Italian National Agency for New Technologies,
Energy and Sustainable Economic Development

Analysis of the Web Graph Aggregated by Host and Pay-Level Domain

Agostino Funel

agostino.funel@enea.it

ENEA DTE-ICT-HPC

Complex Networks - December 11-13, 2018 – Cambridge (UK)

Motivations

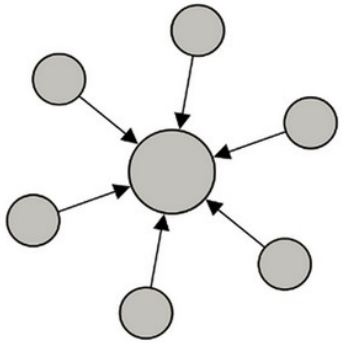
To provide additional information on the web graph structure.

The main distributions of the web (degree/components) are widely supposed to have power law tails. Is it statistically supported by data?

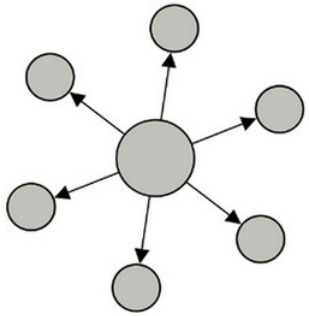
How does the description of the web depend on the scale of observation?

Basic definitions

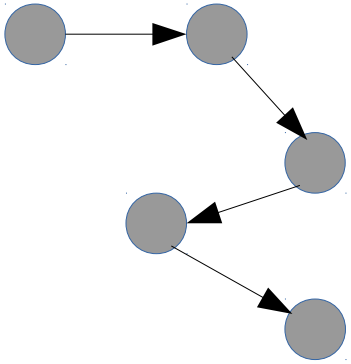
Synonims = {node/vertex, link/arc/connection}
Node degree = # of links attached to a node



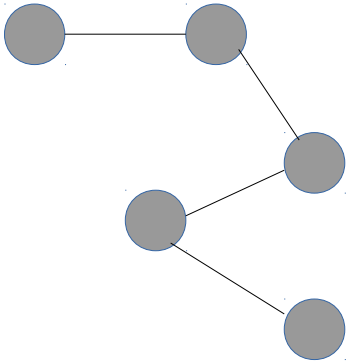
In-degree



Out-degree



Path



Walk

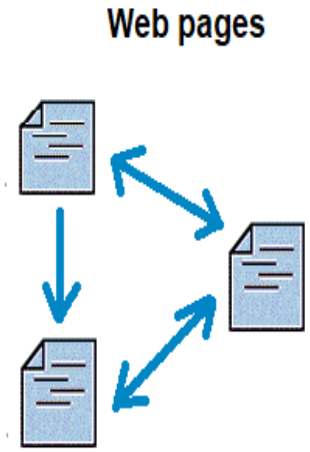
Strongly/Weakly connected component SCC/WCC: a maximal subset in which every pair of nodes is connected by a path/walk

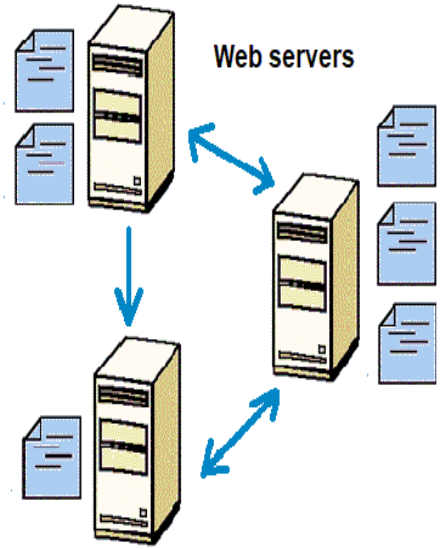
Distance d(x,y): the length of the shortest path/walk from x to y

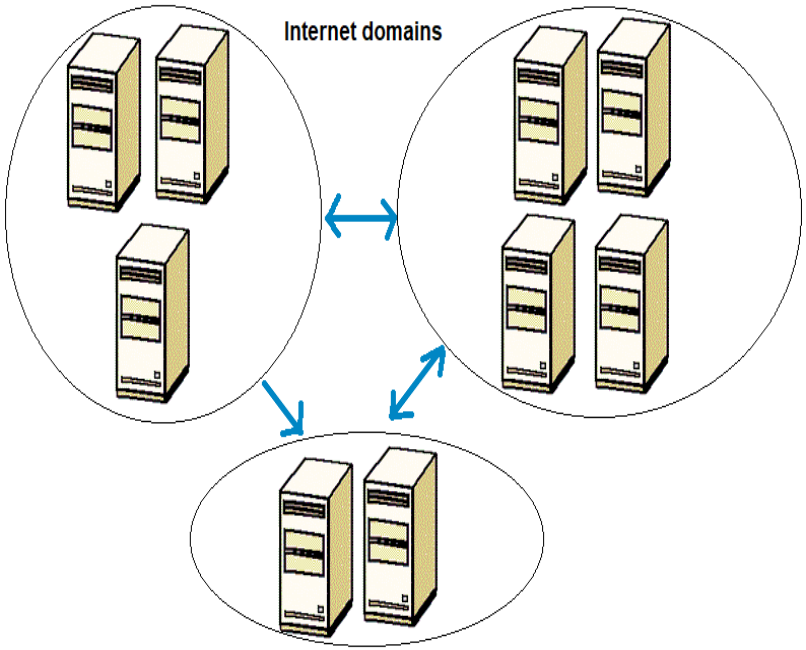
Diameter: the length of the longest shortest path

