



Using graphlets to study the large scale structure of complex networks

Venkatesh. S*

M.Sc. Physics
(Batch of 2019)

Khasim Kallungal

PhD. Physics

Dr. T.K. Shajahan

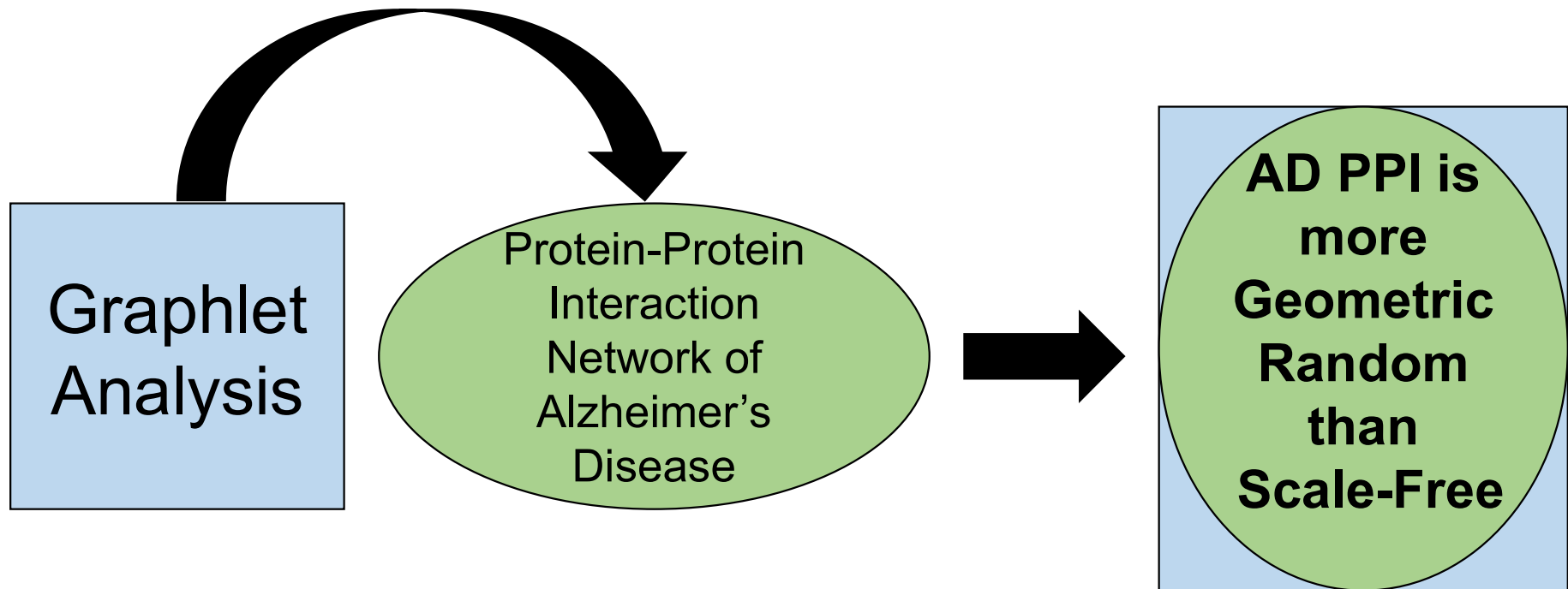
Assistant Professor

Department of Physics, National Institute of Technology Karnataka

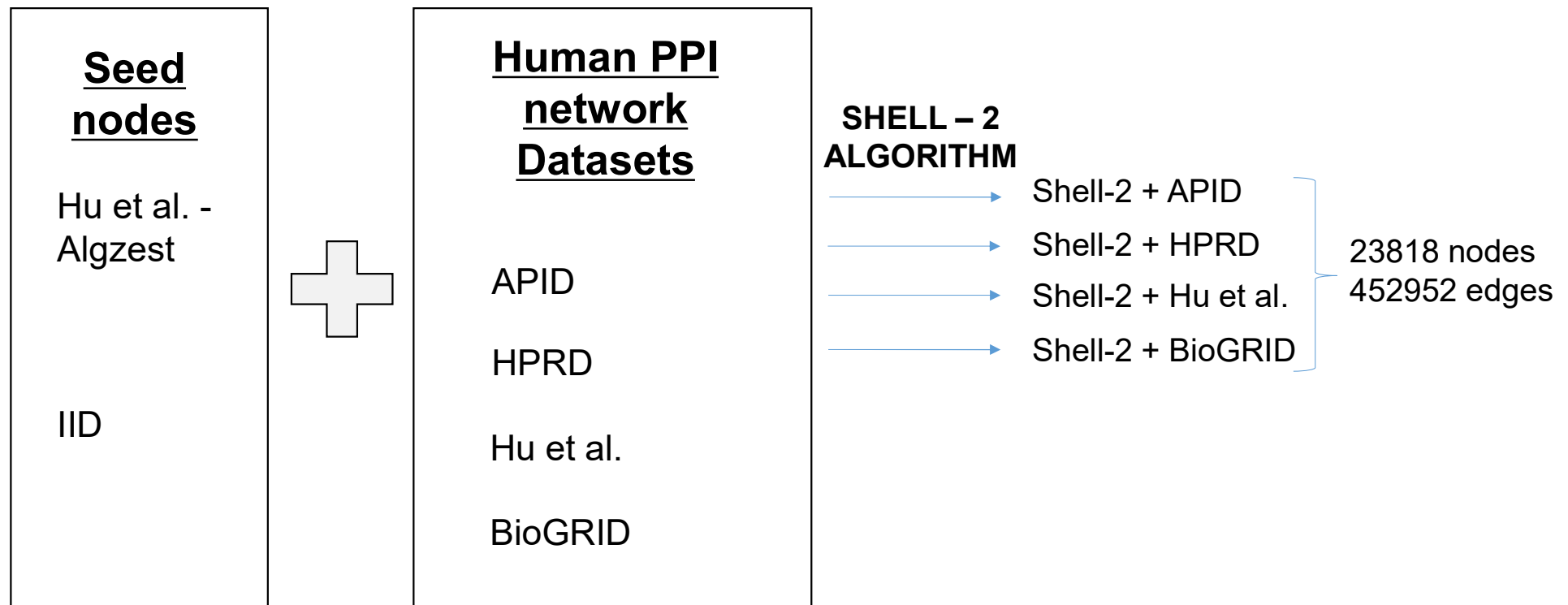
For

Conference on Nonlinear Systems and Dynamics,
Dec 14, 2019

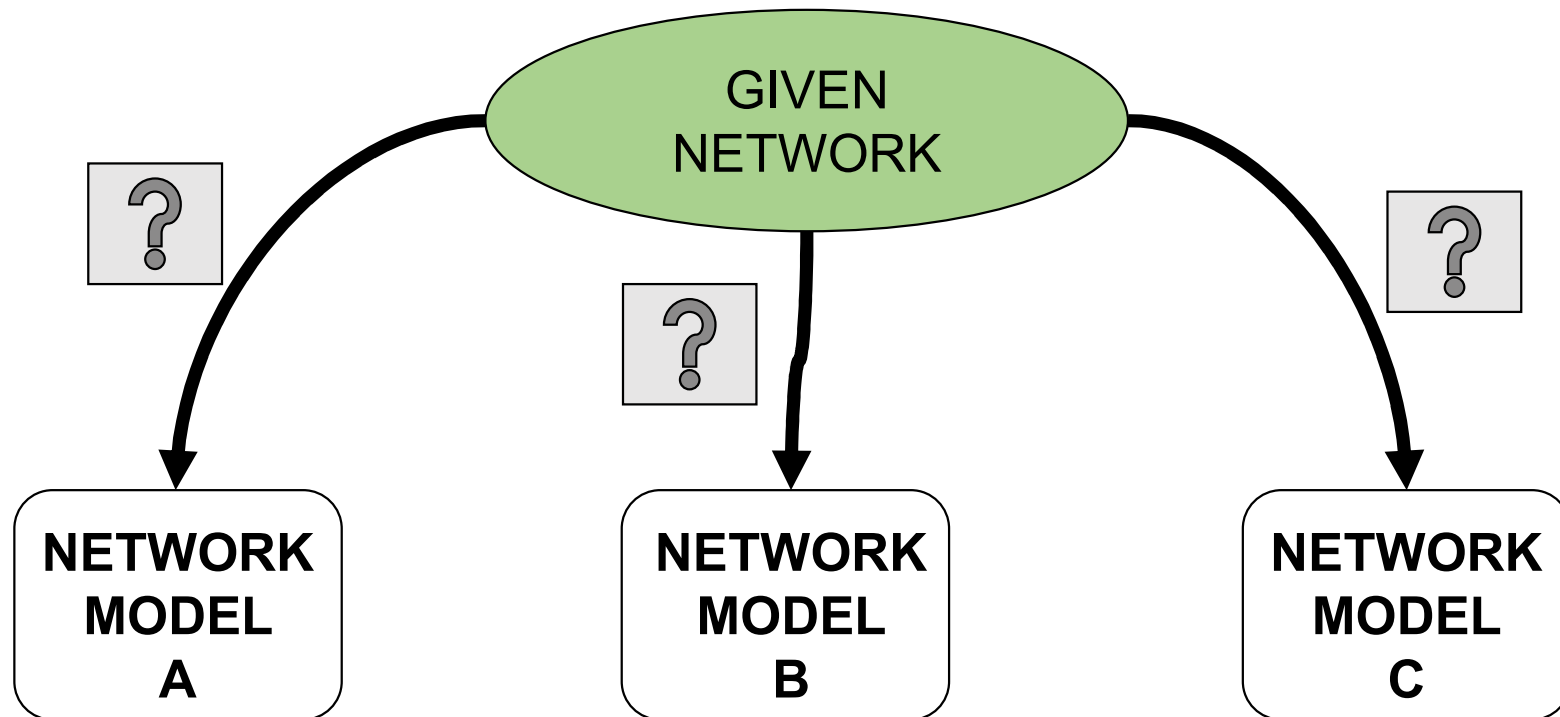
A Gist of the work



Protein-Protein Interaction Network of Alzheimer's Disease was extracted from the larger human PPI networks from different databases.



