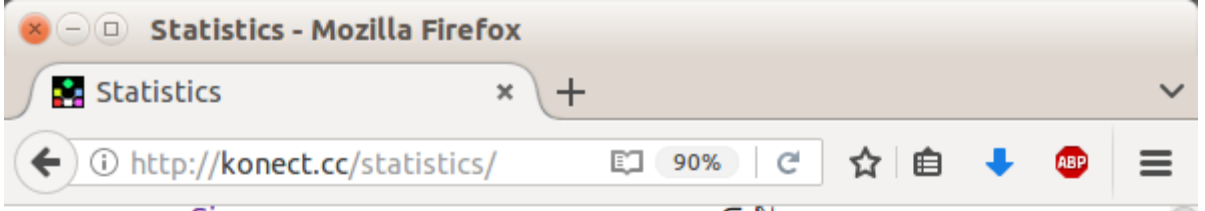
Network Statistics

- A statistic is a real number that characterizes a network
- Examples:
  - Average degree (d)
  - Number of triangles (t)
  - Diameter (δ)
  - Clustering coefficient (c)
  - Gini coefficient of degree distribution (G)
  - Degree assortativity (ρ)

More Statistics

- Number of wegdes (s)
- Number of squares (q)
- Number of claws (z)
- Number of crosses (x)
- Maximum degree (d_{max})
- Relative maximum degree ($d_{MR} = d_{max} / d$)
- Number of degree-1 nodes (d_1)
- 50-percentile effective diameter ($\delta_{0.5}$)
- Relative edge distribution entropy (H_{er})
- Bipartivity ($b_A = 1 – \lambda_{min}[A] / \lambda_{max}[A]$)
- Normalized two-star count ($s_d = s / (n\ d\ (d – 1) / 2)$)
- Eigenvalues of certain matrices ($a = \lambda_2[L]$, $|\lambda_{max}[A]|$, …)
- etc.



n =	Size	$\in \mathbb{N}$
m =	Volume	$\in \mathbb{N}$
$\bar{m}$ =	Unique edge count	$\in \mathbb{N}$
d =	Average degree	$\in \mathbb{R}^+$
d_{max} =	Maximum degree	$\in \mathbb{N}$
p =	Fill	$\in [0, 1]$
$\tilde{m}$ =	Average edge multiplicity	$\in \mathbb{R}^+$
ζ =	Negativity	$\in [0, 1]$
s =	Wedge count	$\in \mathbb{N}$
z =	Claw count	$\in \mathbb{N}$
x =	Cross count	$\in \mathbb{N}$
N =	Size of LCC	$\in \mathbb{N}$
N_s =	Size of LSCC	$\in \mathbb{N}$
ρ =	Degree assortativity	$\in [-1, +1]$
$\rho^{\pm}$ =	In/outdegree correlation	$\in [-1, +1]$
$\|A\|_2$ =	Spectral norm	$\in \mathbb{R}^+$
G =	Gini coefficient	$\in [0, 1]$
γ =	Power law exponent	$\in \mathbb{R}^+$
γ_t =	Tail power law exponent	$\in \mathbb{R}^+$
H_{er} =	Relative edge distribution entropy	$\in [0, 1]$
c =	Clustering coefficient	$\in [0, 1]$
t =	Triangle count	$\in \mathbb{N}$
δ =	Diameter	$\in \mathbb{N}$
$\delta_{0.5}$ =	50-Percentile effective diameter	$\in \mathbb{R}^+$
$\delta_{0.9}$ =	90-Percentile effective diameter	$\in \mathbb{R}^+$
δ_m =	Mean distance	$\in \mathbb{R}^+$
y =	Reciprocity	$\in [0, 1]$
T_4 =	4-Tour count	$\in \mathbb{N}$
q =	Square count	$\in \mathbb{N}$
a =	Algebraic connectivity	$\in \mathbb{R}^+$