Effective diameter δ: given two random chosen nodes there is 90% of probability that they are connected within δ hops

The Web on different levels of aggregation







The page level is the WWW. Each node is a web page (URL) and an arc is a hyper textual link between two web pages.

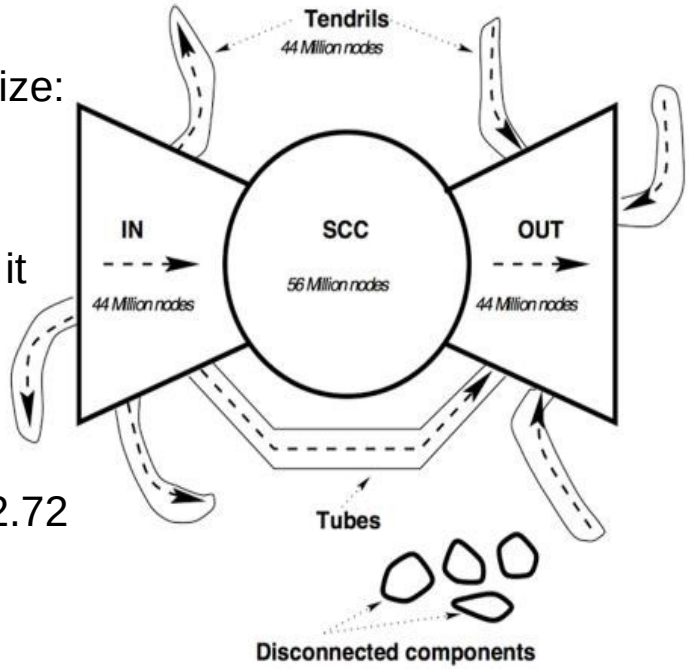
Host level: each node is a web server (IP). Two nodes are connected by an arc if exists at least a link between the web pages hosted on these nodes.

Domain level: each node is an Internet domain. Two domains are connected by an arc if exist at least two web servers, each belonging to one of the domains, connected by an arc.

Previous analysis/1

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata,R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer Networks 33 (1-6), 309-320 (2000)

- Page level. Two Altavista crawls (May and October 1999) each with ~200M pages and ~1.5B links.
- Single giant WCC containing ~90% of the nodes.
- Bow-tie structure: WCC breaks in four pieces roughly of the same size:
  - at the heart there is a SCC containing ~28% of the nodes
  - IN: pages that can reach the SCC but can not be reached from it
  - OUT: pages that can be reached from SCC but do not link back to it
  - TENDRILS: pages completely disconnected fron the SCC
  - TUBES: directed paths from IN to OUT without touching SCC
- Average distance ~16
- in/out degree distributions: power law $P(k) \sim k^{\alpha}$ $\alpha_{in} = -2.10$ $\alpha_{out} = -2.72$
- SCC/WCC size distributions: power law $P(s) \sim s^{\alpha}$ $\alpha = -2.54$



No details about the statistical plausibility of the power laws.

Previous analysis/2

Serrano, M.A., Maguitman, A., Boguñá, M., Fortunato, S., Vespignani, A.: Decoding the Structure of the WWW: A Comparative Analysis of Web Crawls. ACM Transactions on the Web 1(2) (2007)

Four sets of WWW samples with two different crawlers (WebBase, WegGraph projects) from different domains: whole Web (WebBase), .uk (WebGraph), .it (WebGraph). Years 2001-2004.

Data set	WBGC01	WGUK02	WBGC03	WGIT04
# nodes	80,571,247	18,520,486	49,296,313	41,291,594
# links	752,527,660	292,243,663	1,185,396,953	1,135,718,909

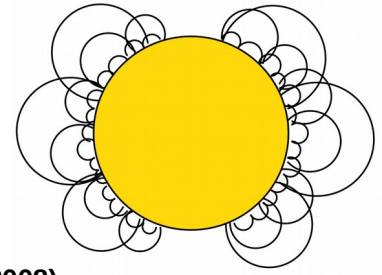
"the statistical measures characterizing these graphs differ quantitatively, and in some cases qualitatively, depending on the domain analyzed and the crawl used for gathering the data"

Donato, D., Leonardi, S., Millozzi, S., Tsaparas, P.: Mining the inner structure of the Web graph. J. Phys. A: Math. Theor. 1(22) (2008)

Four sets of WWW samples with two different crawlers from different domains: Italy, Indochina, UK (UbiCrawler), whole Web (WebBase project).