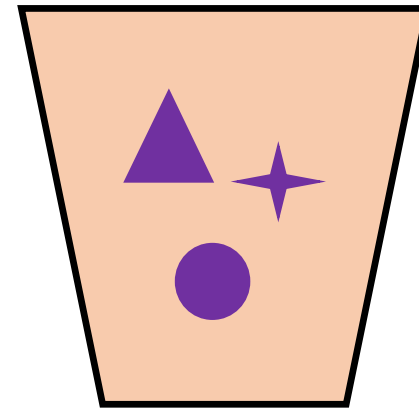
The Problem



Operational Definitions

What is a Network Model?

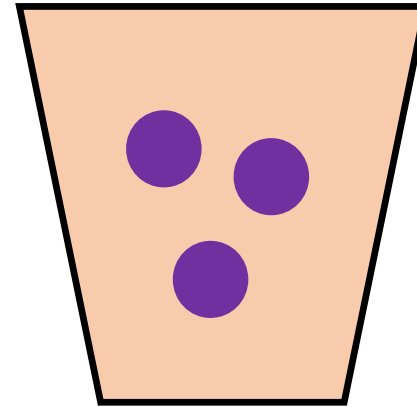
A bucket that holds networks,



Operational Definitions

What is a Network Model?

A bucket that holds networks,
all having the same “large
scale structure”.

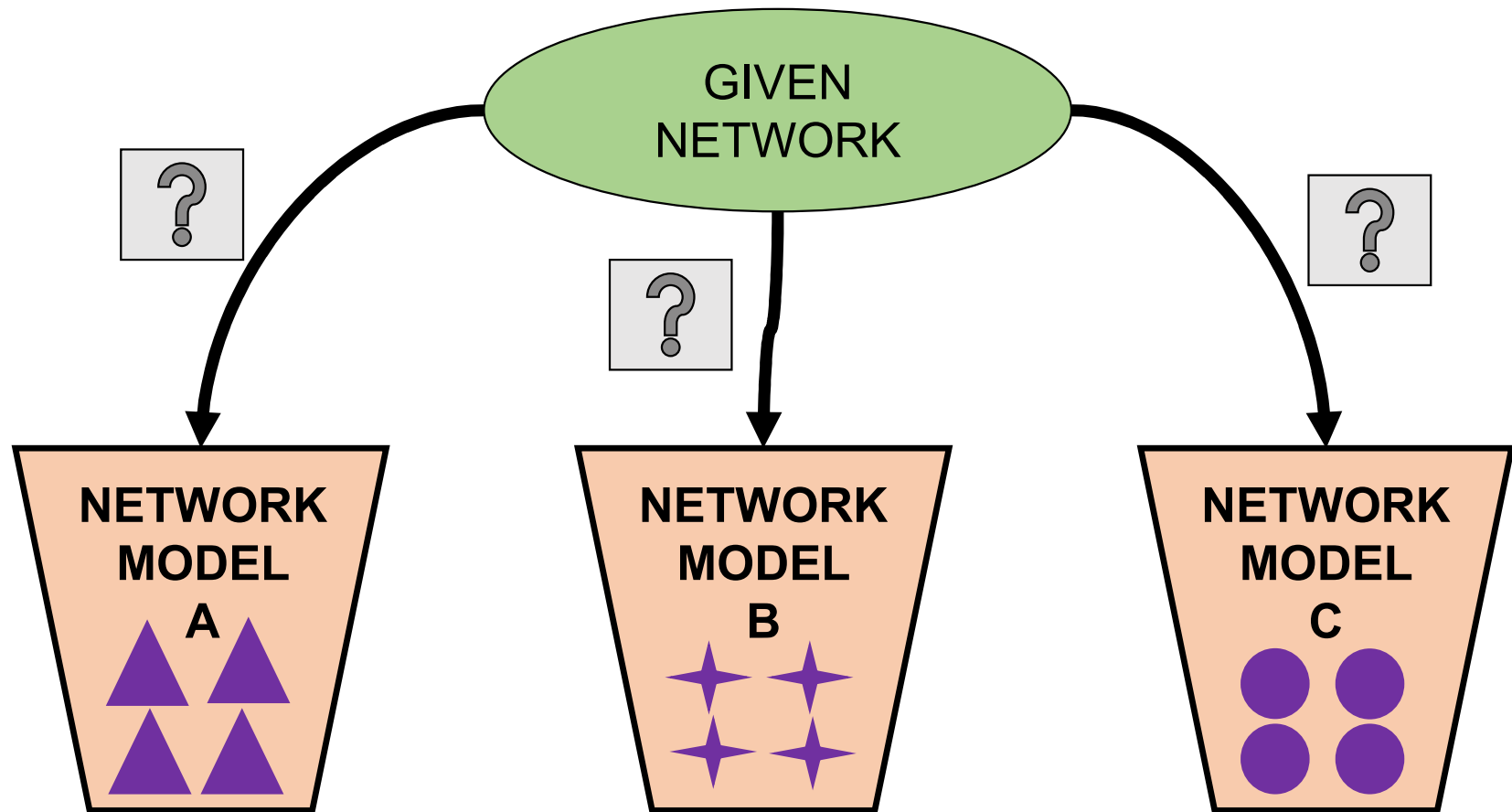


Operational Definitions

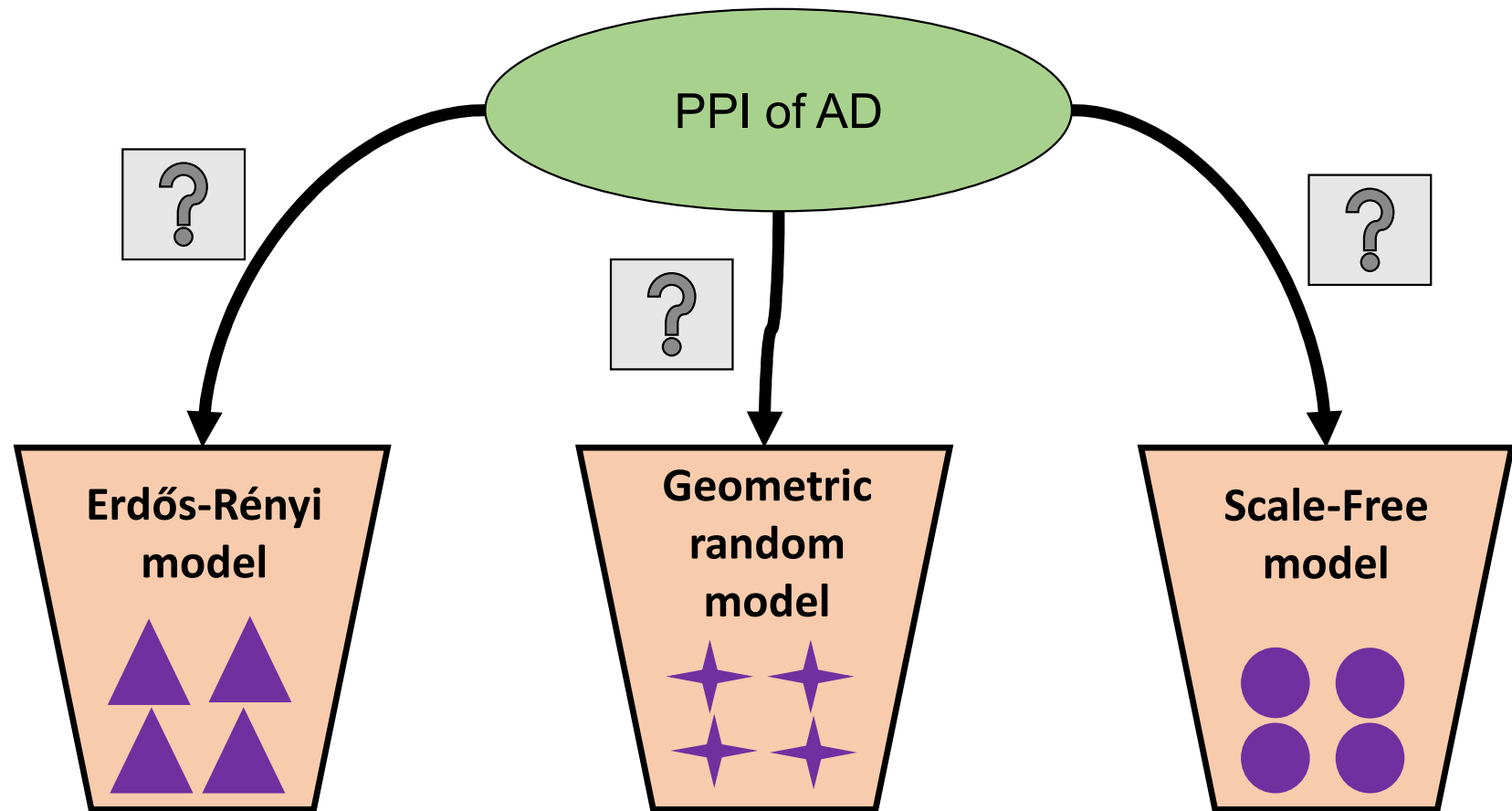
What is “large scale structure”?

A metric which faithfully represents the network,
preferably, in dimensions lower than the network's
dimensions itself

The Problem, concretely



The Problem, specifically



Which metric(s) to use as large scale structure?

