UNIVERSITEIT TWENTE.

# CHARACTERIZATION OF TAIL DEPENDENCE FOR IN-DEGREE AND PAGERANK

Nelly Litvak
University of Twente, The Netherlands
joint work with
Yana Volkovich (Barcelona Media),
Werner Scheinhardt (University of Twente),
Bert Zwart (CWI)

# OUTLINE

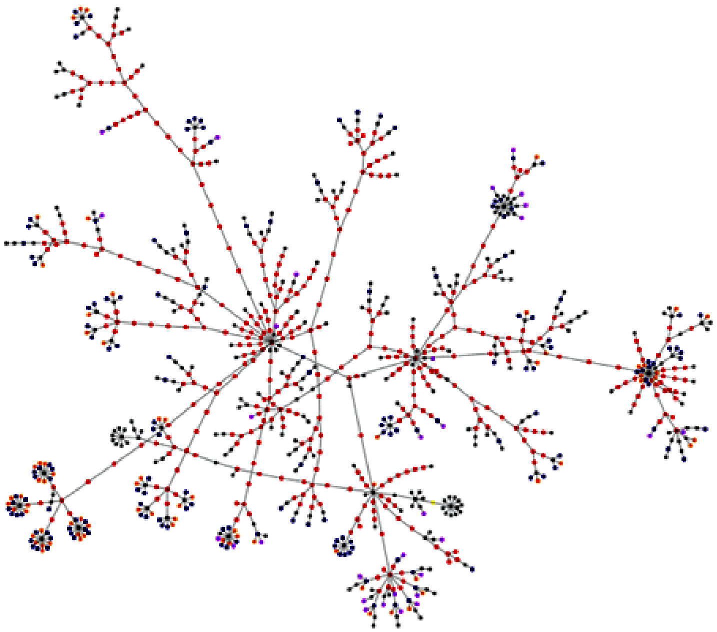
- Power laws in complex networks
- Model for power law distribution of PageRank importance scores
- Dependence between power law graph parameters, angular measure
- Analytical derivations for the angular measure
- Experiments

# COMPLEX NETWORKS

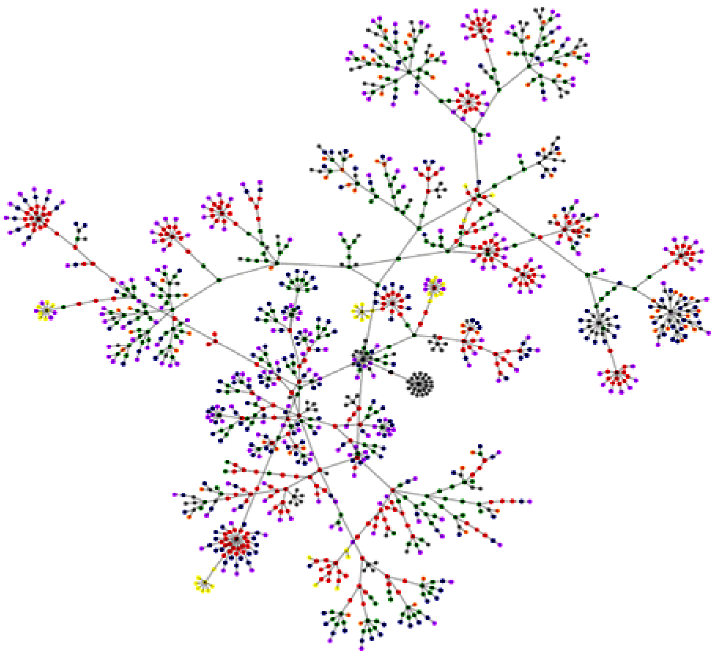
- Examples: Internet, WWW, social networks, food webs
- Typical features: high variability (power laws), clustering, small-world phenomenon, self-similarity

