533.0	521054.791

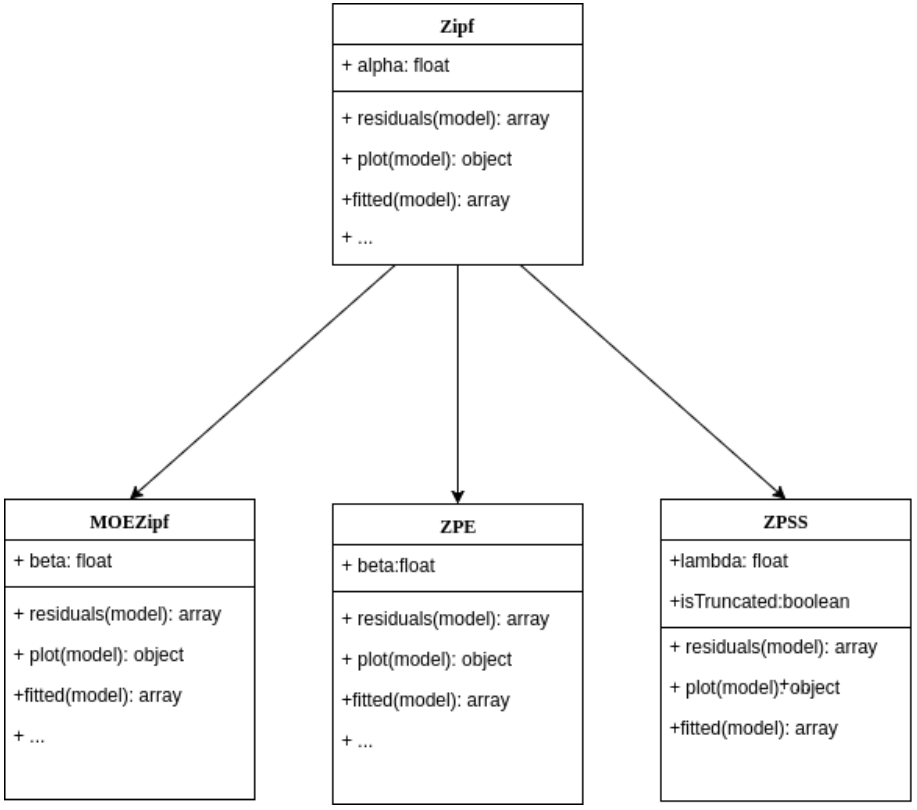
$$\hat{p_0} = 0.0871$$

$$\alpha = 0.05\,(-0.0887,\ 0.2629)$$

$$\alpha = 0.01\,(-0.144,\ 0.3182)$$

OOP

Actualmente las funciones "*Fit" están implementadas utilizando clases **S3**. Lo que permite extender las funcionalidades, *plot*, *summary*, *residuals*, etc.



Bibliografía

[1] Ariel Duarte-López, Arnau Prat-Pérez, and Marta Pérez-Casany. Using the marshall-olkin extended zipf distribution in graph generation. In *European Conference on Parallel Processing*, pages 493–502. Springer, 2015.

[2] Yeşim Güney, Yetkin Tuaç, and Olcay Arslan. Marshall–olkin distribution: parameter estimation and application to cancer data. *Journal of Applied Statistics*, pages 1–13, 2016.

[3] Harry H Panjer. Recursive evaluation of a family of compound distributions. *Astin Bulletin*, 12(01):22–26, 1981.

[4] Ziva Ramsak, Anna Coll, Tjasa Stare, Oren Tzfadia, Spela Baebler, Yves Van de Peer, and Kristina Gruden. Network modelling unravels mechanisms of crosstalk between ethylene and salicylate signalling in potato. *bioRxiv*, 2017.

[5] Bjørn Sundt and William S Jewell. Further results on recursive evaluation of compound distributions. *ASTIN Bulletin: The Journal of the IAA*, 12(1):27–39, 1981.

Agradecimientos

Los autores quieren agradecer a Oracle Labs por el apoyo al proyecto Graphalytics y al apoyo de EC FP-7 a través del proyecto LDBC.

Ariel Duarte-López quiere agradecer a la Dr. Ziva Ramsak (Instituto Nacional de Biología de Slovenia) por los conjuntos de datos aportados. Además, agradece la colaboración de la Agència de Gestió d' Ajuts Universitaris i de Recerca (AGAUR). Grant FI-DGR 2016.

Marta Pérez-Casany agradece al Ministerio de Ciencia e Inovación (España). Grant No. MTM2013-43992-R.

Muchas Gracias!!!