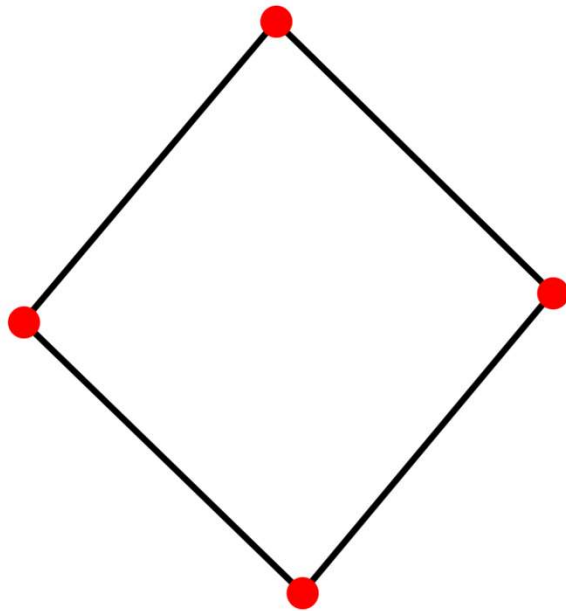
Some metrics found in the literature are:

- Average Path Length and Average Clustering coefficient
 - Watts & Strogatz¹
- Degree Distribution
 - Commonly used in network theory literature

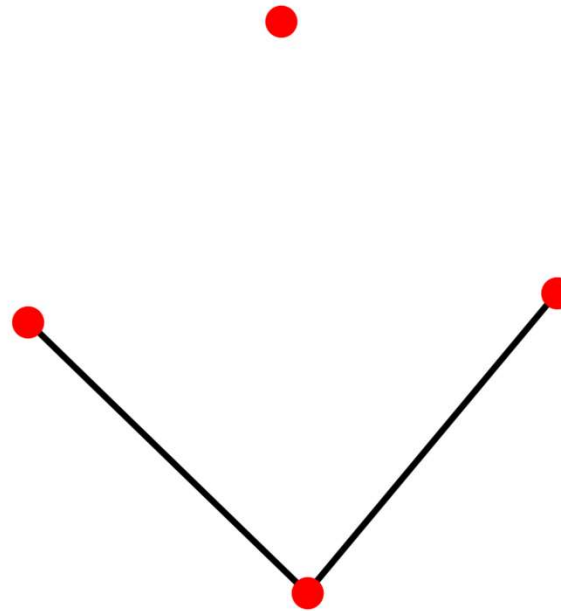
The problem with Path Length

We should be able to compare against a disconnected model graph too

Characteristic Path Length = 1.3



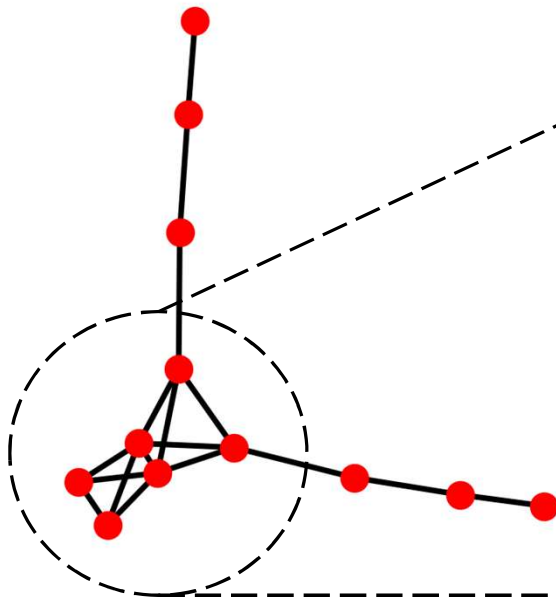
Characteristic Path Length : UNDEFINED



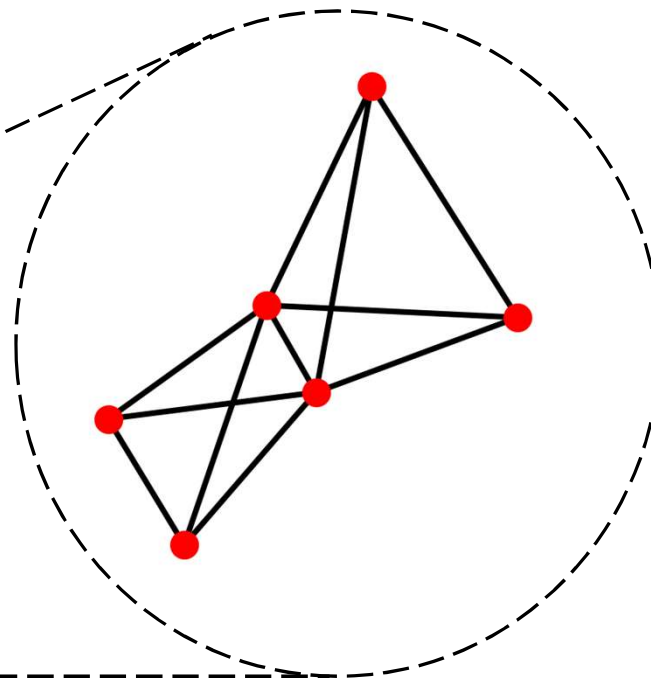
The problem with Clustering coefficient

Global (Average) Clustering coefficient is misrepresentative of local properties

Average Clustering coefficient = 0.35

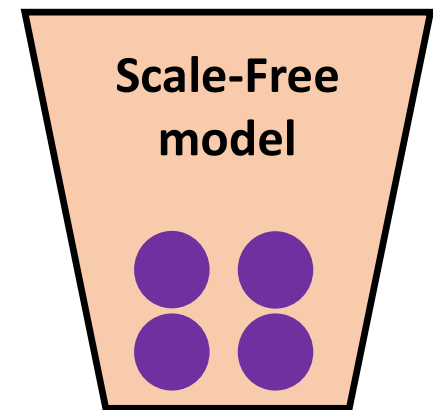
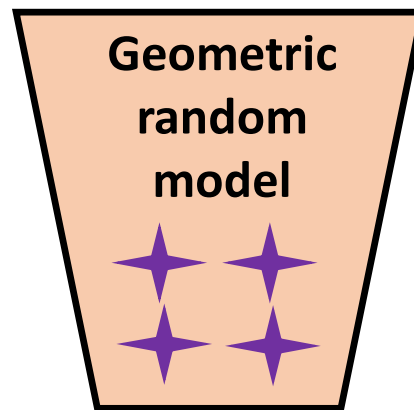
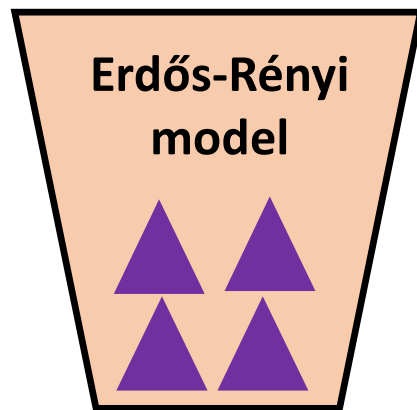


Average Clustering coefficient = 0.87



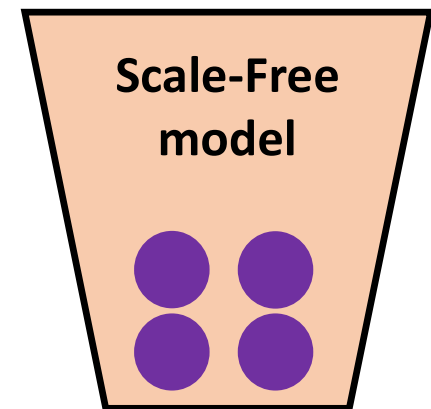
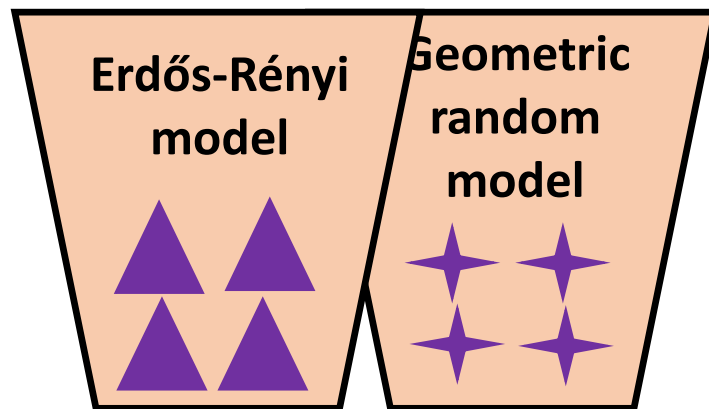
The problem with Degree Distribution

Lets compare the DD of the models we are interested in so that we can find out if we can get 3 buckets in the first place.



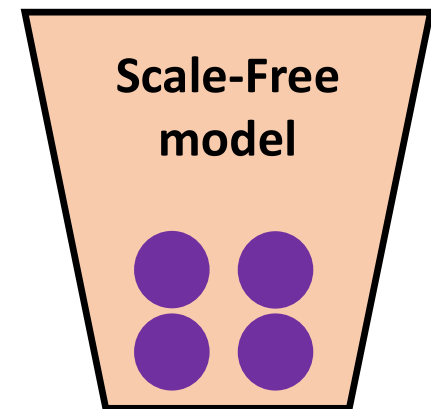
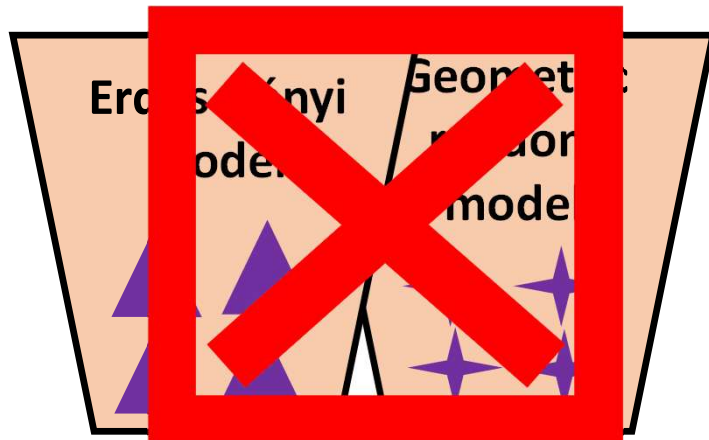
The problem with Degree Distribution

2 sample KS test between DD of ER and
GEO2D of same edge density as the
empirical graph fails to distinguish them



The problem with Degree Distribution

If we are lucky, empirical data follows a power law and we don't have to worry about the other two buckets!



The problem with Degree Distribution



“.. estimating the degree exponent is still not yet an exact science. We continue to lack methods that would estimate the statistical significance in a manner that would be acceptable to a practitioner.”

- Albert-László Barabási²

The problem with Degree Distribution

Results from Clauset-Shalizi-Newman method³

α estimate	3.306
$X_{\min}$ estimate	504
% of data beyond $X_{\min}$	0.48%

- The value of α is > 3
- Feels like $X_{\min}$ is too high...