	Italy	Indochina	UK	WebBase
nodes	41.3M	7.4M	18.5M	135.7M
edges	1.15G	194.1M	298.1M	1.18G

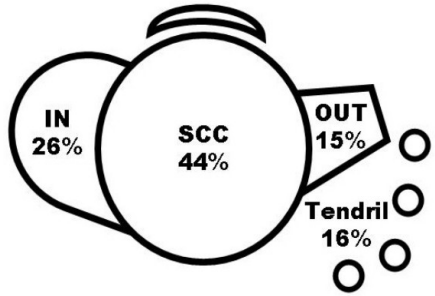
"The picture that emerges from our work can better be described by the shape of a daisy: the IN and OUT regions are fragmented into large number of small and shallow petals (the WCCs) hanging from the central dense CORE"



Zhu, J.J.H., Meng, T., Xie, Z., Li, G., Li, X.: A Teapot Graph and Its Hierarchical Structure of the Chinese Web. Proc. WWW '08 (2008)

China WWW, 837 M pages, 43 B links. Period: January and February 2006.

"A Teapot Graph is constructed as alternative to the classic Bow Tie or Daisy Graphs"



What we have learned so far

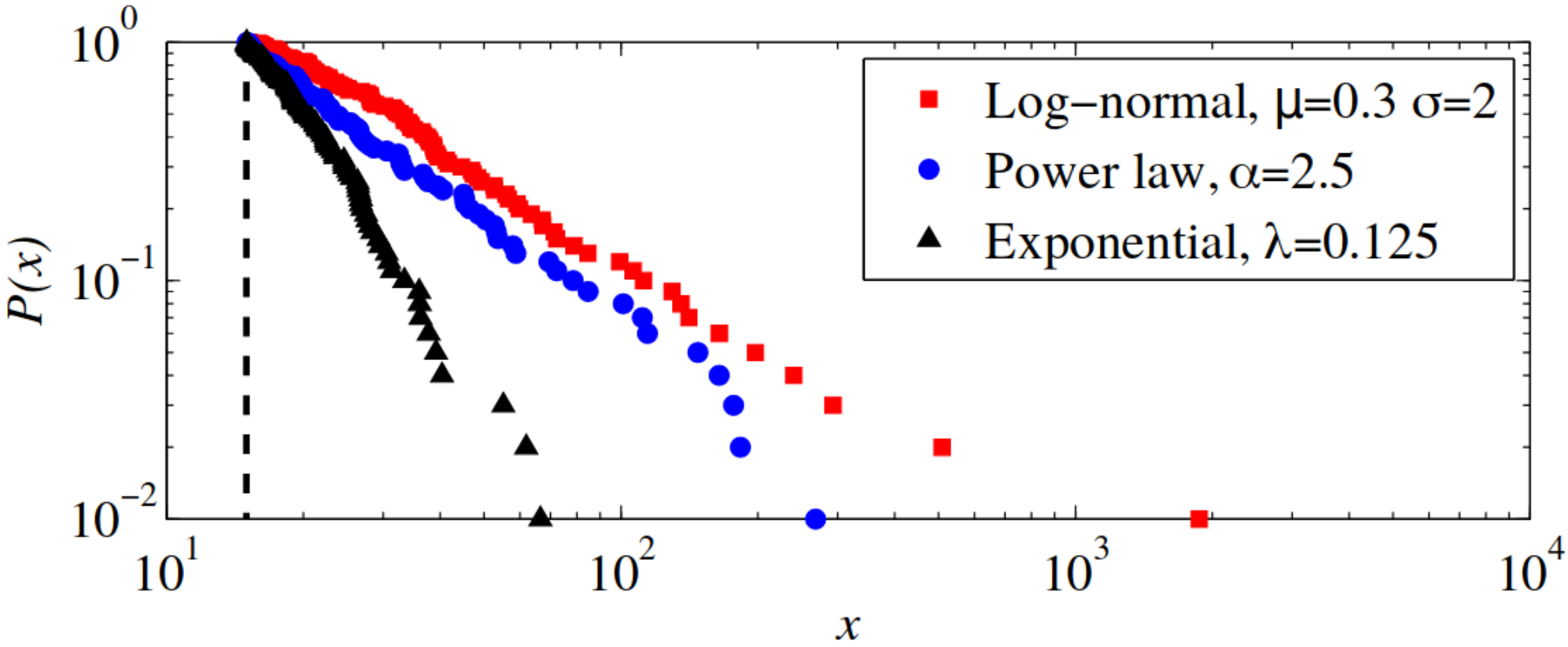
The large scale structure of the Web is not exactly known. However, a giant connected component in the core seems to exist.

The main distributions (degree, components) are heavy tailed and the best description is a power law model (even with noises/cuts).

Power laws deduction: linear fit of log-log plots, maximum likelihood.

Lack of in-depth statistical investigations on the plausibility of the power law model.