# POWER LAW GRAPHS

Web graphs *[http://www.aharef.info]*



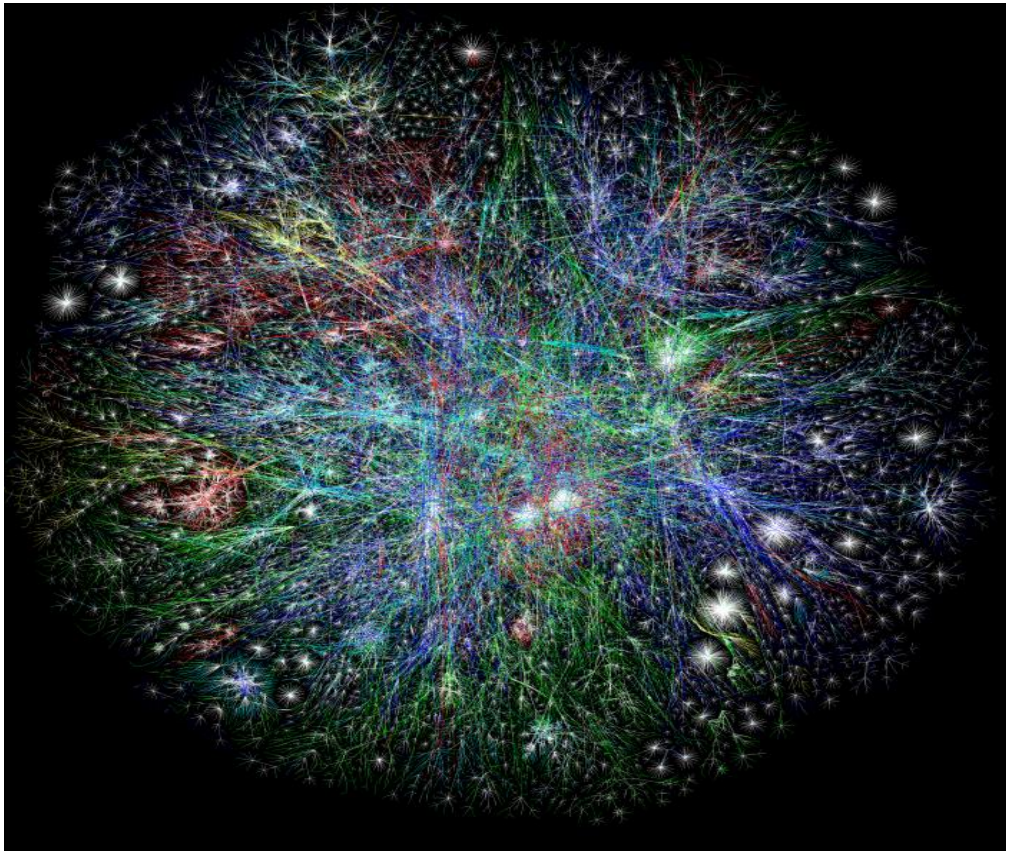
*CNN*



*Yahoo!*

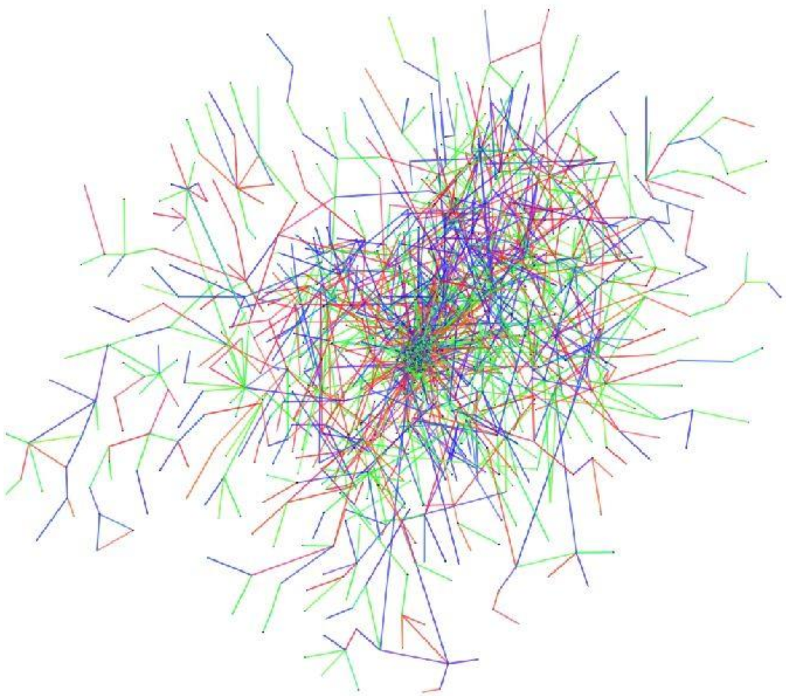
# POWER LAW GRAPHS-2

internet graphs *[http://www.opte.org]*



# POWER LAW GRAPHS-3

Social networks



Collaboration network: node=authors, edge=co-authors of a paper
source: http://www.jacobsschool.ucsd.edu/

# POWER LAW FORMALIZATION

- **Regular variation**

  X is ***regularly varying random variable*** with index $\alpha$

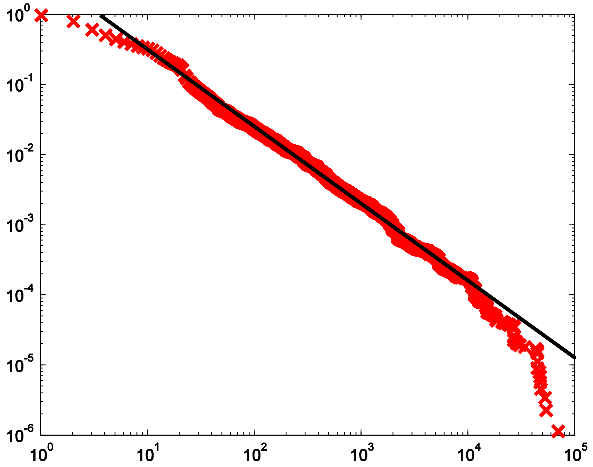
  if **$P(X>x) \sim L(x)x^{-\alpha}$** as $\mathbf{x} \to \infty$ (here a~b if a/b $\to$1)

L(x) is ***slowly varying*** if for every t>0:

$\mathbf{L(tx)/L(x)} \to \mathbf{1}$ as $\mathbf{x} \to \infty$

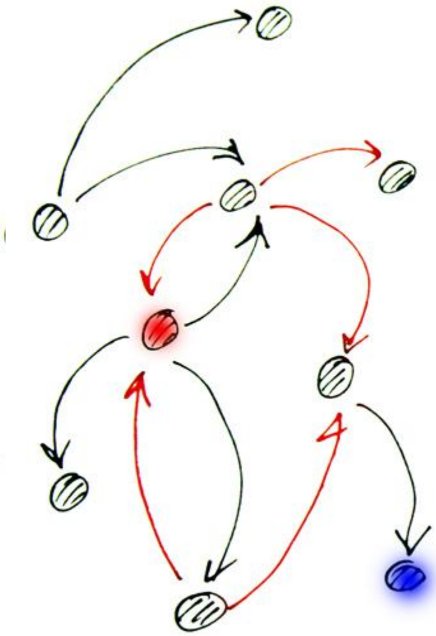
- **log-log plot:** straight line

  $$\log [P(X>x)] = - \alpha \log(x)+\log(c)$$



# GRAPH'S PARAMETERS

- In-degree (number of incoming links)
- Out-degree (number of outgoing links)
- PageRank (importance scores)



$$\boxed{PR(i) = c\sum_{j\to i}\frac{1}{d_j}PR(j) + \frac{c}{n}\sum_{j\in\mathcal{D}}PR(j) + (1-c)T(i)}$$

**$d_j$ is number of outgoing links of page j**
**c is damping factor (c=0.85)**
**n is the number of pages in the graph**
**D is the set of dangling nodes (outdegree zero)**
**T(i) probability to jump on page i**
**(classical example T(i)=1/n)**

- Broder et al. (2000) In-degree obeys power laws with $\alpha \approx 1.1$. Out-degree follows power law with exponent $\alpha \approx 1.6$
- Panduragan et al. [2002] and other authors: PageRank scaled as R(i)=nPR(i) obey power laws with $\alpha \approx 1.1$

# MOTIVATION-1

***Pandurangan et al.***[2002]: In-degree and PageRank have a similar asymptotic behavior

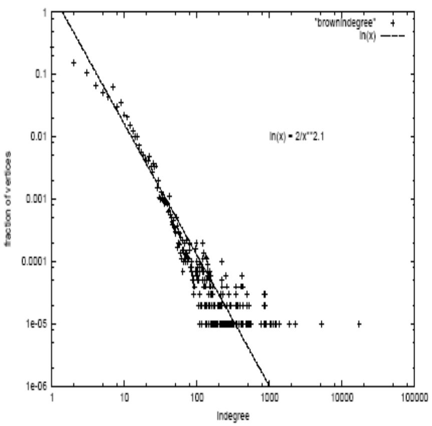



Figure 1: Log-log plot of the in-degree distribution of the Brown domain (*.brown.edu). The in-degree distribution follows a power law with exponent close to 2.1.

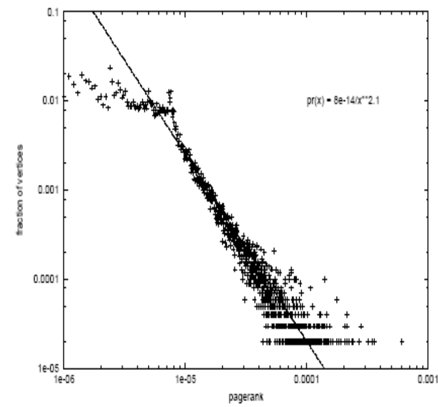



Figure 3: Log-log plot of the PageRank distribution of the Brown domain (*.brown.edu). A vast majority of the pages (except those with very low PageRank) follow a power law with exponent close to 2.1. The plot almost flattens out for pages with very low PageRank.

# STOCHASTIC EQUATION FOR PAGERANK

- PageRank definition

$$R(i) = c\sum_{j\to i}\frac{1}{d_j}R(j) + \frac{c}{n}\sum_{j\in\mathcal{D}}R(j) + (1-c)nT(i)$$

- Consider the PageRank **R** of **randomly chosen page**

PageRank R is a solution of stochastic equation

$$R = c\sum_{j=1}^{N}\frac{1}{D_j}R_j + cp_0 + (1-c)nT$$

**N is the in-degree of the randomly chosen page**

**D is the out-degree of page that links to the randomly chosen page**
**(have no restrictions on the out-degree distribution)**

**$p_0$ is the fraction of PageRank mass concentrated in the dangling nodes**

**$R_j$ is distributed as R; N, D, and $R_j$ are independent ; N and T can be dependent**

# TAIL BEHAVIOR OF R

- If **P(nT > x) = o(P(N > x)),** then **$P(R > x) \sim C_N P(N > x)$ as $x \to \infty$,**

where $$C_N = c^{\alpha_N}(1-p_0)^{\alpha_N}(E(N))^{-\alpha_N}\left[1 - c^{\alpha_N}E(N)E\left(\frac{1}{D^{\alpha_N}}\right)\right]^{-1}$$

- If **P(N > x) = o(P(nT > x)),** hen **$P(R > x) \sim C_T P(nT > x)$ as $x \to \infty$,**

where $$C_T = (1-c)^{\alpha_T}\left[1 - c^{\alpha_T}E(N)E\left(\frac{1}{D^{\alpha_T}}\right)\right]^{-1}$$

- If **$P(nT > x) \sim C_{BN}(1-c)^{\alpha_N}P(N > x)$** for some constant **$C_{BN}$**, then **$P(R > x) \sim CP(N > x)$ as $x \to \infty$,**

where
$$C = \left[C_{BN} + c^{\alpha_N}(1-p_0)^{\alpha_N}(E(N))^{-\alpha_N}\right]\left[1 - c^{\alpha_N}E(N)E\left(\frac{1}{D^{\alpha_N}}\right)\right]^{-1}$$