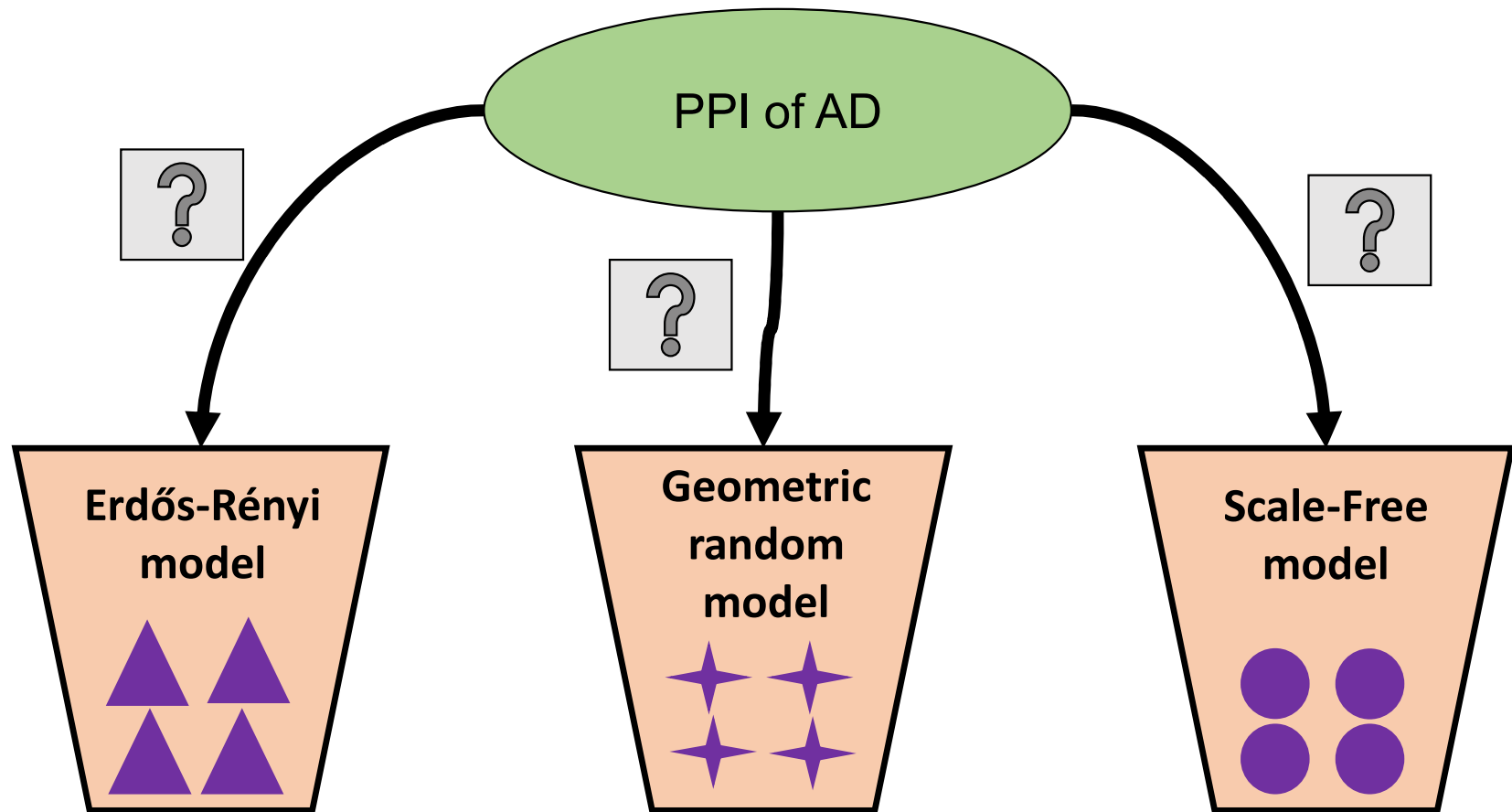
The problem with Degree Distribution

Results from Clauset-Shalizi-Newman method

Loglikelihood ratio test results (Positive implies Power law)

Exponential	Normalized R	1.4257	Only Moderate evidence !
	p-value	0.154	
Lognormal	Normalized R	0.312	
	p-value	0.755	
Truncated power law	Normalized R	0.0006	
	p-value	0.9997	

So the problem remains.
How do we proceed next?



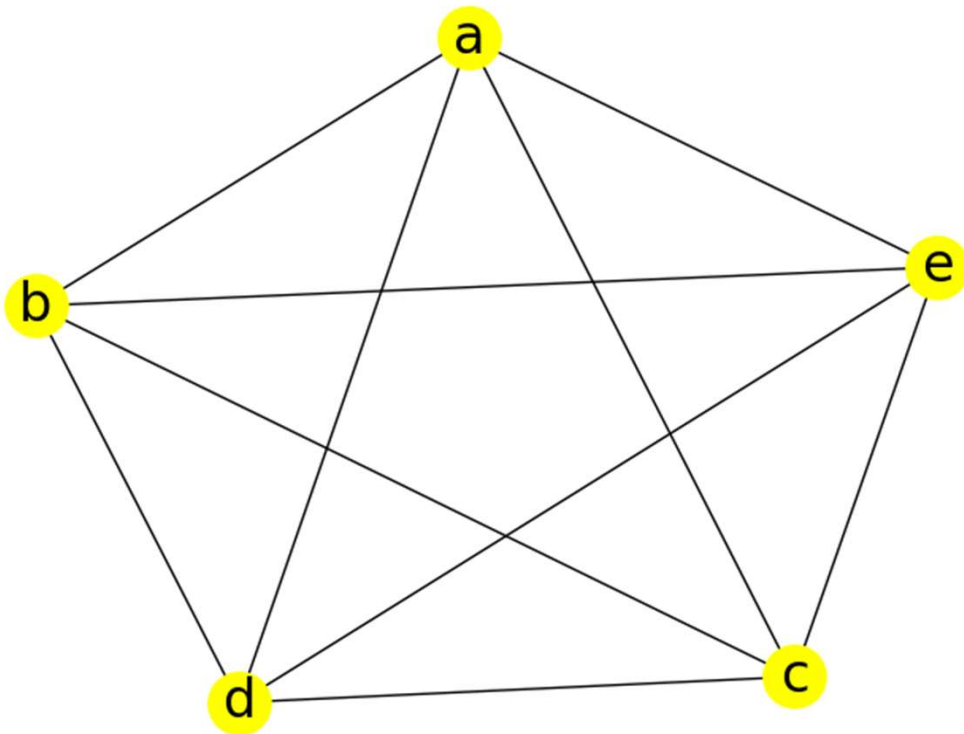
Why did de we fail?

- None of the metrics seem to faithfully represent the intricacies of the whole network.
- Ideally, we shouldn't be comparing metrics of the networks, but the networks themselves...

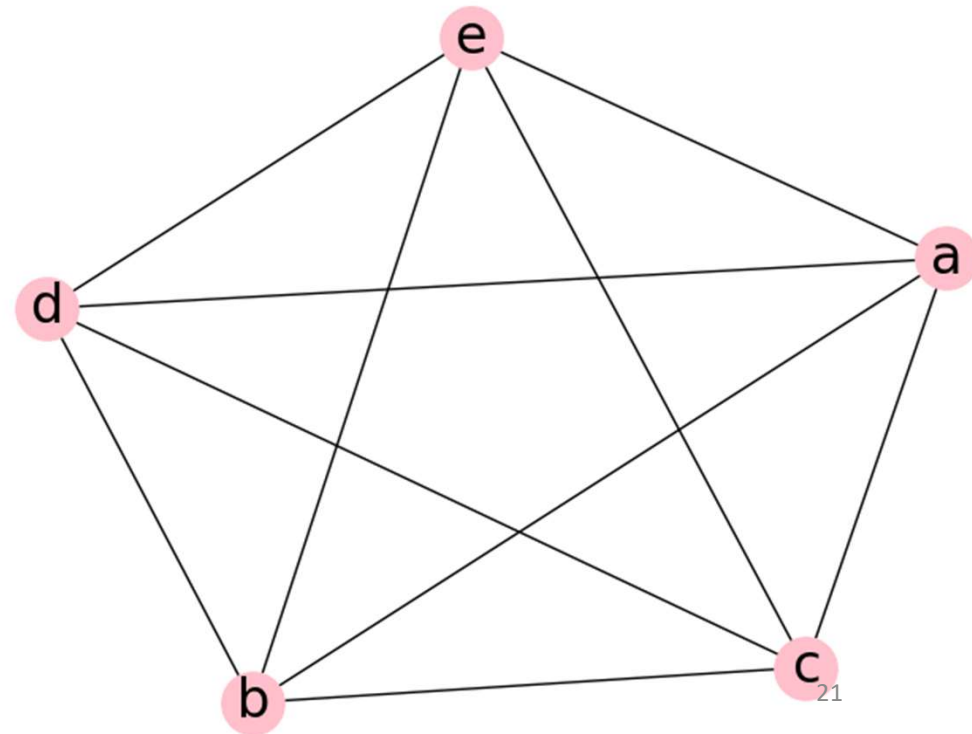
Checking for similarity - Graph Isomorphism

Consider the two graphs here. Are they similar?