Example: different distributions look linear on a log-log plot



Previous analysis/3

Meusel, R., Vigna, S., Lehmberg, O., Bizer, C.: The Graph Structure in the Web Analyzed on Different Aggregation Levels. The Journal of Web Science 1, 33–47 (2015)

Granularity	# Nodes in millions	# Arcs in millions
Page Graph	3 563	128 736
Host Graph	101	2 043
PLD Graph	43	623

Data publicly available from Common Crawl Foundation, gathered in the first half of 2012 and released in August of the same year.

Authors used statistical methods to test power laws for the distributions of degree and sizes of SCC/WCC.

Clauset, A., Shalizzi, C.R., Newman, M.E.J.: Power law distributions in empirical data. SIAM Rev. 51(4), 661–703 (2009)

	Page	Host	PLD
Power Law	No	No	Yes (indegree)
WCC	94%	87%	92%
SCC	51%	47%	52%
Average Distance	12.8	5.3	4.3
Diameter (lower bound)	5282	261	48

Power laws exist only for the indegree distribution of the PLD graph.

A giant WCC exists in the Web for page/host/PLD Web graphs.

There is a SCC containing ~50% of the nodes of the page/host/PLD Web graphs.

The WWW is more connected than what estimated by Broder et al.

Open questions

Does really the power law behavior of the Web distributions depend on the scale of observation?

Are there alternative models which better fit data?

Can we confirm the existence of a giant connected component in the Web?

On page level the web seems to become more connected over time. Is it true also for the other levels of aggregation?

Data sets and definitions for this analysis

Data publicly available from Common Crawl Foundation[1], gathered during May-June-July 2017

Granularity	# Nodes (M)	# Arcs (M)
Host	1307	5268
PLD	91	1071

Host: The name or address of a web server can be extracted by its URL by excluding protocol, authentication, port, path, query and fragment substrings.

Pay-Level Domain (PLD): This level of aggregation is based on the Public Suffix List, an initiative of Mozilla[2]. It is a catalog of Internet domain name suffixes that can be directly registered by users. The PLD of a host is obtained by aggregating one dot above the public suffix.

URL	http://www.example.1.com
Host	www.example.1.com
PLD	1.com

[1] http://commoncrawl.org
[2] https://www.publicsuffix.org

Methodology of analysis

For each in/out degree and sizes of SCC/WCC distribution

Make a best fit to a power law $P(x) \sim x^{-\alpha}$ (maximum likelihood) → (x_{min}, α)
Goodness of fit test (Kolmogorov-Smirnov statistic) → p-value

0 ≤ p < 0.1	Power law hypothesis rejected
0.1 ≤ p ≤ 1	Power law hypothesis accepted

If power law is accepted as a further check compare experimental data with synthetic data randomly generated from a power law with (x_{min}, α) parameters

Comparison of power law with alternative models: $e^{-\lambda x}$ **(exp),** $\frac{1}{x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$ **(logn),** $x^{-\alpha} e^{-\lambda x}$ **(tpl),** $x^{\beta-1} e^{-\lambda x^{\beta}}$ **(sexp)**

Comparison between two distributions f_A and f_B: normalized loglikelihood ratio $R(f_A/f_B)$

$R(f_A/f_B) > 0$	f_A is the favorite distribution
$R(f_A/f_B) = 0$	There is not a favorite model
$R(f_A/f_B) < 0$	f_B is the favorite distribution

Statistical significance of the sign of R → q-value

0 ≤ q < 0.1	The sign of R is a reliable indicator
0.1 ≤ q ≤ 1	The sign of R is not a reliable indicator

Statistical plausibility of f as a better fit distribution compared to the power law.

In the case of many candidates models alternative to the power law compare them pairwise to find out the best one.

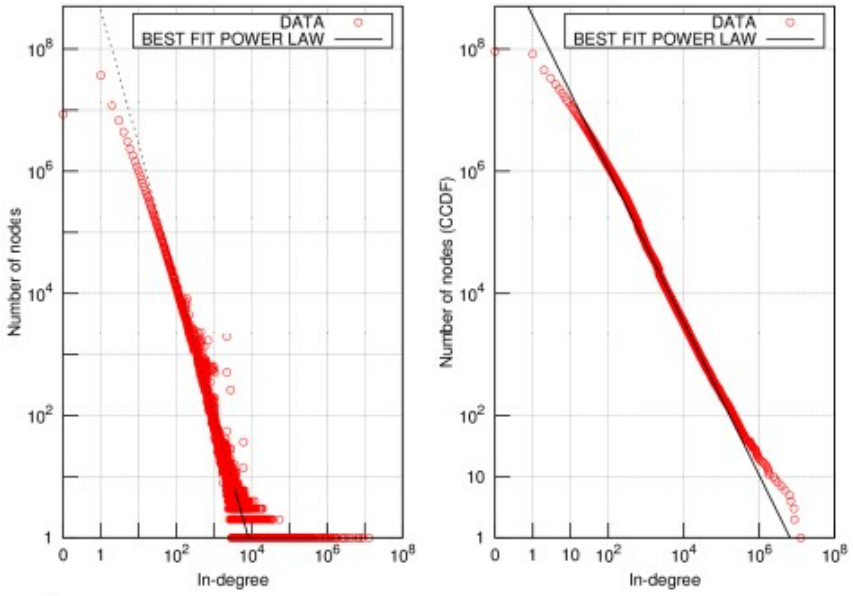
R(pl/f) > 0	0 ≤ q < 0.1	None
- ∞ < R(pl/f) < + ∞	0.1 ≤ q ≤ 1	Undecidable
R(pl/f) = 0	0 ≤ q ≤ 1	Undecidable
R(pl/f) < 0	0 ≤ q < 0.1	Strong