# DEPENDENCE IN COMPLEX NETWORKS

- How the graph parameters depend on each other?

  no agreement on the dependence between in-degree and PageRank in the Web

- The correlation coefficient

$$\boxed{\rho(X,Y) = \frac{E[(X-E(X))(Y-E(Y))]}{\sigma(X)\sigma(Y)},}$$

where **E(X)** and **E(Y)** are expected values,

**σ(X)** and **σ(Y)** are standard deviations of **X** and **Y**

# DEPENDENCE IN COMPLEX NETWORKS (CONTINUED)

We want to measure a dependence between two heavy-tailed parameters X and Y

⬇

We are mainly interested in the dependence between extremely large values of X and Y

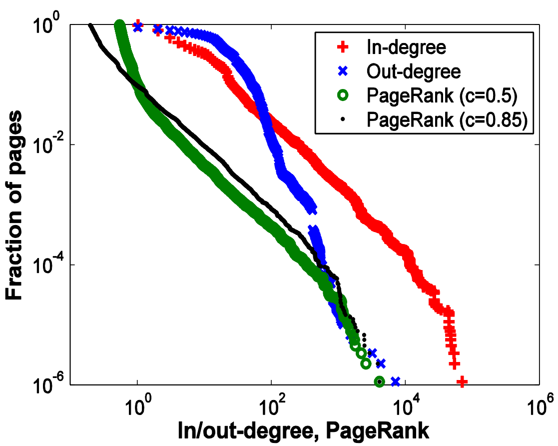
- extremal dependence is a well-developed notion of dependence that is designed for **power law tails**

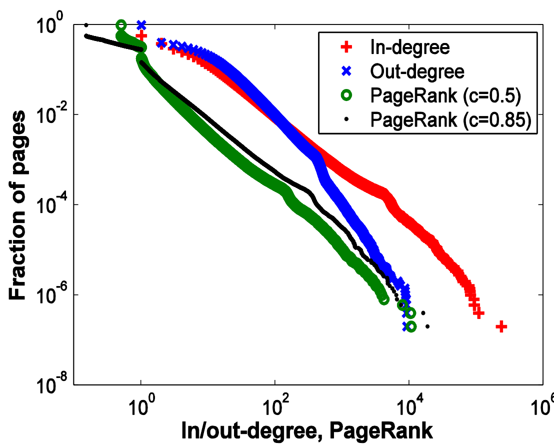
S.I. Resnick **"Heavy-tail phenomena: probabilistic and statistical modelling"**, Springer, 2007

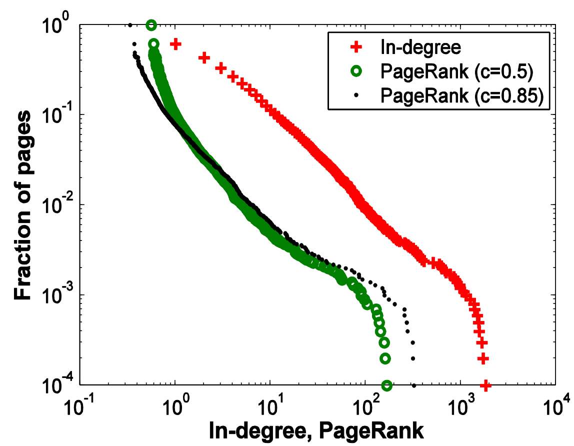
[telecommunications and mathematical finance]

# DATA SETS

**Eu-2005** contains 862.664 nodes and 19.235.140 links
**Wikipedia** contains 4.881.983 nodes and 42.062.836 links
**Growing Network** contains 10.000 nodes and 79.992 links




x-axes: values of parameter; y-axes: proportion of pages for which this parameter is greater than x

# MATHEMATICAL FRAMEWORK

- X,Y are r.v's; $F_X$,$F_Y$ are distribution functions
- 1-$F_X$(X) = fraction of occurrences of the value >X (rank)
- P(1-$F_X$(X) ≤ 1/t)=1/t, and if t is large then X is large
- Define

$$\boxed{(\mathrm{R},\Theta) = \mathrm{POLAR}\left(\frac{1}{1-\mathrm{F_X(X)}}, \frac{1}{1-\mathrm{F_Y(Y)}}\right)}$$

where

$$\mathrm{POLAR(x,y)} = \left(\| (x,y) \|, \Theta\right)$$

- Then $\lim_{t\to\infty} \mathrm{tP(R > t, \Theta \in A) = S(A)}$
- S(A) is the angular measure
- Independence: R is large if X or Y is large

Dependence: X and Y can be large together

# INTERPRETATION OF THE ANGULAR MEASURE

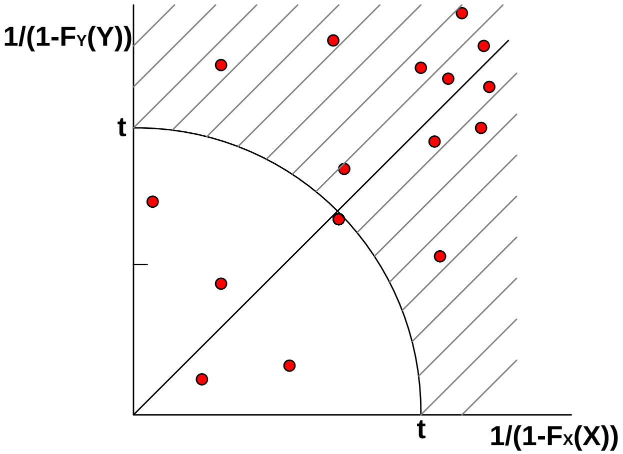
- Define $\left(R_{j,k}, \Theta_{j,k}\right) = \mathrm{POLAR}\left(\dfrac{1}{1-F_X(X)}, \dfrac{1}{1-F_Y(Y)}\right)$
- Angular measure S(A): $\lim_{t\to\infty} tP(R > t, \Theta \in A) = S(A)$