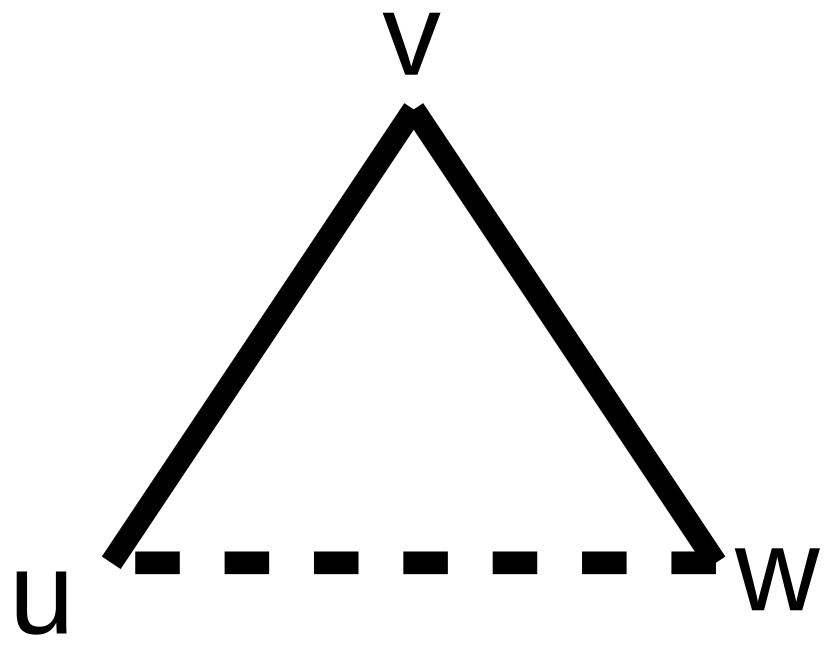
http://konect.cc/statistics/

Clustering Coefficient

$$c = 3t / s$$

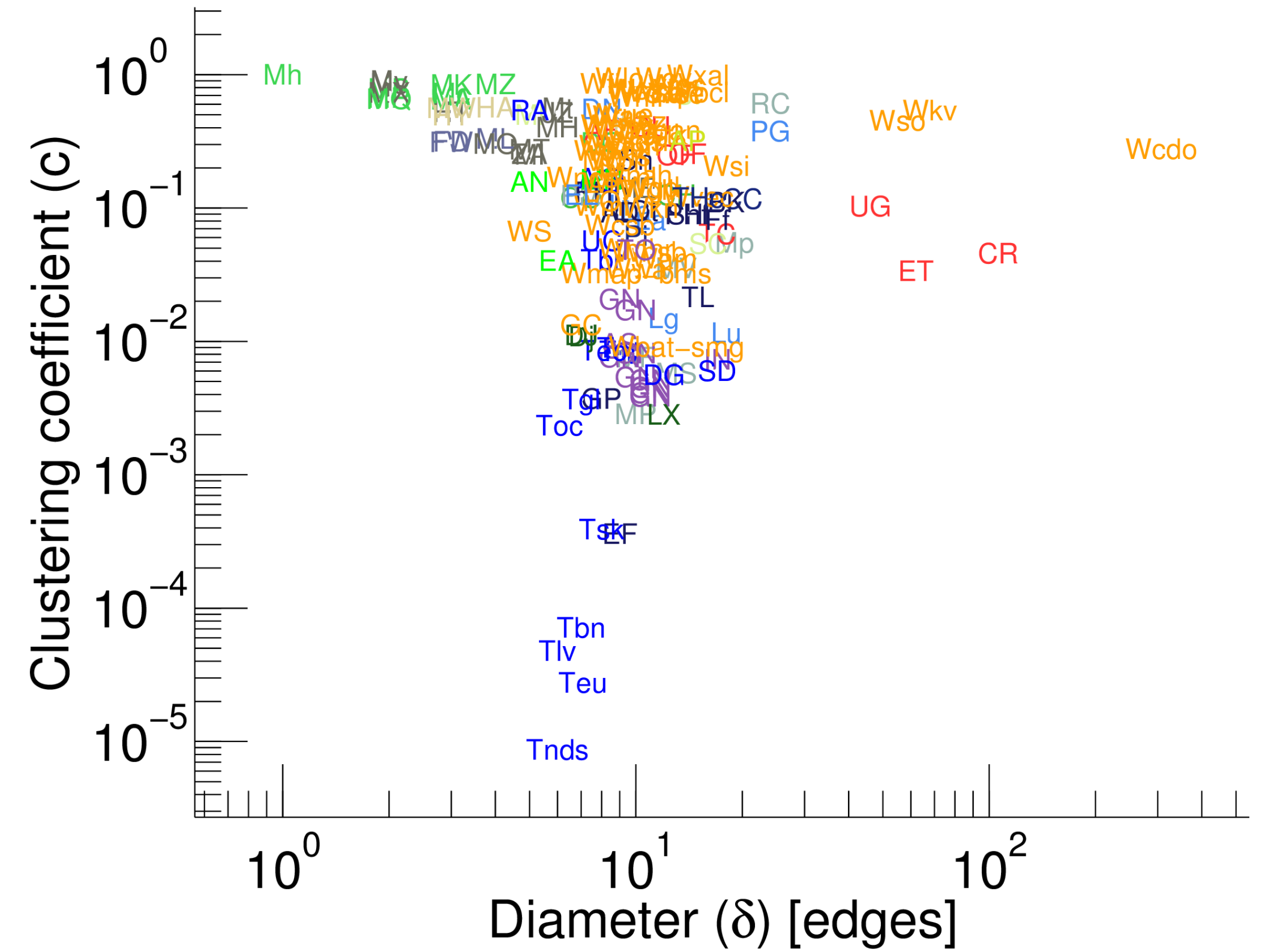
s : number of wedges

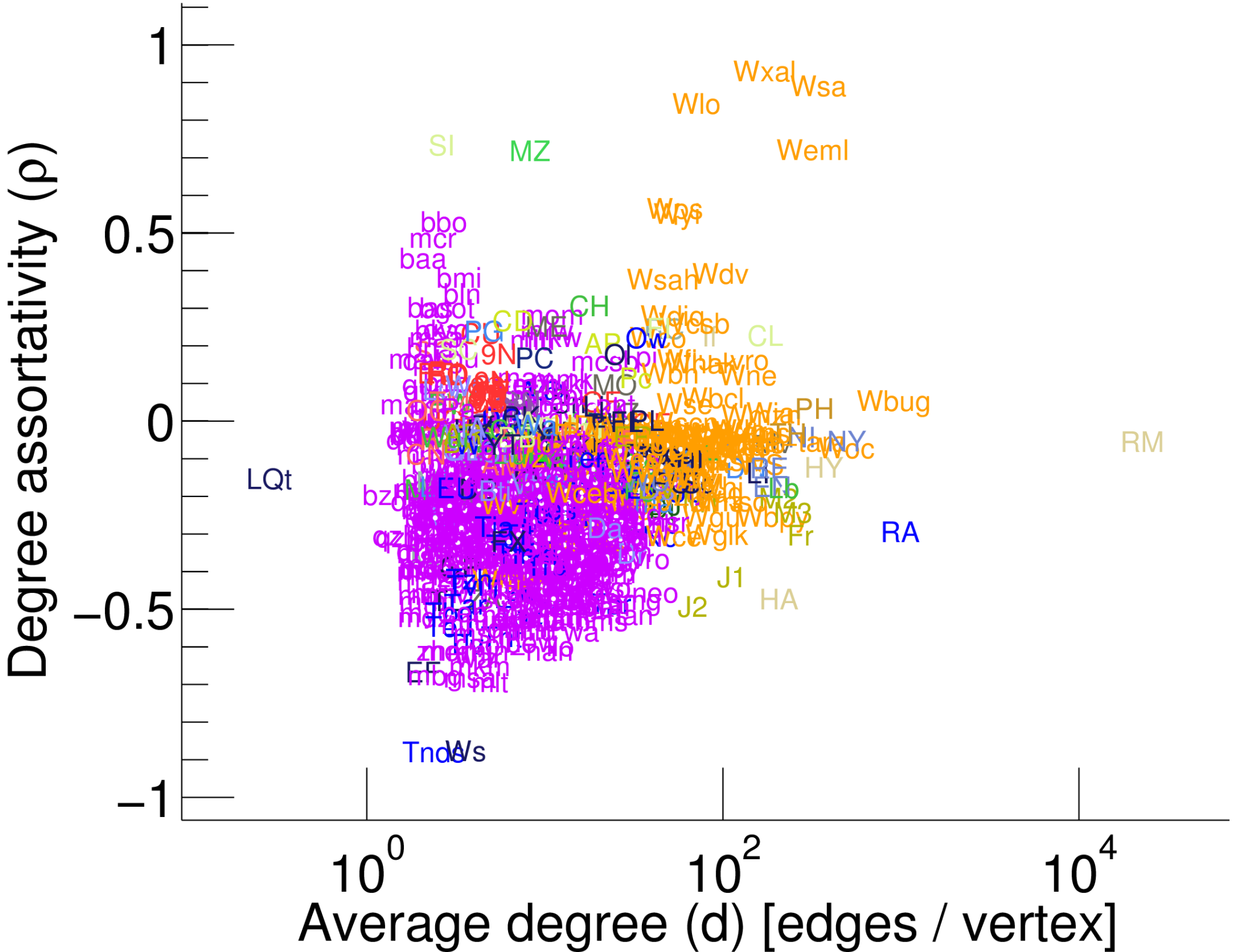
t : number of triangles

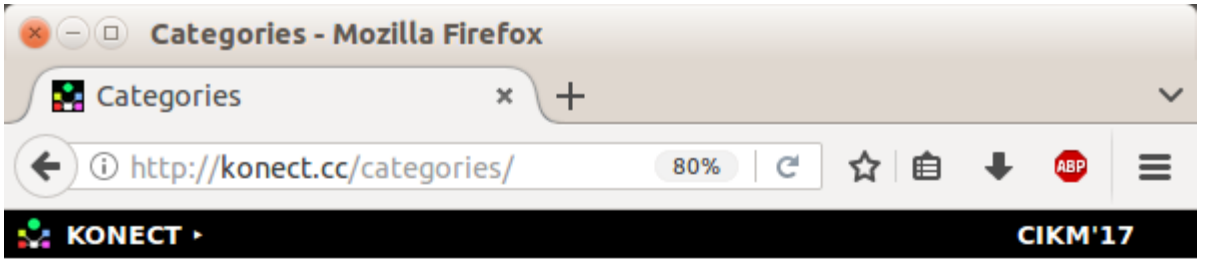


$$c = P(uw \mid uv \wedge vw)$$

http://konect.cc/statistics/clusco/


Clustering coefficient (c)
Diameter (δ) [edges]
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^0
10^1
10^2
Mh
MK
MZ
RC
PG
Wso
Wkv
Wcdo
Wsi
UG
CR
ET
WS
EA
TL
Lg
Lu
SD
Toc
Tbn
Tlv
Teu
Tnds


Degree assortativity (ρ)
1
0.5
0
−0.5
−1
Average degree (d) [edges / vertex]
10^0
10^2
10^4
Wxal
Wsa
Wlo
Sl
MZ
Weml
bbo
mcr
baa
bmi
Wsah
Wdv
CH
PC
CL
Wbn
Wne
Wbcl
Wbug
PH
NY
RM
LQt
HY
RA
Fr
J1
HA
J2
Tnde
Ws