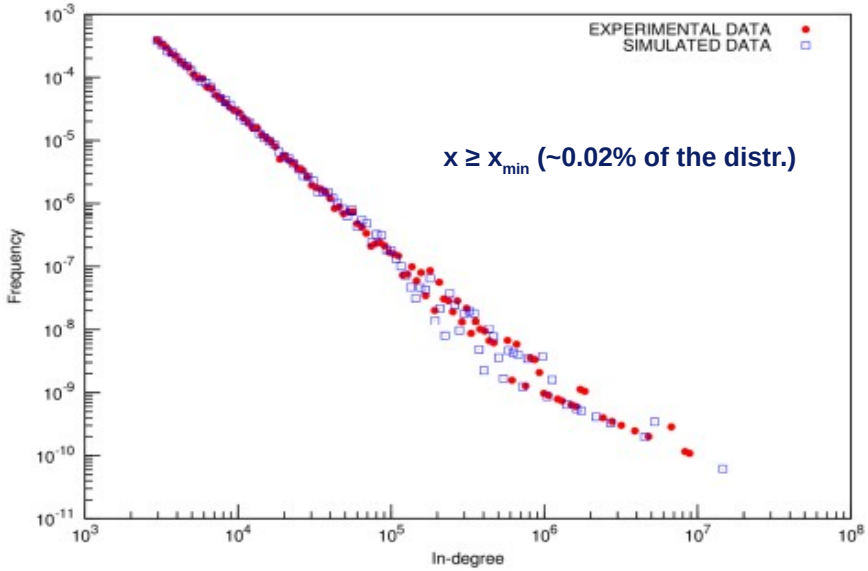
Average distance: shortest path lengths distribution. Diameter: BFS with 10000 starting nodes

Software

Library	Reference
Goodness of fit tests: plfit http://github.com/ntamas/plfit	Clauset, A., Shalizzi, C.R., Newman, M.E.J.: Power law distributions in empirical data. SIAM Rev. 51(4), 661–703 (2009)
Models comparison: powerlaw https://pypi.python.org/pypi/powerlaw	Alstott, J., Bullmore, E., Plenz, D.: powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1) (2014)
Structural properties: SNAP https://snap.stanford.edu/snap/	Leskovec, J., Sosič, R.: Snap: A general-purpose network analysis and graph-mining library. ACM Transactions on Intelligent Systems and Technology (TIST) 8(1), 1 (2016)

Analysis of the PLD Web graph: indegree distribution





$x \geq x_{min}$ (~0.02% of the distr.)

# Nodes	~91 M
# Arcs	~1.1 B
# PLD in 2Q 2017 ~332 M[1]	Analyzed ~27.4% of all PLD
Zero degree nodes	~0.1% of the total
Average degree	~23.56

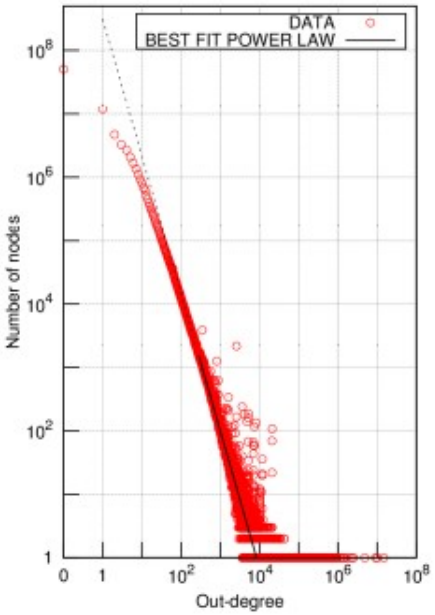
nodes with indegree zero: ~8.6· 10^6 (9.4% of the total)
Max indegree: 12896169
Power law: X_{min} = 2858 α = 2.21 ± 0.01 p = 0.69 ± 0.01