Graph 1



Graph 2



Checking for similarity - Graph Isomorphism

They indeed are! A bijective map exists between the two.
Such graphs are said to be isomorphic.

$$a \iff e$$

$$e \iff a$$

$$c \iff c$$

$$b \iff d$$

$$d \iff b$$

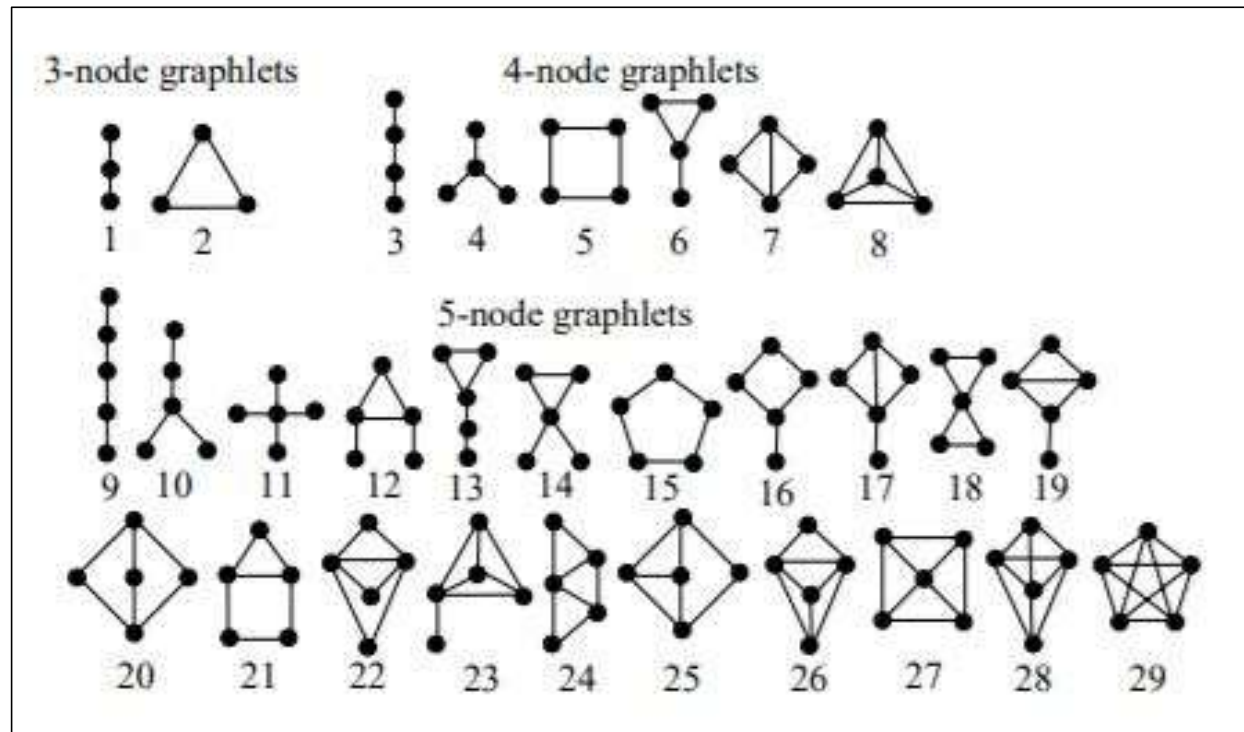
Checking for similarity - Graph Isomorphism

But Polynomial time algorithms don't exist!

Are there heuristic algorithms?

Fortunately, yes - **Graphlets**

Basic Graphlet Analysis

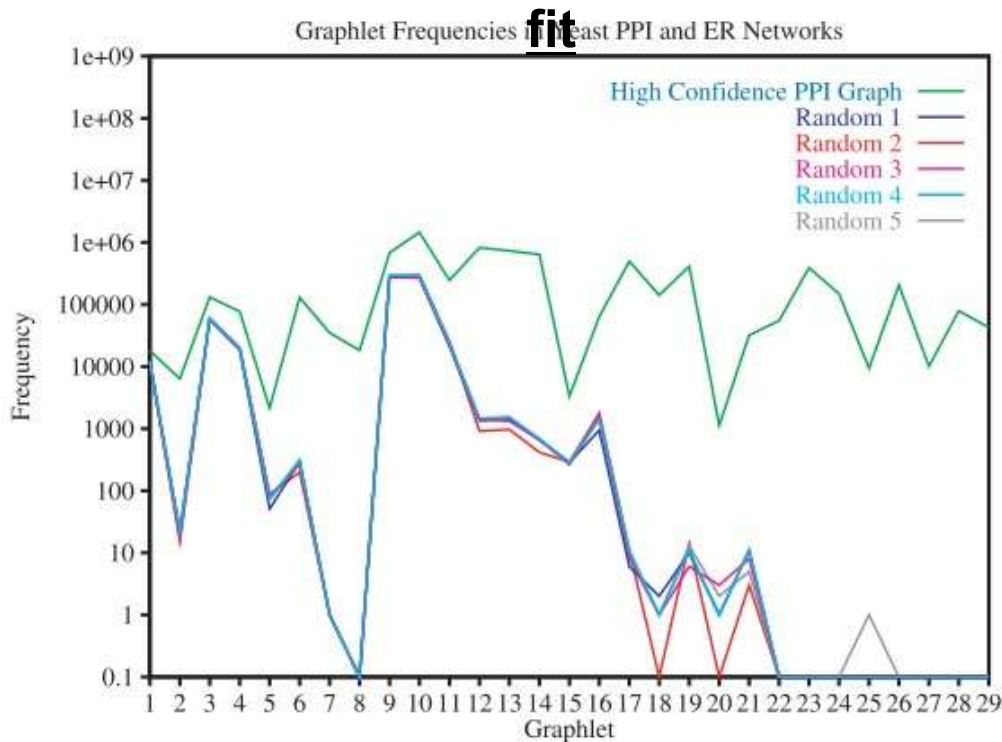


The 29 possible graphlets for 3-5 nodes⁴

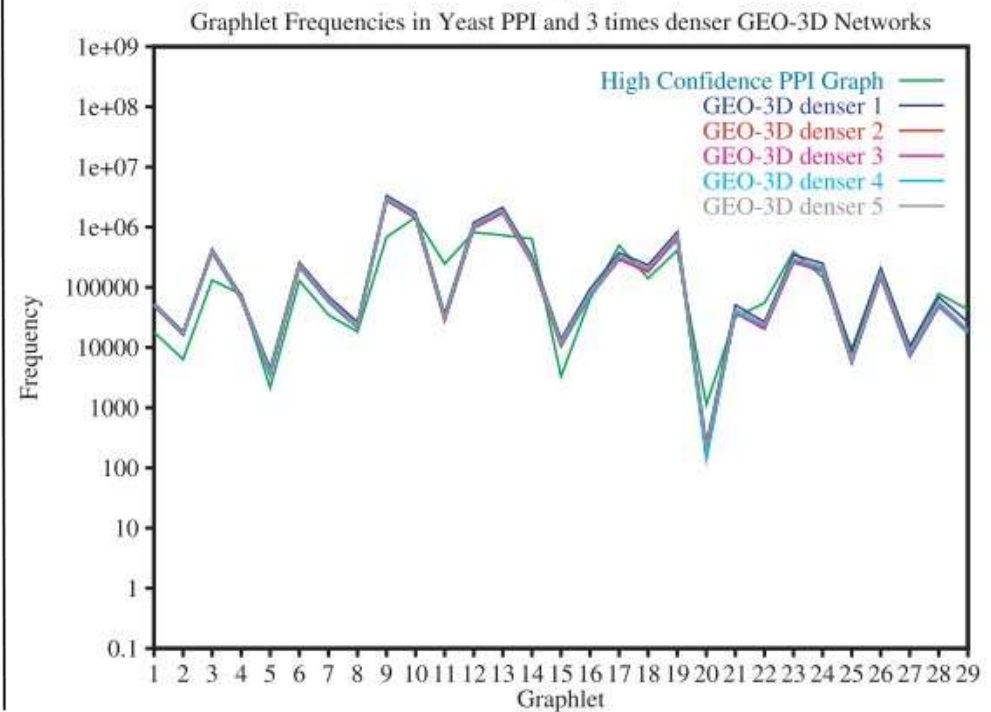
**Lets count how many of each of these graphlets
are there in a given network...**

The basic idea of graphlet analysis is comparing frequencies of graphlets⁴

Graphlet frequency distributions don't fit



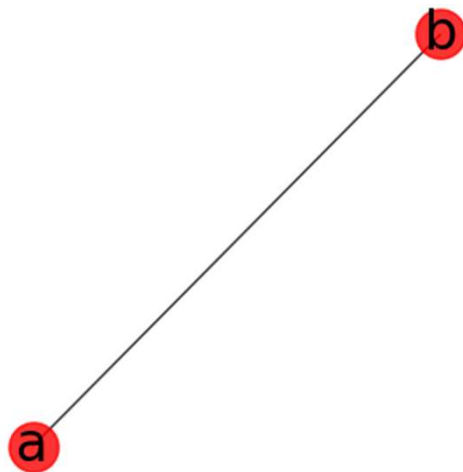
Graphlet frequency distributions do fit



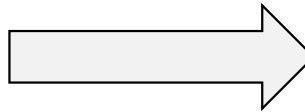
But further developments have been made...

Degree Distributions have been generalized to Graphlet degree distributions⁵

A 2-node graphlet is
nothing
but an edge!

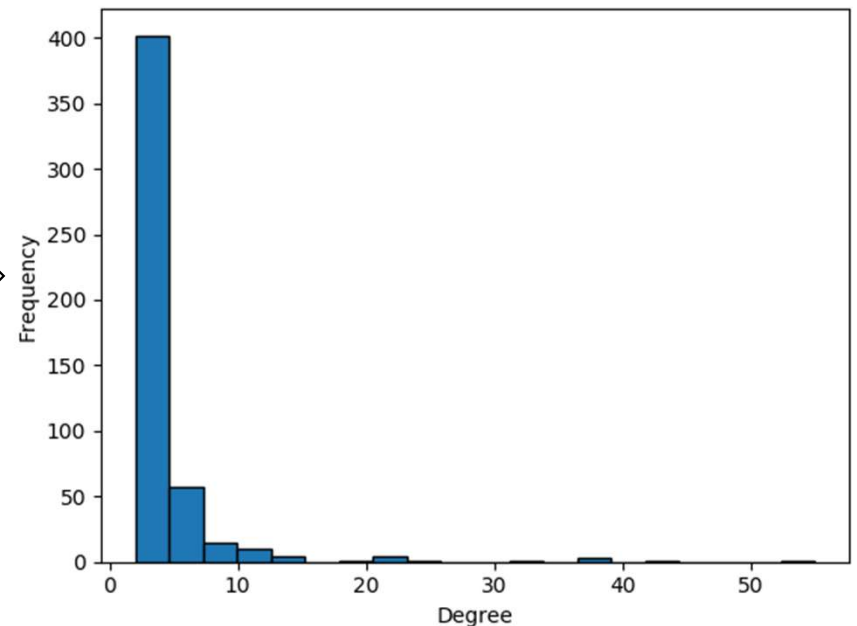


Its distribution



is nothing but...