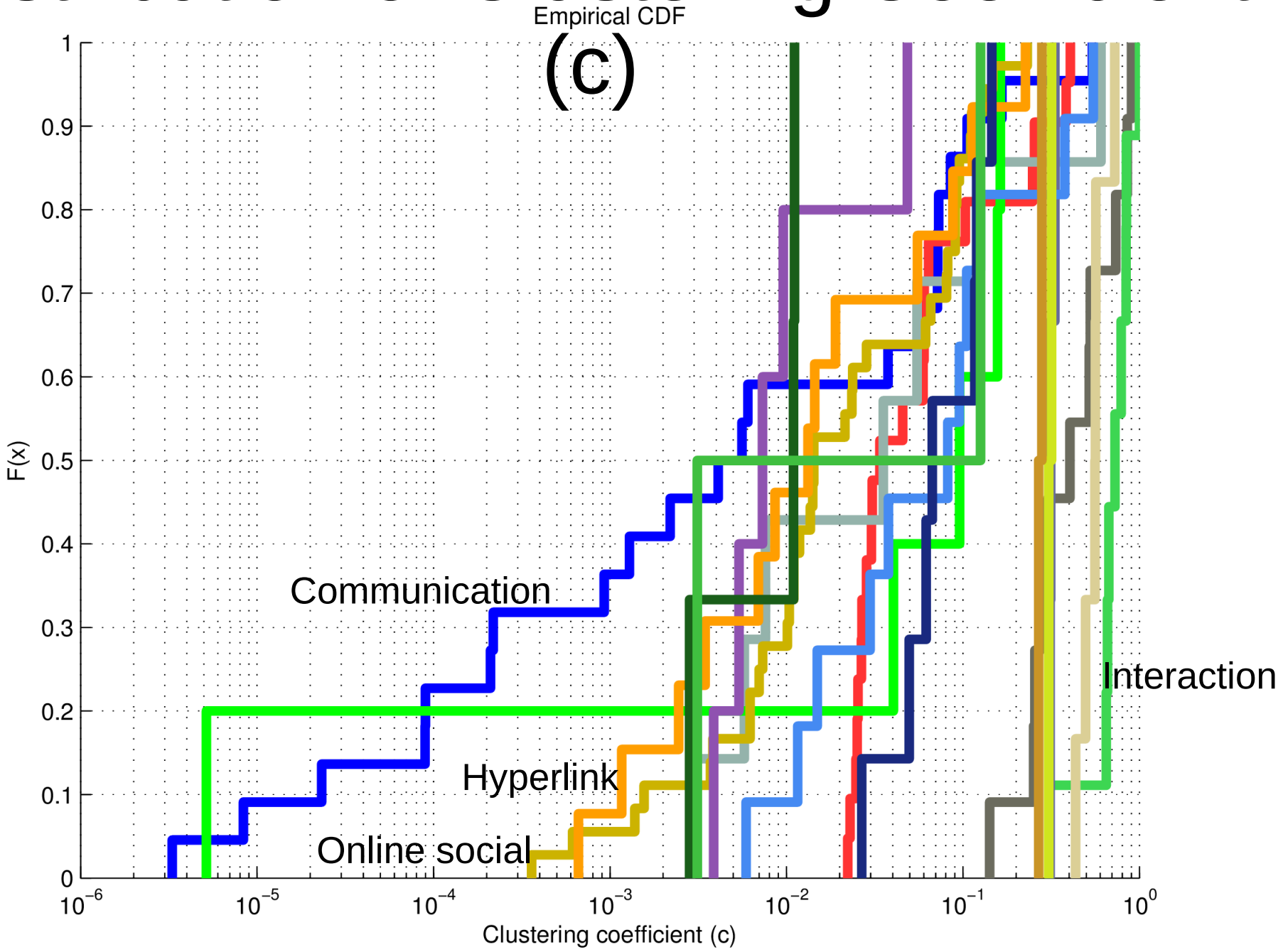
Categories

In KONECT, each network belongs to a **category** which denotes its semantics.
Examples of categories are social networks, road networks, and citation networks.

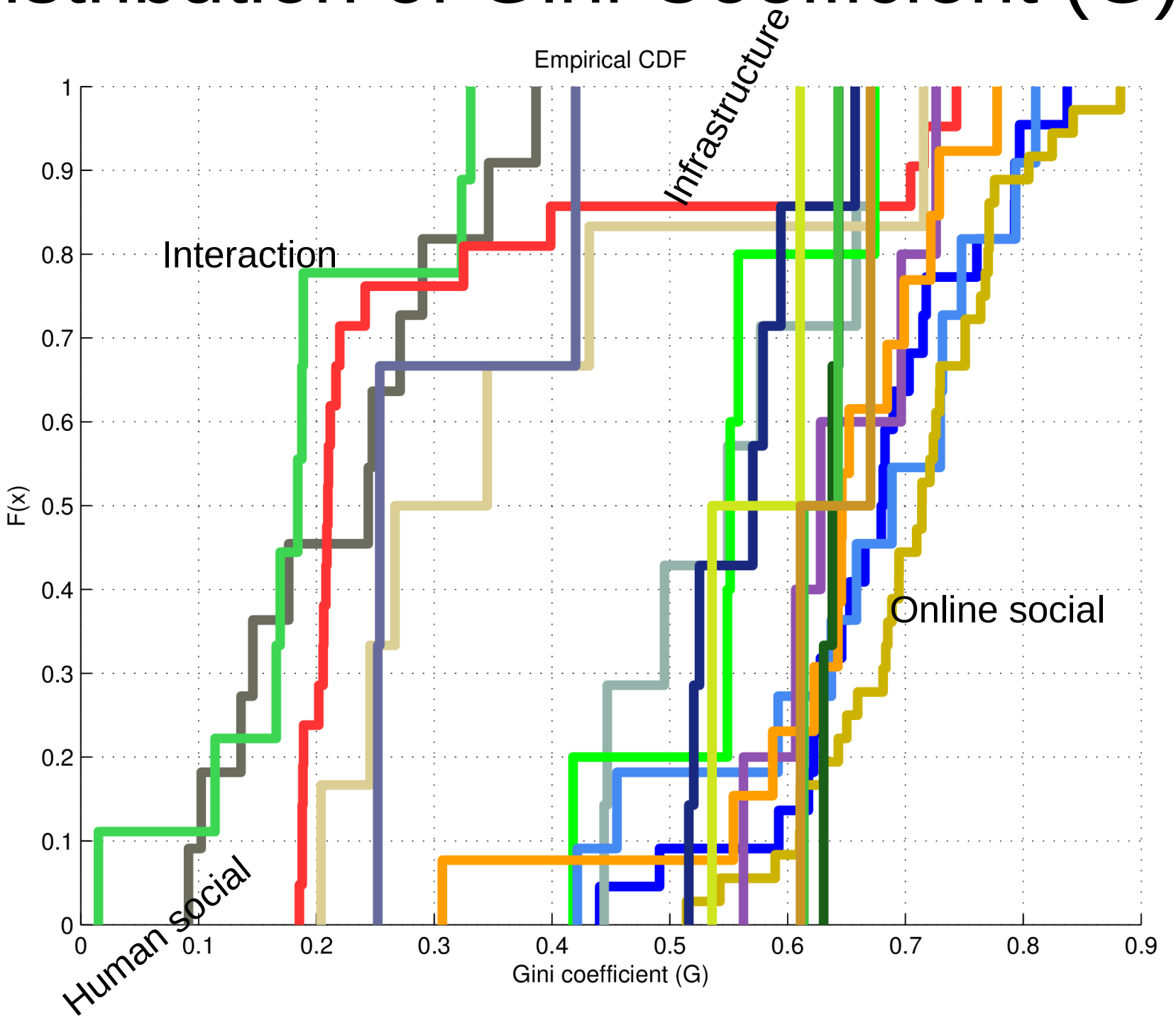
Category	Count
Affiliation	17
Animal	9
Authorship	809
Citation	7
Co-authorship	3
Co-citation	2
Communication	42
Computer	13
Feature	17
Human contact	5
Human social	12
Hyperlink	191
Infrastructure	23
Interaction	25
Lexical	5
Metabolic	7
Miscellaneous	12
Online contact	15
Online social	46
Rating	15
Software	3
Text	10
Trophic	3