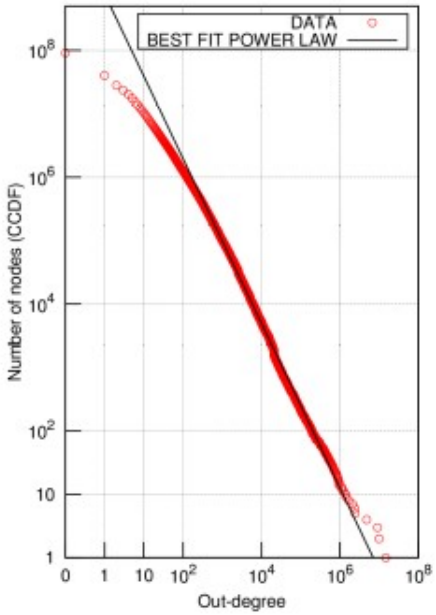
- Concavity in the region 1 ≤ x ≤ 30 also visible in the CCDF. Spikes in the region 2000 ≤ x ≤ 6000
- Even if the tail of the CCDF is not linear in the log-log plot there is statistical evidence of power law
- Good agreement with synthetic data in $x \geq x_{min}$
- Fit calculations: power law is the most statistically plausible model

f	$R(pl/f)$	q	Statistical Plausibility of f as Alternative to the Power Law
exp	9.375661	0	None
logn	1.558554	0.119102	Undecidable
tpl	0.055049	0.988221	Undecidable
sexp	76.482004	0	None
f_A/f_B	$R(f_A/f_B)$	q	Comment
logn/tpl	-1.548861	0.121415	None of the tested models is favorite

[1] https://www.verisign.com/assets/domain-name-report-Q22017.pdf

Analysis of the PLD Web graph: outdegree distribution





$x \geq x_{min}$ (~0.5% of the distr.)

nodes with outdegree zero: ~50.7· 10^6 (55.7% of the total)
Max outdegree: 14903607
Power law: X_{min} = 279 α = 2.164 ± 0.02 p = 0

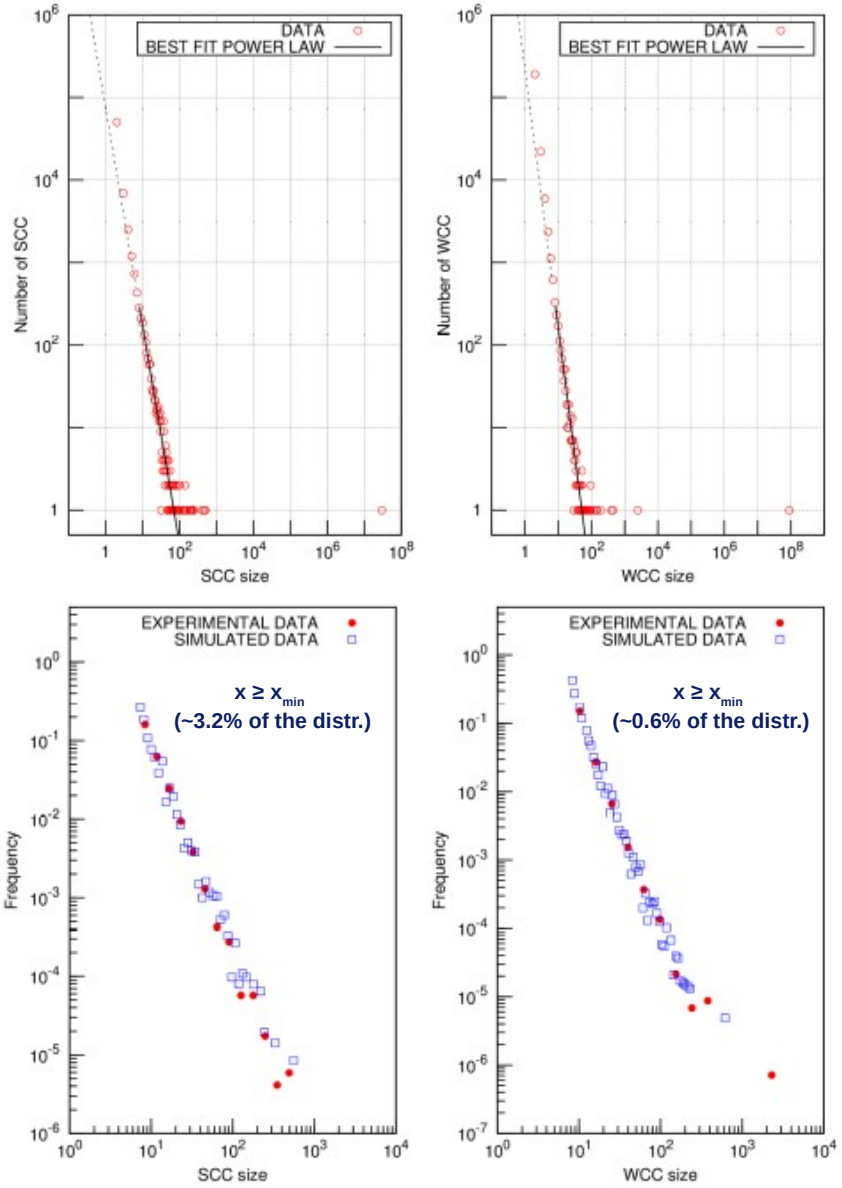
f	$R(pl/f)$	q	Statistical Plausibility of f as Alternative to the Power Law
exp	149.413009	0	None
logn	-59.471894	0	Strong
tpl	-4.724054	0	Strong
sexp	17.653723	0	None
f_A/f_B	$R(f_A/f_B)$	q	Comment
logn/*tpl*	10.801170	0	Strong support for the *logn*