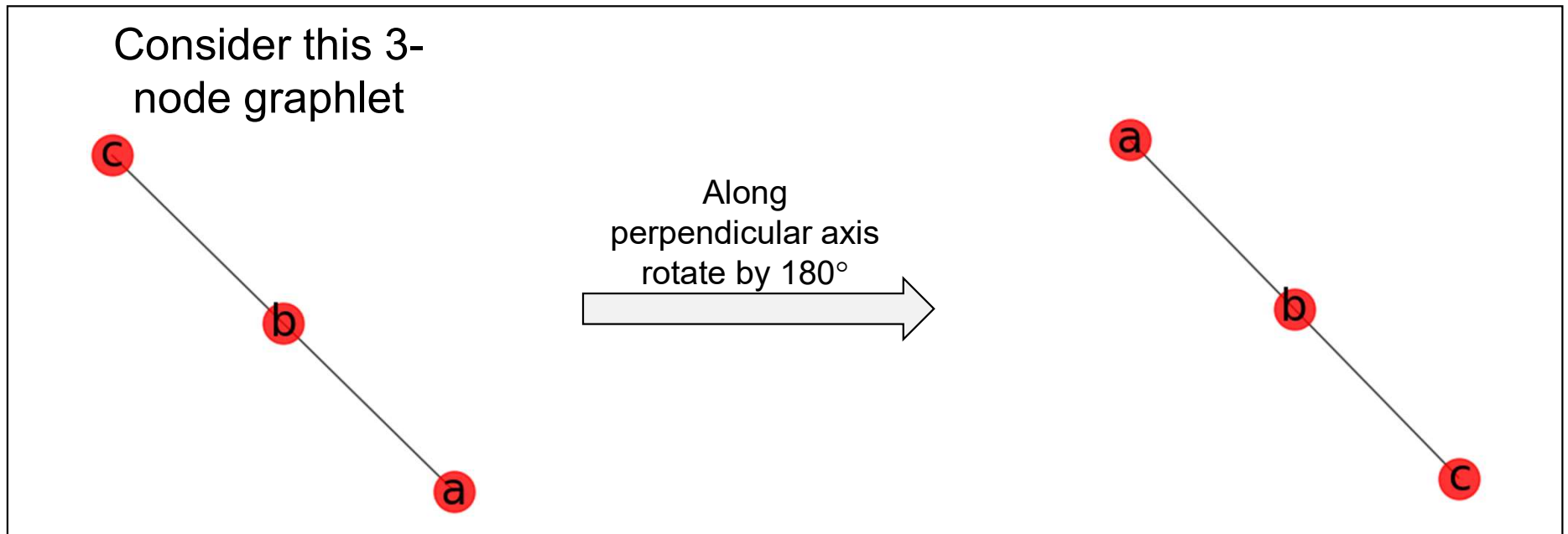
... Degree distribution!



Why not make distributions of structures with more than 2 nodes?
Thus, we generalize to graphlet degree distributions.

But EVEN MORE developments have been made...

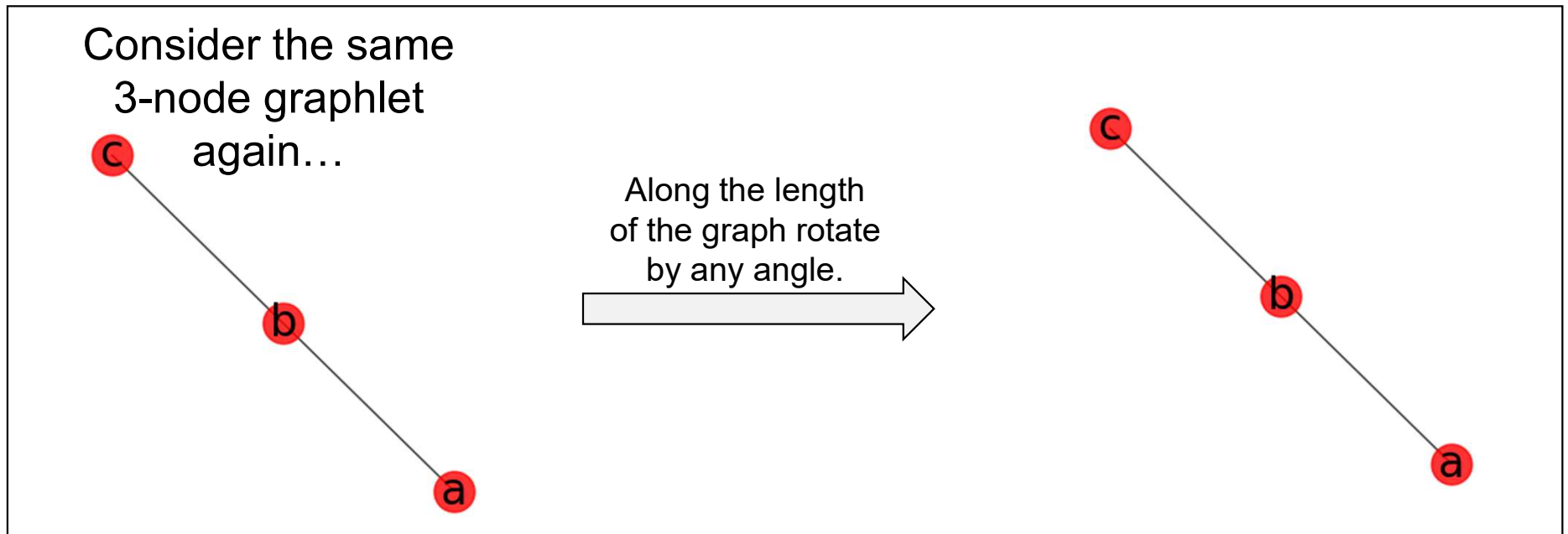
Each graphlet has some symmetry and hence is *Automorphic*.



A bijection exists between the two:

$$\alpha_1 = \{ac\}\{b\}$$

Each graphlet has some symmetry and hence is *Automorphic*.



Another bijection exists between these two:

$$\alpha_2 = \{a\}\{b\}\{c\}$$

Each graphlet has some symmetry and hence is *Automorphic*.

Collecting the two bijections into an *Automorphic group*:

$$Aut(\alpha) = \{\alpha_1, \alpha_2\} \quad (1)$$

$$= \{\{ac\}\{b\}, \{a\}\{b\}\{c\}\} \quad (2)$$

Defining *Automorphic orbit* for each node:

$$Orb(a) = \{c, a\} \quad (1)$$

$$Orb(b) = \{b\} \quad (2)$$

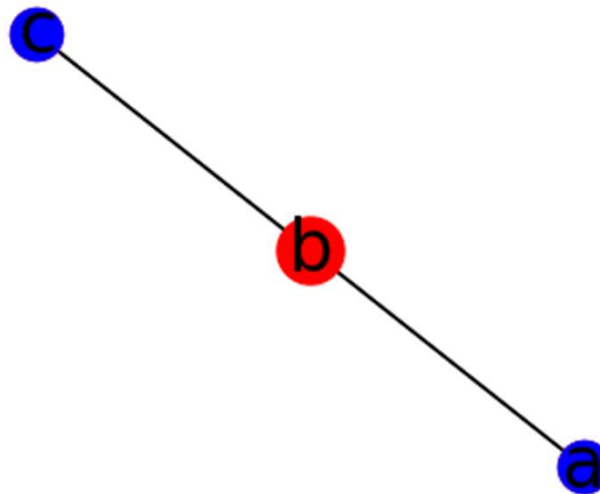
$$Orb(c) = \{a, c\} \quad (3)$$

Note that:

$$Orb(a) = Orb(c)$$

**Each graphlet has some symmetry and hence is
Automorphic.**

So, c and a are
in the same
orbit while b is
in a different
one.



That is,
touching c or
a is
topologically
the same
while
touching b is
different.

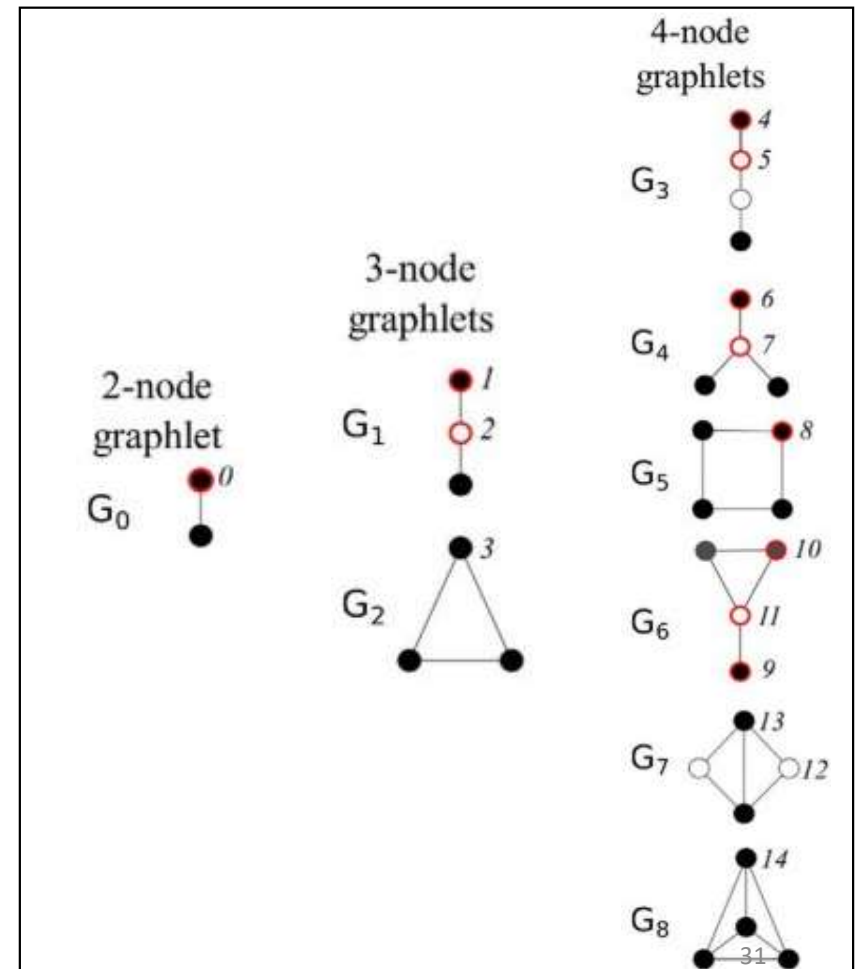
Automorphism of graphlets increases constraints⁵

The same can be done for all the graphlets.