http://konect.cc/categories/

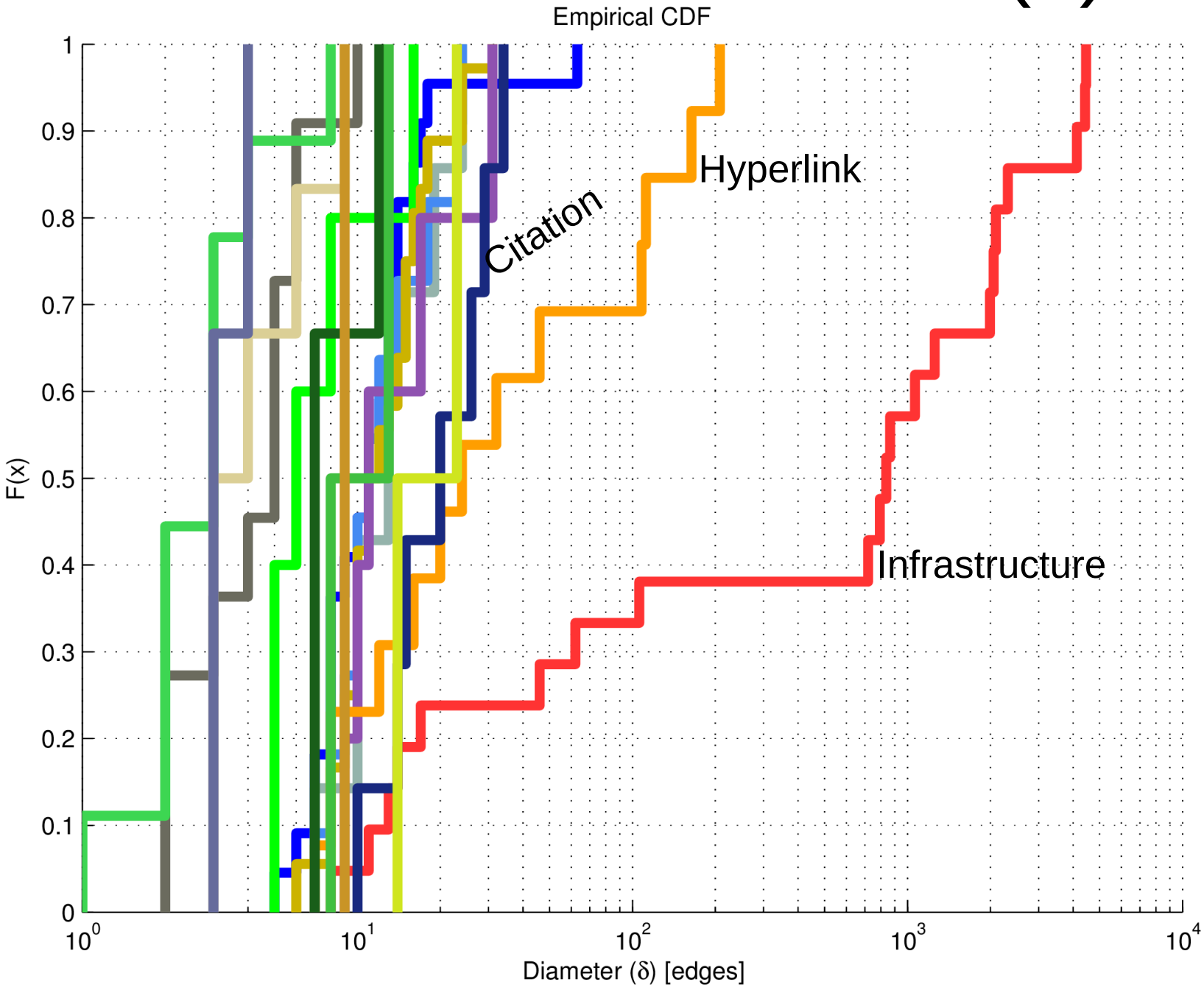
Distribution of Clustering Coefficient



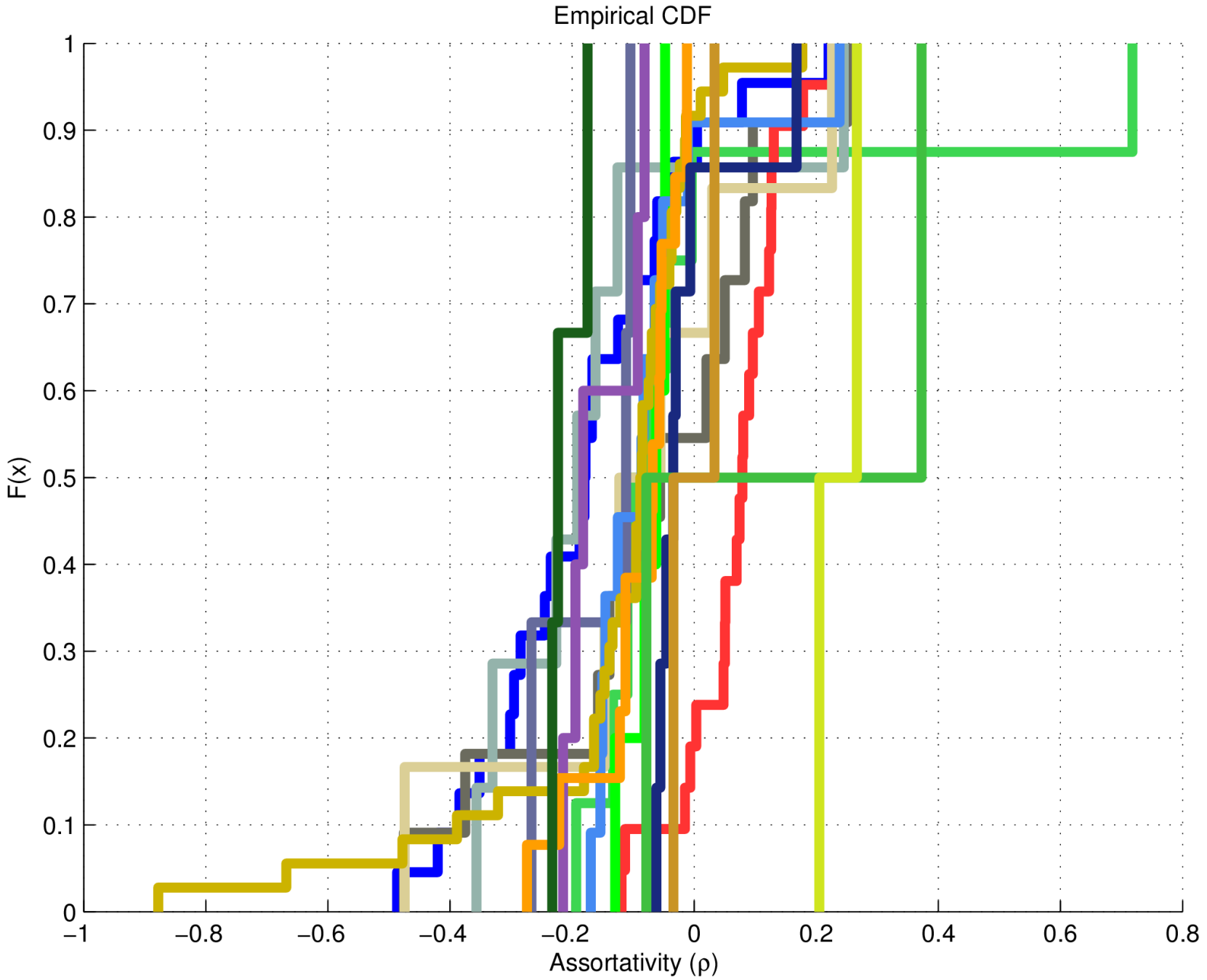
Distribution of Gini Coefficient (G)