Concavity in the region 1 ≤ x ≤ 60 also visible in the CCDF. Spikes in the region 230 ≤ x ≤ 20000

There is no statistical evidence of power law

Fit calculations: strong support for the lognormal

Analysis of the PLD Web graph: components distributions



Largest SCC: ~32.7%

Power law: X_{min} = 7 α = 2.63 ± 0.04 p = 0.41 ± 0.01

f	$R(pl/f)$	q	Statistical Plausibility of f as Alternative to the Power Law
exp	1.755426	0.079186	None
logn	0.320868	0.748310	Undecidable
tpl	0.999271	0	None
sexp	46.723807	0	None

Largest WCC: ~99.4%

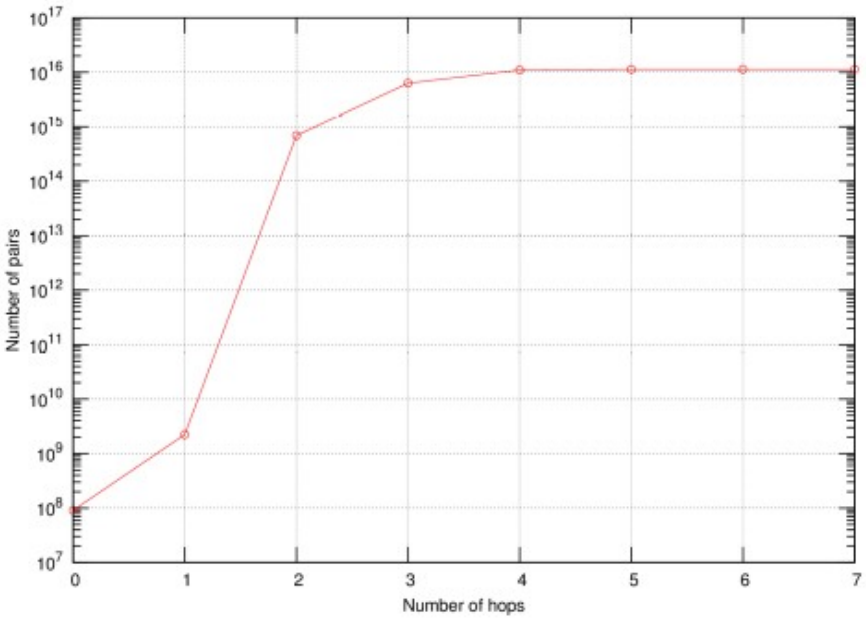
Power law: X_{min} = 8 α = 3.12 ± 0.06 p = 0.34 ± 0.01

f	$R(pl/f)$	q	Statistical Plausibility of f as Alternative to the Power Law
exp	1.786023	0.074096	None
logn	0.431812	0.665878	Undecidable
tpl	0.139838	0.530106	Undecidable
sexp	45.928895	0	None
f_A/f_B	$R(f_A/f_B)$	q	Comment
logn/tpl	-0.100918	0.919616	None of the tested models is favorite

There is statistical support for power law tails for SCC and WCC size distributions

We confirm a giant WCC in the PLD Web graph

Analysis of the PLD Web graph: distances and diameter



Hop-plot - Cumulative distribution of the shortest path lengths

Y-axis: number of pair of nodes with distance within h hops

Very memory consuming calculation. The SNAP library adopts a fast and memory efficient algorithm based on an approximation of the neighbourhood funcion[1]

In the PLD Web graph ~90% of all pairs of nodes have distance within 3.8 ±0.4

The lower bound of the full diameter estimated with a BFS using 10000 random starting nodes is 34

[1]**Palmer, C.R., Gibbons, P.B., Faloutsos, C.: ANF: A Fast and Scalable Tool for Data Mining in Massive Graphs. Proc. KDD '02 (2002)**

Summary and conclusions

May-June-July 2017	Host Graph	PLD Graph
# Nodes	1306661614	91034128
# Arcs	5268397861	1071173924
Indegree		
α	2.193 ± 0.001	2.21 ± 0.01
x_{min}	250	2858
largest	23055296	12896169
Outdegree		
α	2.3242 ± 0.0001	2.164 ± 0.002
x_{min}	23	279
largest	15090917	14903607
SCC		
α	2.367 ± 0.005	2.63 ± 0.04
x_{min}	4	7
largest	$\sim$4.5%	$\sim$32.7%
WCC		
α	1.684 ± 0.001	3.12 ± 0.06
x_{min}	22	8
largest	$\sim$99.7%	$\sim$99.4%
Effective diameter	5.6 ± 0.6	3.8 ± 0.4
Full diameter (lower bound)	970	34