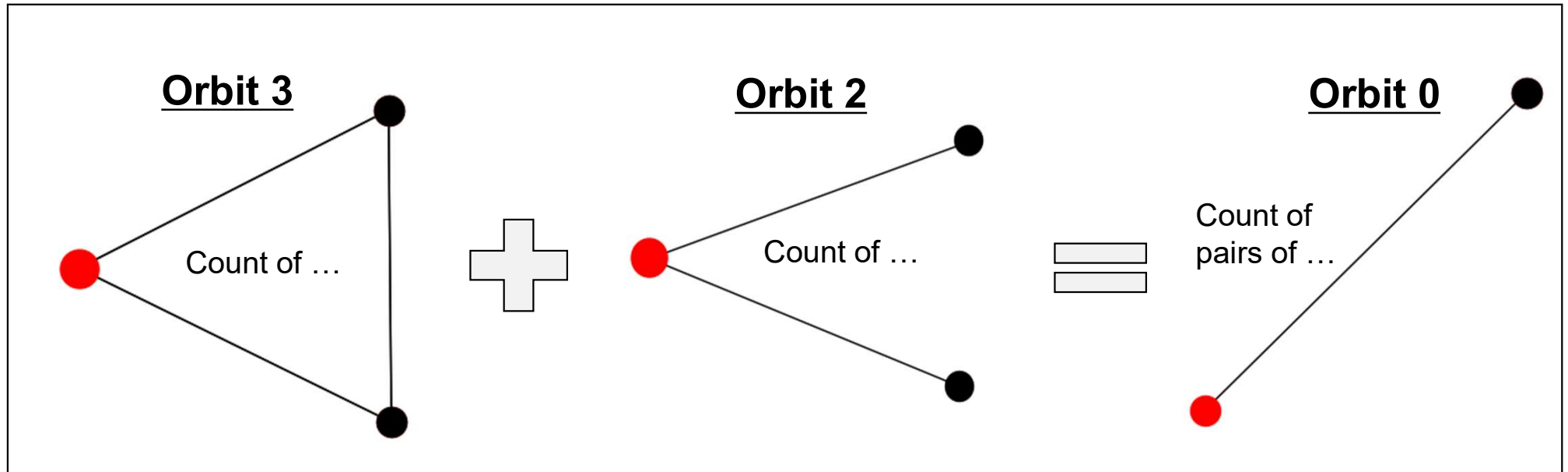
So instead of graphlet degree distribution we now have orbit degree distribution.

Thus we have increased the number of constraints we are putting when we do the classification.

But what if this is too much noise?



There is redundancy in orbit counts⁶



Mathematically,

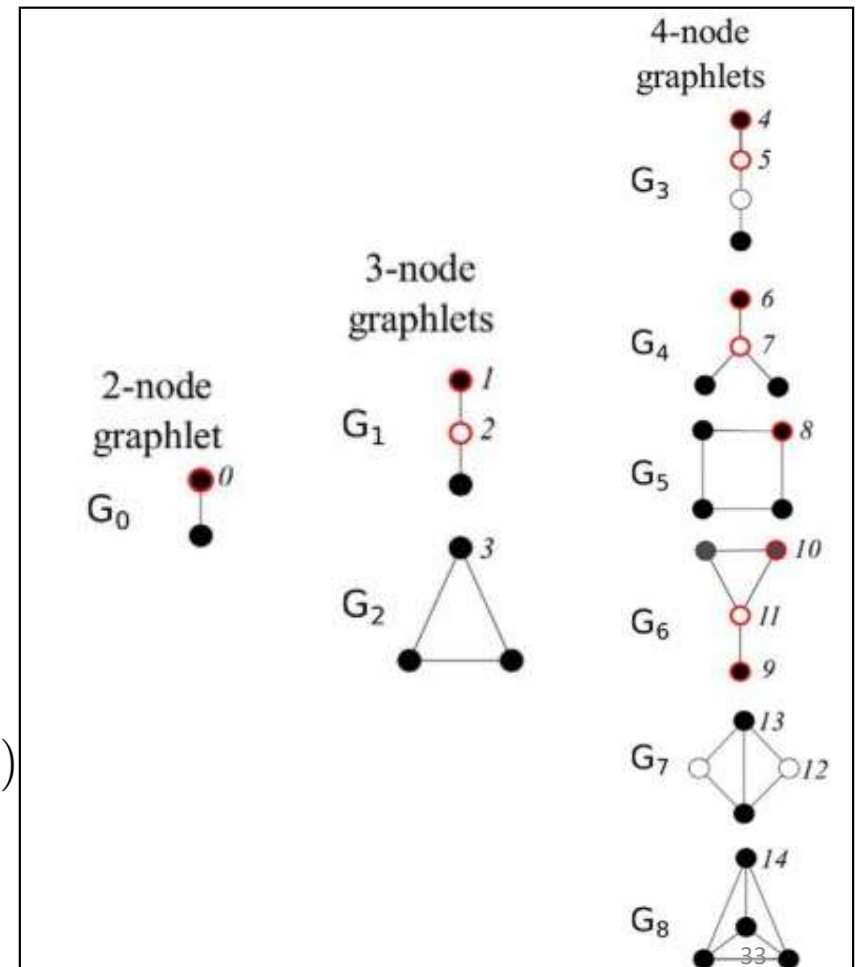
$$C_3 + C_2 = \binom{C_0}{2} \Rightarrow \text{It is redundant to count one of these 3 orbits!}$$

We remove noise in our metric by not counting the redundant orbits⁶.

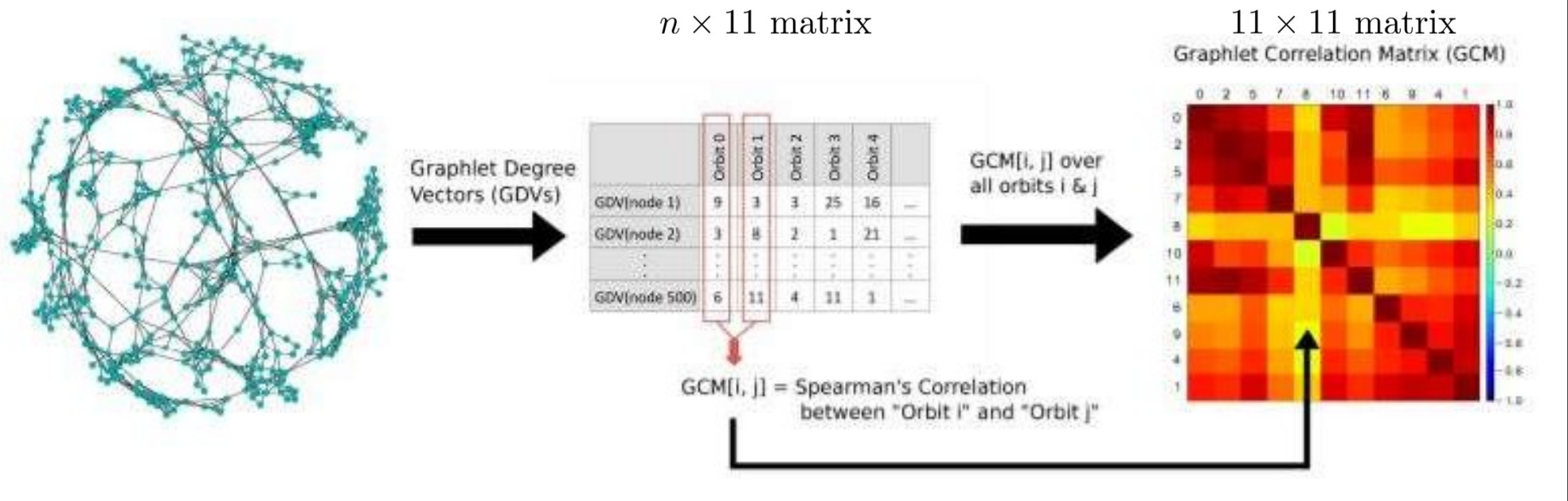
There are 4 equations like the one we just saw. So we can remove 4 orbits as redundant. So there are 11 non redundant orbits (as highlighted).

We put the number of non redundant orbits each node is touching into a vector and assign it to the node. Call this vector Orbit Degree Vector (ODV). That is,

$$ODV_i = (C_0, C_1, C_2, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11})$$