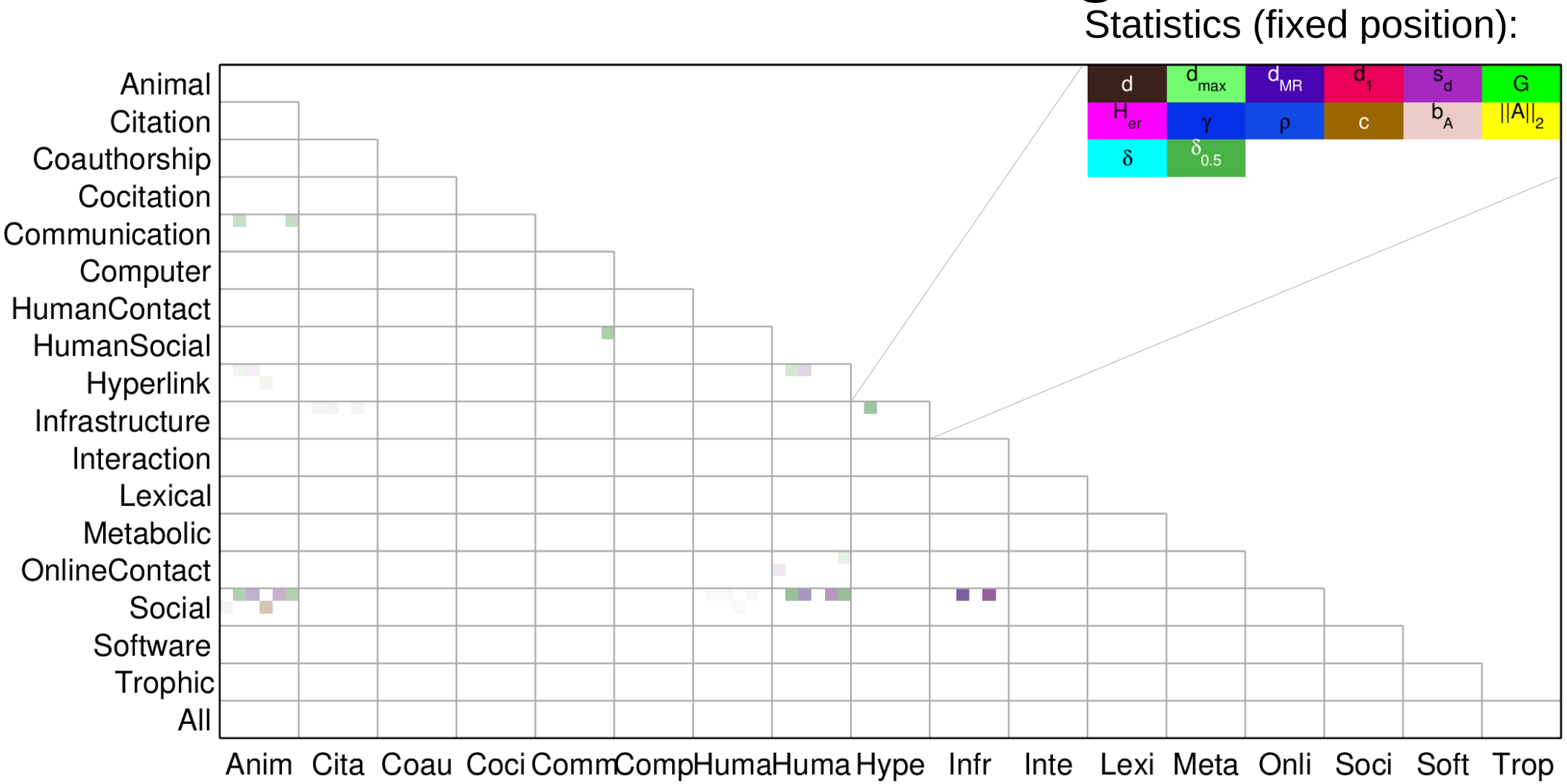
Distribution of Diameter (δ)



Degree Assortativity (ρ)