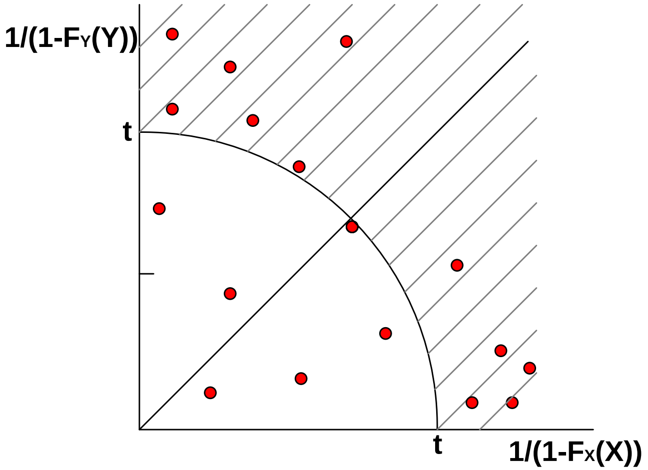
Dependence

(measure is concentrated around **π/4)**

Independence

**(**measure is concentrated around **0** and **π/2)**



# STATISTICAL DEPENDENCIES

- graph's parameters:

$$X= (X_1,\ldots, X_n) \text{ and } Y= (Y_1,\ldots, Y_n)$$

$$\text{node } j \to (X_j ,Y_j)$$

- **rank transform**

$$\{(X_j ,Y_j), 1\leq j \leq n\} \longrightarrow \{(r_j^x , r_j^y ), 1\leq j \leq n\},$$

$r_j^x$ is the descending rank of $X_j$ in $(X_1,\ldots, X_n)$

$r_j^y$ is the descending rank of $Y_j$ in $(Y_1,\ldots, Y_n)$

# POLAR COORDINATE TRANSFORM

## Polar coordinate transform

k=1,…n  is the number of upper statistics

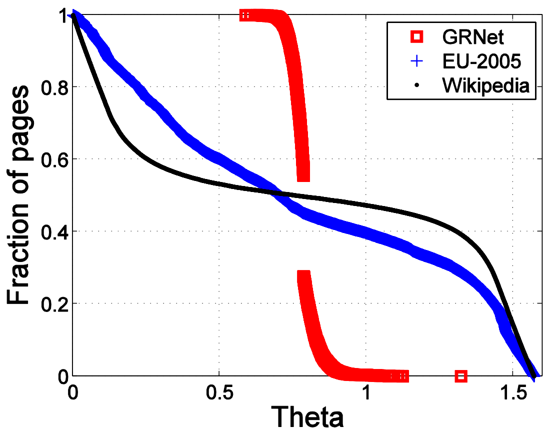
$$\mathrm{POLAR}\left(\frac{k}{r_j^x}, \frac{k}{r_j^y}\right) = \left(R_{j,k}, \Theta_{j,k}\right)$$

where $$\mathrm{POLAR}(x, y) = \left(\sqrt{x^2 + y^2}, \arctan(y/x)\right)$$

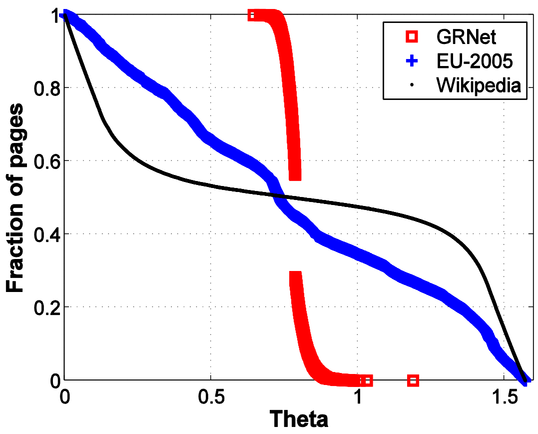
- empirical distribution of **Θ** for **k** largest values of **R**
- cumulative distribution function **$\{\Theta_{j,k} : R_{j,k} > 1\}$**

# DEPENDENCIES

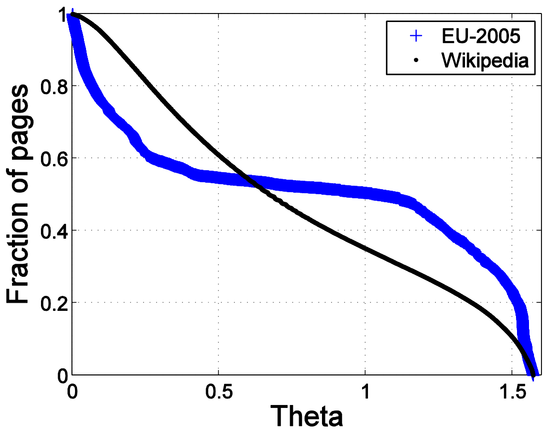
in-degree and PageRank (c=0.85)