		This analysis Host graph	Meusel et. al Host graph	This analyis PLD graph	Meusel et al. PLD graph
# Nodes (M)		1307	101	91	43
# Arcs (M)		5268	2043	1071	623
Power law statistical support	Indegree	No	No	Yes	Yes
	Outdegree	No	No	No	No
	SCC	No	No	Yes	No
	WCC	No	No	Yes	No
Largest SCC (%)		4.5	47	32.7	52
Largest WCC (%)		99.7	87	99.4	92
Effective diameter		5.6±0.6	5.3±0.001	3.8±0.4	4.27±0.085

Host Web Graph Distribution	Statistical Support for the Power Law	Statistical Support for Alternative Models (Discrete Fit)	Statistical Support for Alternative Models (Continuous Fit)
Indegree	None	Lognornal (strong)	None
Outdegree	None	Lognornal (strong)	Lognornal (strong)
SCC	None	Lognornal (strong)	None
WCC	None	Lognornal (strong)	None

PLD Web Graph Distribution	Statistical Support for the Power Law	Statistical Support for Alternative Models (Discrete Fit)	Statistical Support for Alternative Models (Continuous Fit)
Indegree	Yes	None	None
Outdegree	None	Lognornal (strong)	Lognornal (strong)
SCC	Yes	None	Lognornal (strong)
WCC	Yes	None	None

- **There is no statistical evidence of power law tails on host level for in/out degree and SCC/WCC size distributions.**
- **Power laws emerge on PLD aggregation for indegree, SCC and WCC size distributions.**
- **For both host and PLD graphs, among all tested models alternative to the power law the lognormal is the only which is not ruled out by the likelihood ratio test.**
- **The fraction of nodes in the largest SCC varies considerably and is ~4% in the host graph and ~33% in the PLD graph. Maybe this is due to crawling artifacts and/or aggregation processes.**
- **This analysis confirms the presence of a giant WCC in the host and PLD Web graphs.**
- **On host and PLD levels the effective diameter remains almost the same even if the size of the Web graph increases (even by one order of magnitude as in the case of host level).**

Thank you!

Back slide: data format and hardware

Data Format	Host	PLD
Txt (GB) (from_id, to_id)	100	18
Binary (GB)	82	11

Processor	Intel Xeon E5-2643 3.5 GHz
Number of cores	12
Memory (GB)	750
Swap memory (GB)	250
OS	CentOS 6.4

Back slide: Common Crawl Foundation

Common Crawl is a non-profit organization dedicated to providing a copy of the internet to internet researchers, companies and individuals at no cost for the purpose of research and analysis.

The main dataset is released on a monthly basis and consists of billions of web pages stored in WARC format on Amazon S3.

Common Crawl uses Apache Nutch based web crawler to generate their datasets.

Data are stored in different three types of files:
- **WARC files store the raw crawl data.**
- **WAT files store computed metadata for the data stored in the WARC.**
- **WET files store extracted plaintext from the data stored in the WARC.**

Projects for processing data and provide final web graphs:
- **https://github.com/commoncrawl/cc-webgraph**
- **https://github.com/commoncrawl/cc-pyspark**

Back slide: statistics

Power law: $p(x)\,\mathrm{d}x = \Pr(x \le X < x + \mathrm{d}x) = Cx^{-\alpha}\,\mathrm{d}x$ C from normalization $p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha}.$

CCDF $P(x) = \Pr(X \ge x)$ $P(x) = \int_x^\infty p(x')\,\mathrm{d}x'$

Likelihood: probability that the data were drawn from the model $p(x \mid \alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha}$

To obtain α maximize the log likelihood $\mathcal{L} = \ln p(x \mid \alpha) = \ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha}$

Kolmogorov-Smirnov staistic: the maximum distance D between the CCDF of the data S(x) and P(x), that of the power-law model that best fits the data in the region $x \ge x_{min}$. The best x_{min} minimizes D.

$$D = \max_{x \ge x_{\min}} |S(x) - P(x)|$$

Goodness of fit: based on measurement of the “distance” between the distribution of the empirical data and the hypothesized model. This distance is compared with distance measurements for comparable synthetic data sets drawn from the same model, and the p-value is defined to be the fraction of the synthetic distances that are larger than the empirical distance. If p is large, then the difference between the empirical data and the model can be attributed to statistical fluctuations alone; if it is small, the model is not a plausible fit to the data.

Likelihood of two competing distributions: $L_1 = \prod_{i=1}^{n} p_1(x_i), \qquad L_2 = \prod_{i=1}^{n} p_2(x_i)$ Let R be the log of the ratio L_1/L_2

The favorite distribution is the one with higher likelihood. If p_1 is favorite than $L_1 > L_2$ and $R > 0$. If p_2 is favorite $L_2 > L_1$ and $R < 0$. If $R = 0$ there is not a favorite distribution because $L_1 = L_2$.