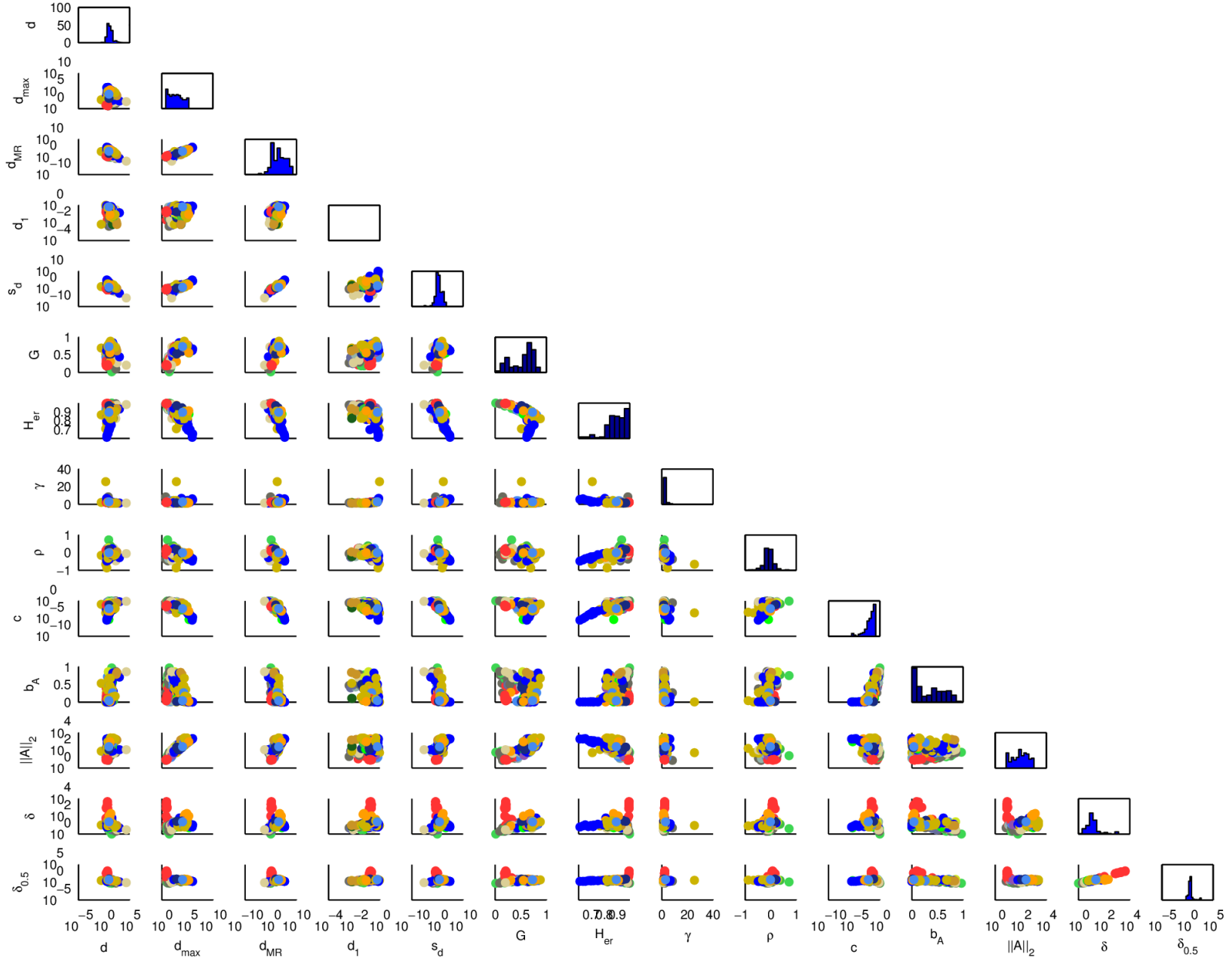


Statistical Testing

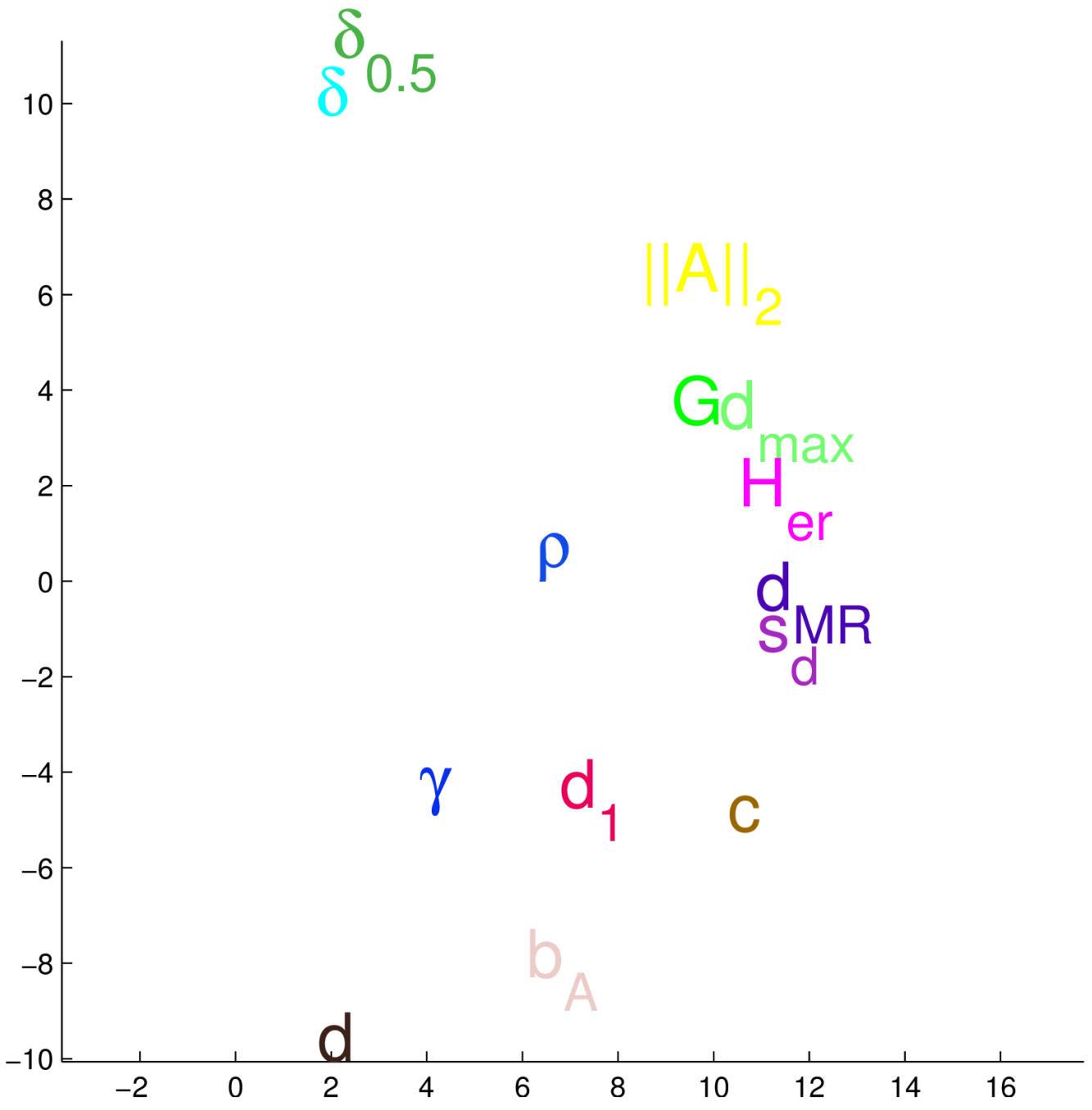


Kolmogorov–Smirnov test on each pair of categories; non-white cell when statistic is significantly different ($p < 0.10$). Base colour by HSL: Hue denotes network statistic; S & L is constant. Shown colour is interpolated between base colour and white for $0 \leq p \leq 0.10$.

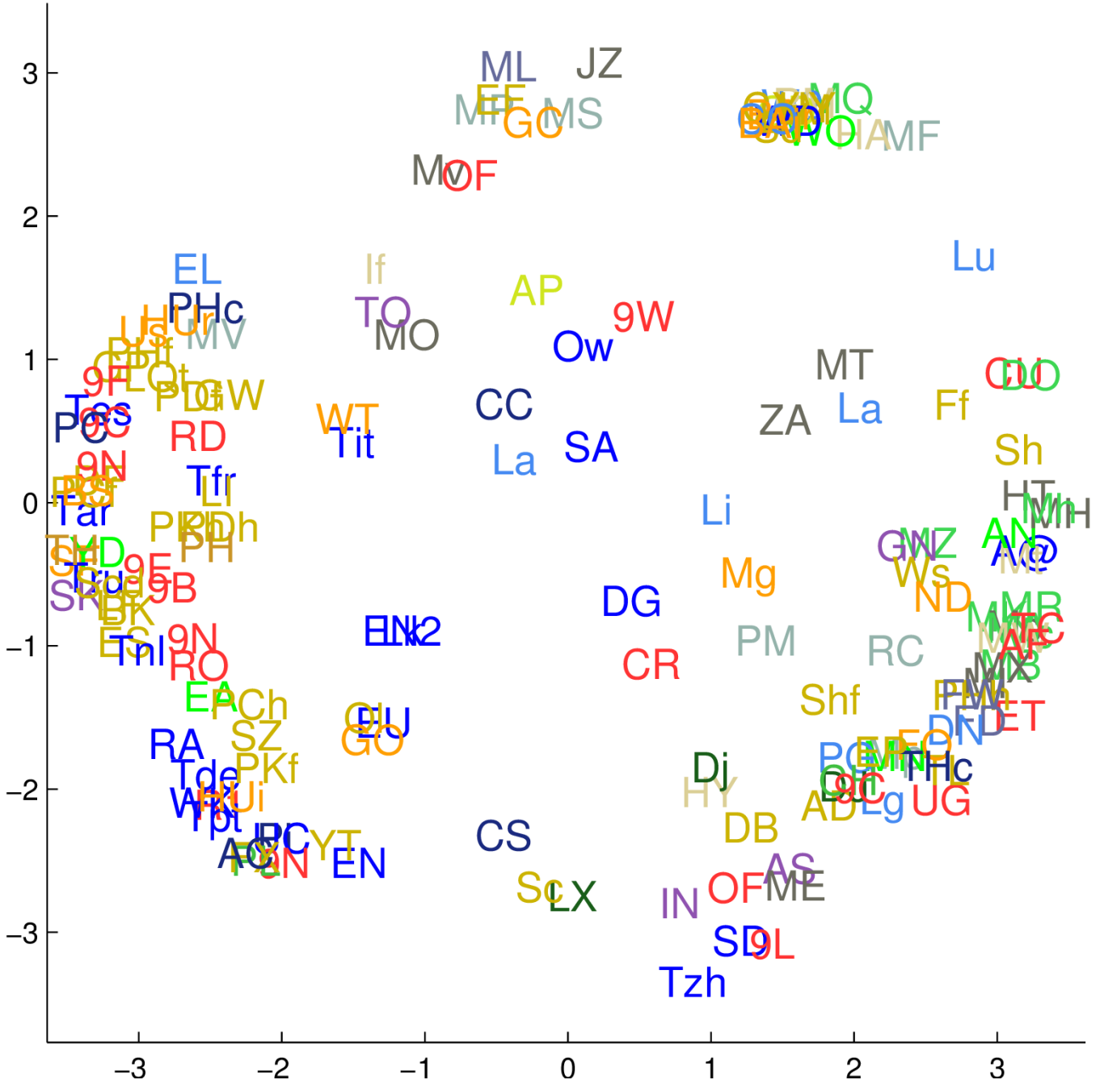
Statistics Are Not Uncorrelated



Principal Component Analysis of Statistics



PCA of Network Datasets



Feature Engineering

- Find size-independent formulations of statistics
  - E.g., c instead of t
- Avoid highly correlated statistics
  - E.g., keep only one of G and P
- Find statistics that are easy to compute
  - E.g., algebraic connectivity (a) needs $O(n^2)$ runtime

Related Work (In Progress)

- Characterizing the structural diversity of complex networks across domains, Kansuke Ikehara, Aaron Clauset

[https://arxiv.org/abs/1710.11304]