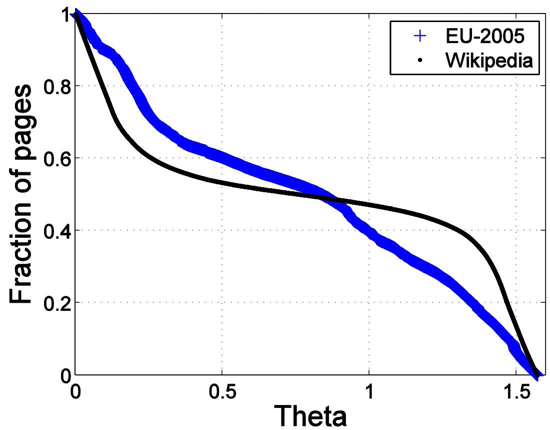

in-degree and PageRank (c=0.5)




in-degree and out-degree (c=0.85)




out-degree and PageRank (c=0.85)



# ANGULAR MEASURE: ANALYTICAL DERIVATION

- Our stochastic model

$$\boxed{R =^{d} \sum_{j=1}^{N} A_j R_j + B}$$

- N is regularly varying, **P(N>x) ~ L(x)x$^{-\alpha}$**

- Can we analytically determine the angular measure between N and R?

- It turns out that this can be done with the results from the extreme value theory:

Beirlant, Goegebeur, Segers, Teugels (2004): Statistics of Extremes: Theory and Applications

# INSIGHT THROUGH THE LAW OF LARGE NUMBERS

$$\boxed{R =^{d} \sum_{j=1}^{N} A_j R_j + B}$$

where $P(N>x) \sim L(x)x^{-\alpha}$

**Lemma.** As $u \to \infty$, for any constant $K>0$,

$$P(N > u, R > Ku) \sim \min\{1, [E(A)/K]^{\alpha}\} P(N > u)$$

\`Proof': By the SLLN we have $R \sim E(A)N$ when N is large.

- when $E(A) > K$, the event $\{R>Ku\}$ is \`implied' by $\{N>u\}$;
- when $E(A) < K$, then N needs to be larger than $Ku/E(A)$ for $\{R>Ku\}$ to hold.

# TAIL DEPENDENCE FUNCTION

- Remember that the angular measure is defined as

$$\lim_{t\to\infty} tP(R > t, \Theta \in A) = S(A)$$

where

$$\boxed{(R,\Theta) = \mathrm{POLAR}\left(\frac{1}{1-F_X(X)}, \frac{1}{1-F_Y(Y)}\right)}$$

- Using that P(R>u)~CP(N>u) for large u, we can compute the tail dependence function (which is closely related to a copula):

$$r(x,y) = \lim_{t\to 0} t^{-1} P((1-F_N(N)) \le tx, (1-F_R(R)) \le ty)$$

- There is a one-to-one correspondence between S(A) and r(x,y)

## DERIVATION OF THE DEPENDENCE FUNCTION

**Theorem.** Dependence function between N and R is:

$$r(x, y) = \min\{x, y[E(A)]^{\alpha} / C\}$$

**Proof.** By rewriting r(x,y) in the form as in the Lemma and then using the Lemma.

$$\begin{aligned}
&P(\bar{F}_1(N) \le tx, \bar{F}_2(R) \le ty) \\
&\quad = P(N \ge \bar{F}_1^{-1}(tx), R \ge \bar{F}_2^{-1}(ty)) \\
&\quad = P\left(N \ge \bar{F}_1^{-1}(tx), R \ge \left(\frac{y}{Kx}\frac{L(\bar{F}_1^{-1}(tx))}{L(\bar{F}_2^{-1}(ty))}\right)^{-1/\alpha} \bar{F}_1^{-1}(tx)\right) \\
&\quad \sim P\left(N \ge \bar{F}_1^{-1}(tx), R \ge \left(\frac{y}{Kx}\right)^{-1/\alpha} \bar{F}_1^{-1}(tx)\right)
\end{aligned}$$

# BACKGROUND FROM THE EXTREME VALUE THEORY

Choose any norm $||\cdot||$, then a unique (nonnegative) measure $H(\cdot)$ exists on $\Xi = \{\omega \in \mathbb{R}_+^2 \;:\; ||\omega|| = 1\}$, such that

$$r(x, y) = \int_\Xi \min(\omega_1 x, \omega_2 y) H(d\omega).$$

Normalization:

$$\int_\Xi \omega_1 H(d\omega) = \int_\Xi \omega_2 H(d\omega) = 1,$$

# THE ANGULAR MEASURE IN $L_1$

- Denote by H(.) the angular measure in $L_1$ –norm
- From the extreme value theory we have:

$$\boxed{r(x,y)=\int_0^1 \min\{wx,(1-w)y\}H(dw)}$$

with normalization

$$\int_0^1 \omega H(d\omega)=\int_0^1 (1-\omega)H(d\omega)=1 \Rightarrow \int_0^1 H(d\omega)=2$$

- It is easy to check that the following two-point measure satisfies the conditions above and corresponds to the obtained r(x,y):

$$\boxed{\begin{aligned} &H(0)=1-\frac{[E(A)]^{\alpha}}{C} \text{ in } 0,\\ &H(a)=1+\frac{[E(A)]^{\alpha}}{C} \text{ in } a=\frac{C}{C+[E(A)]^{\alpha}} \end{aligned}}$$

# INTERPRETATION OF THE ANGULAR MEASURE

- The model:

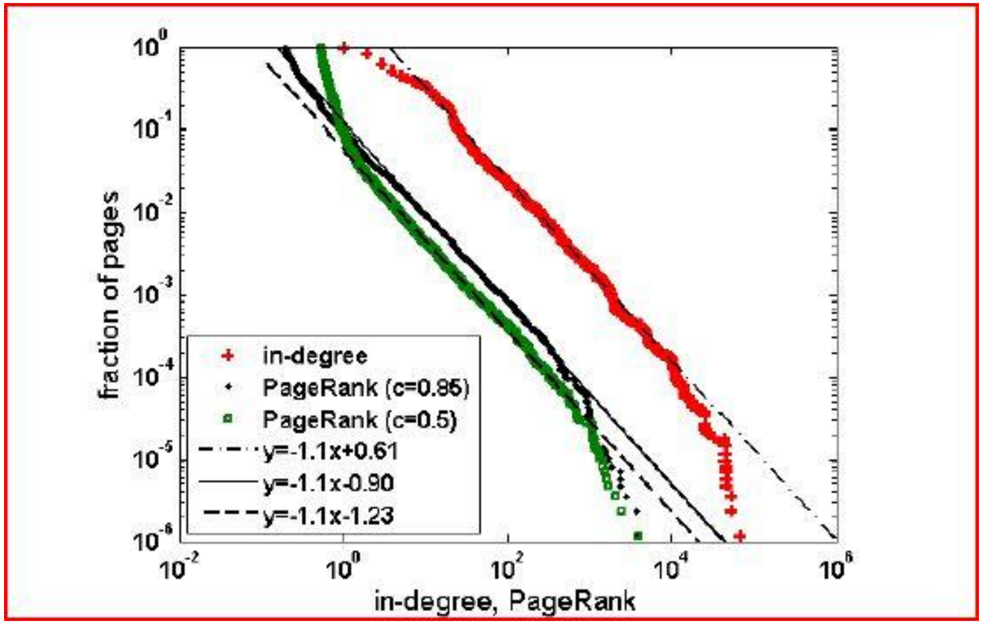
$$\boxed{R \stackrel{d}{=} \sum_{j=1}^{N} A_j R_j + B}$$

- H is the dependence measure between in-degree N and PageRank R
- The total weight of H(.) is 2
- H is concentrated in two points: 0 and a
- Interpretation:

  - fraction H(a)/2 of pages has large PR due to large in-degree
  - fraction H(0)/2 of pages has a high PR due to important links

# EXPERIMENTS: WEB

- EU-2005 data set due to the Laboratory for Web Algorithmics (LAW) of the Universit`a degli studi di Milano, Boldi and Vigna (2004)
- Total of 862,664 nodes and 19,235,140 links
- Fitting gives α =1.1, both for In-degree and PageRanks, see log-log plots (with c=0.85 and c=0.5):

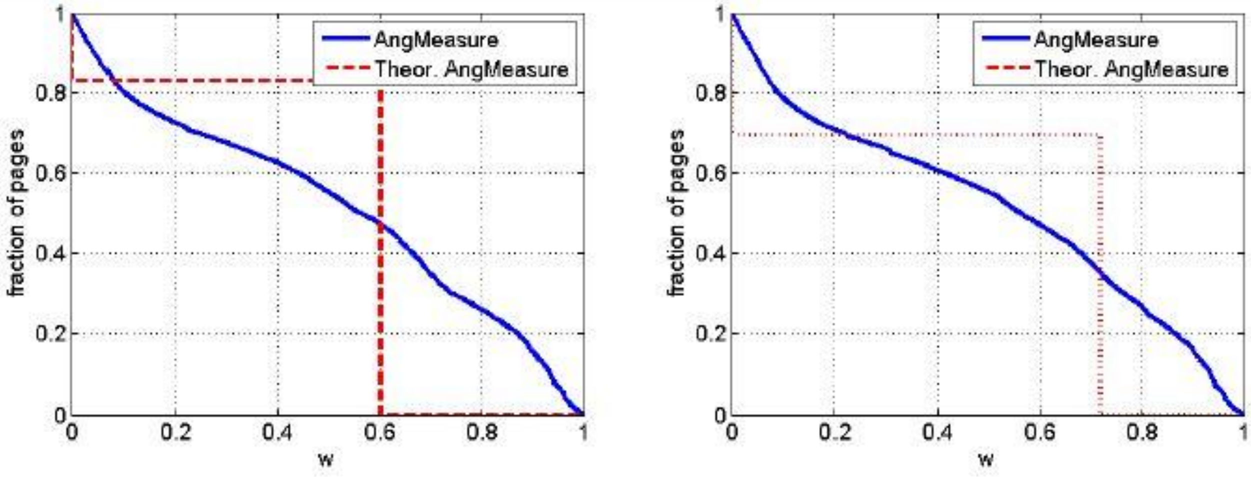


# EXPERIMENTAL RESULTS: WEB

- Theoretical result for a two-point measure:

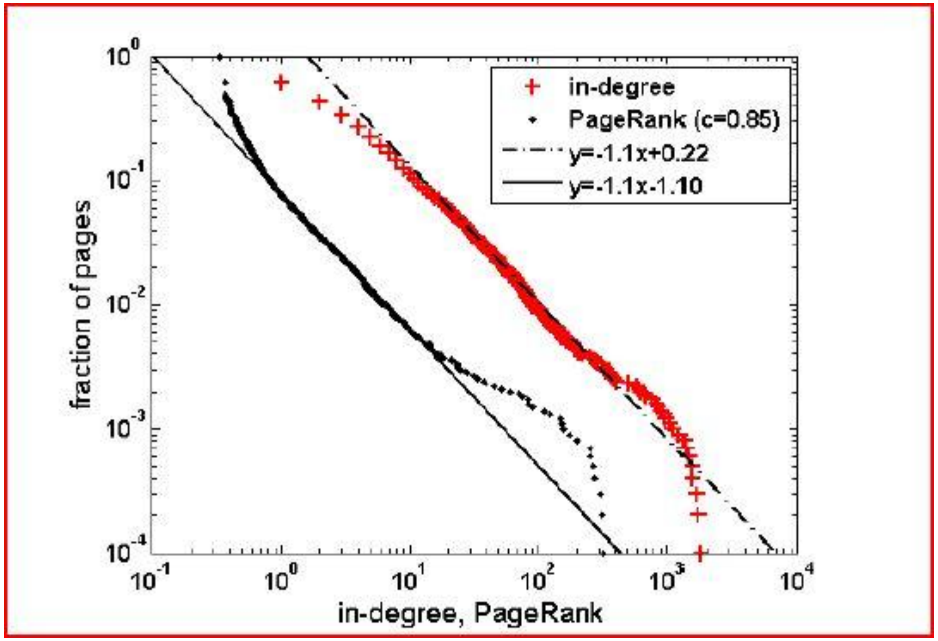
| $c$ | $a_c$ | $H(a_c)/2$ |
|---|---|---|
| 0.5 | 0.6031 | 0.8290 |
| 0.85 | 0.7210 | 0.6934 |

- Experimental comparison…



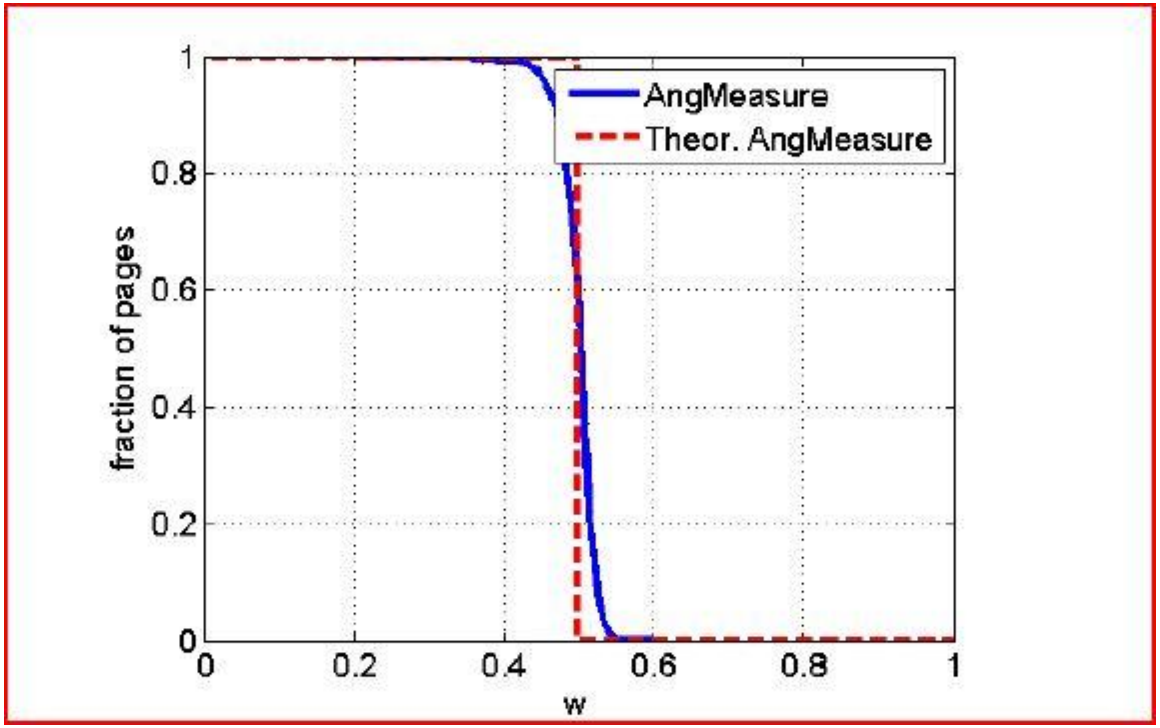
# EXPERIMENTS: GROWING NETWORKS

- Network of 10.000 nodes with constant out-degree d = 8
- With prob. 0.1, new node links to random page, with 0.9, new node follows preferential attachment
- Fitting gives α=1.1, both for In-degree and PageRank



# EXPERIMENTAL RESULTS: GROWING NETWORKS

- Assuming $P(R_i > u) = o(P(N > u))$ we find that $H(.)$ is a one-point measure:
- $a = 1/2$, $H(a) = 2$



# SUMMARY

- We propose a new approach to modeling and analysing the relations between various parameters of complex networks
- Extremal dependencies reveal that Web, Wikipedia and preferential attachment graphs have a totally different dependence structure between different graph parameters
- Our stochastic model is too rough to capture the dependencies in the Web

# THANK YOU!

?