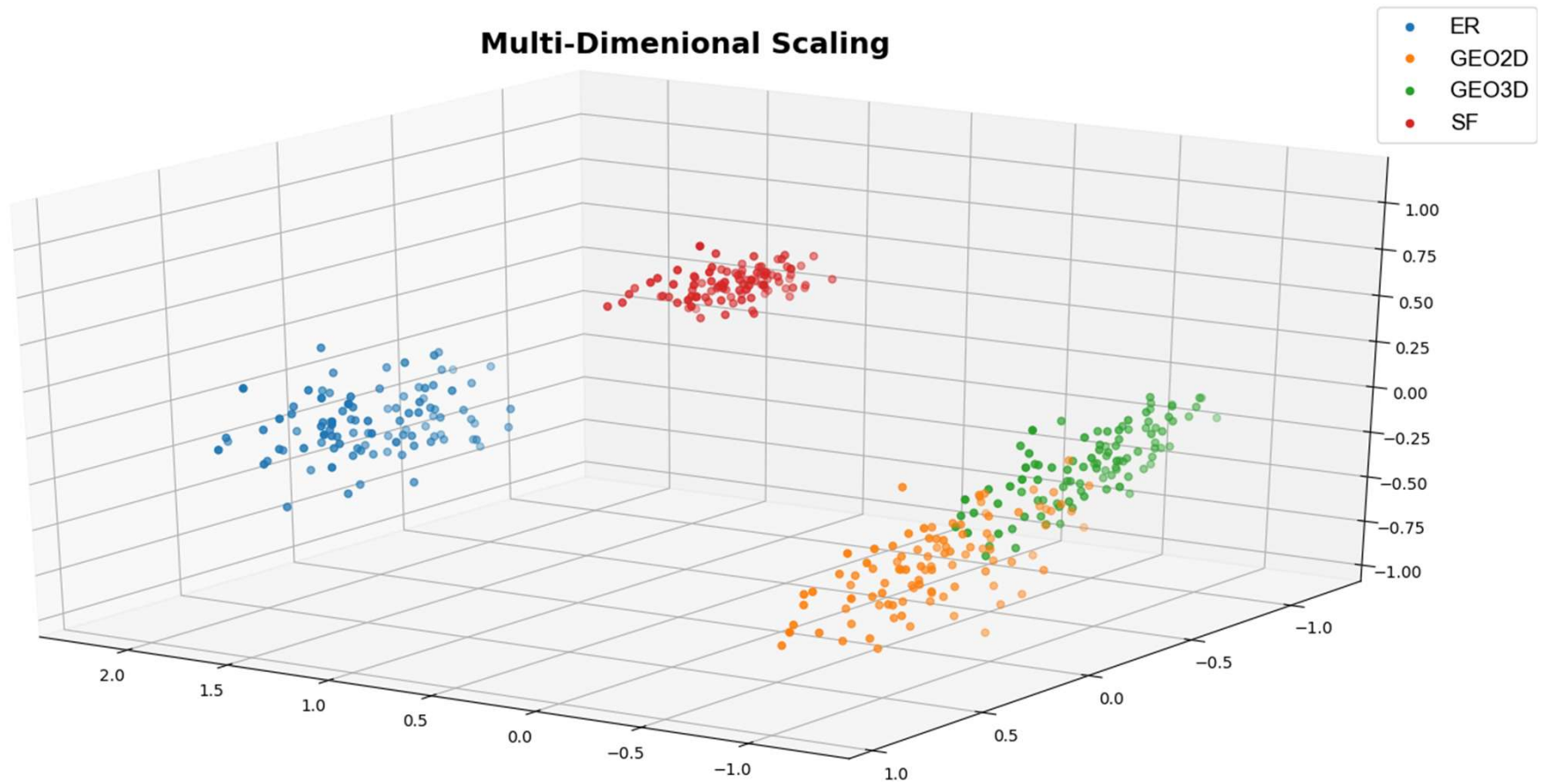
We use the correlations between the non redundant orbits as the new graphlet metric⁶ – GCD11



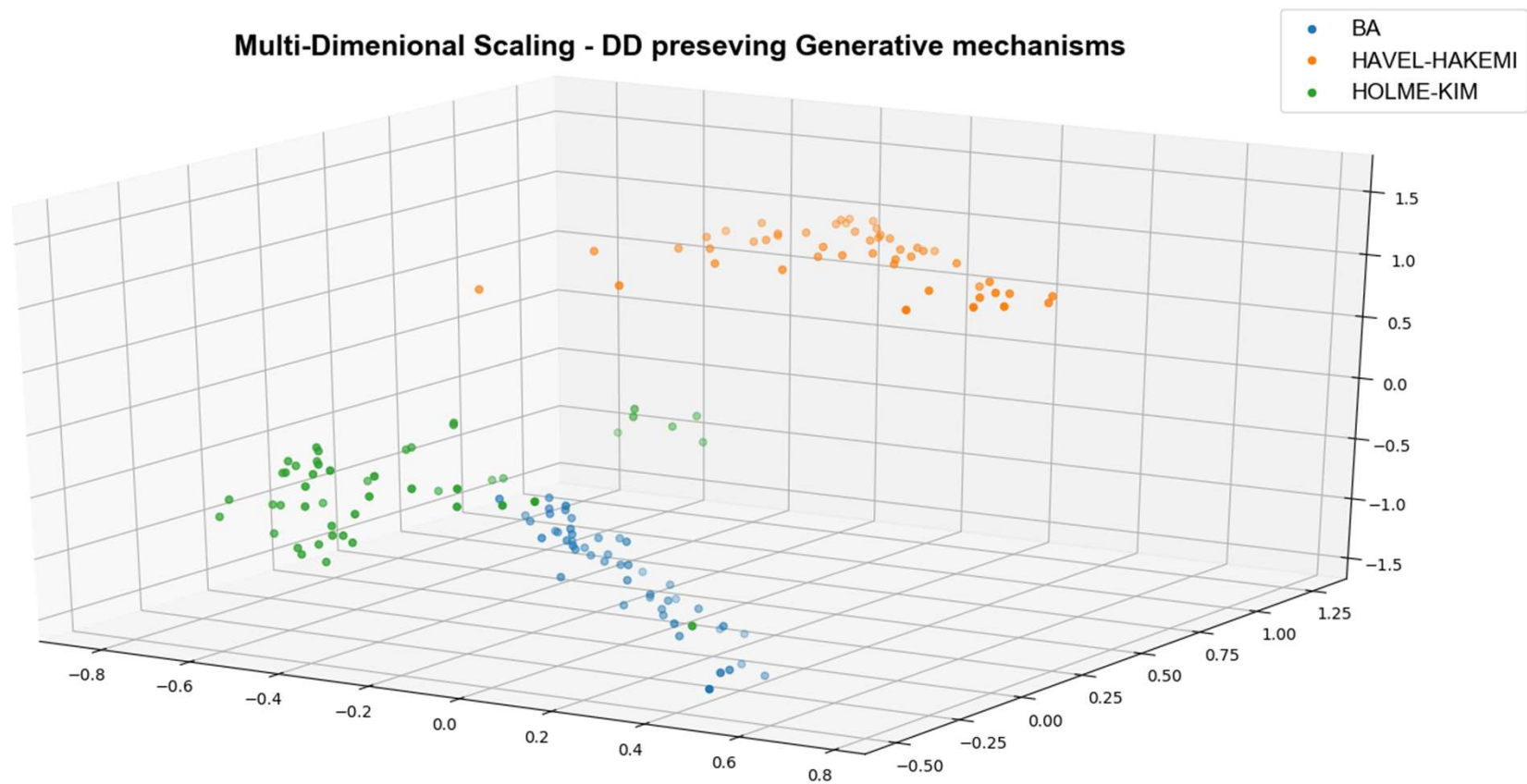
Graphlet Correlation Distance - 11:

$$GCD11(G, H) = \sqrt{\sum_{i=1}^{11} \sum_{j=i+1}^{11} [GCM_G(i, j) - GCM_H(i, j)]^2}$$

GCD11 gives us the three buckets!



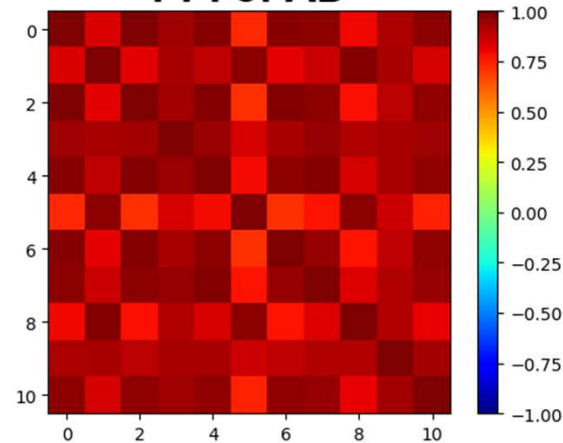
GCD11 is good at classifying different network models



Now that we have validated GCD11 as a metric that faithfully represents the large scale structure, lets use it to compare study the empirical data

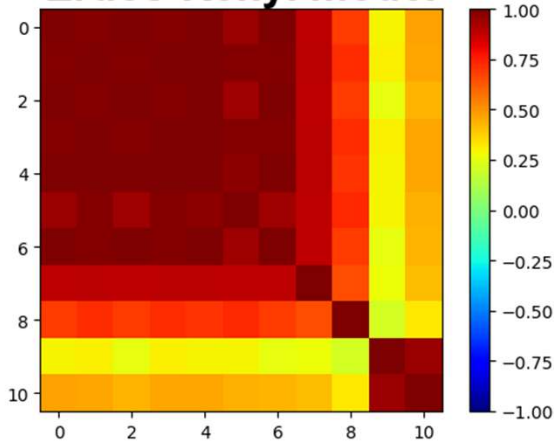
Visualizing GCM

PPI of AD

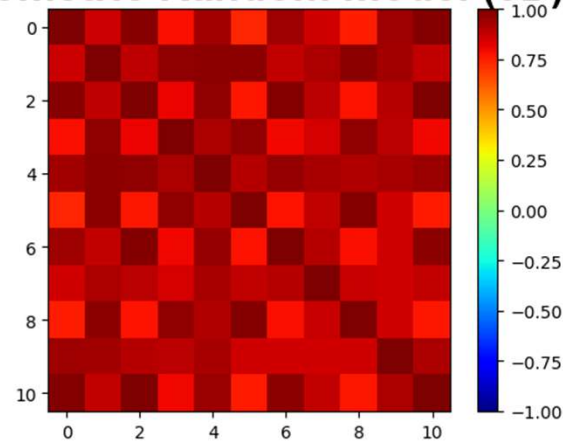


It looks like it
matches
Geometric
Random model...

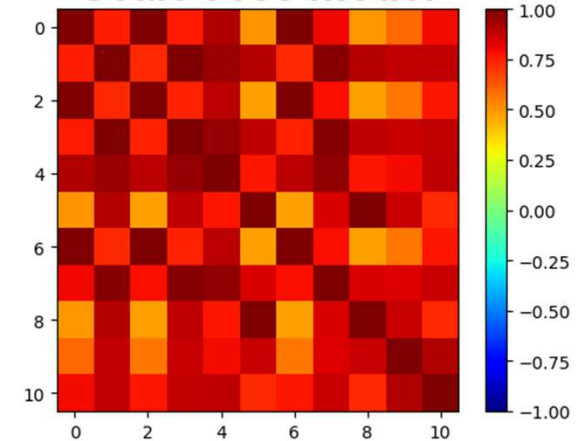
Erdos-Renyi model



Geometric Random model (3D)



Scale-Free model



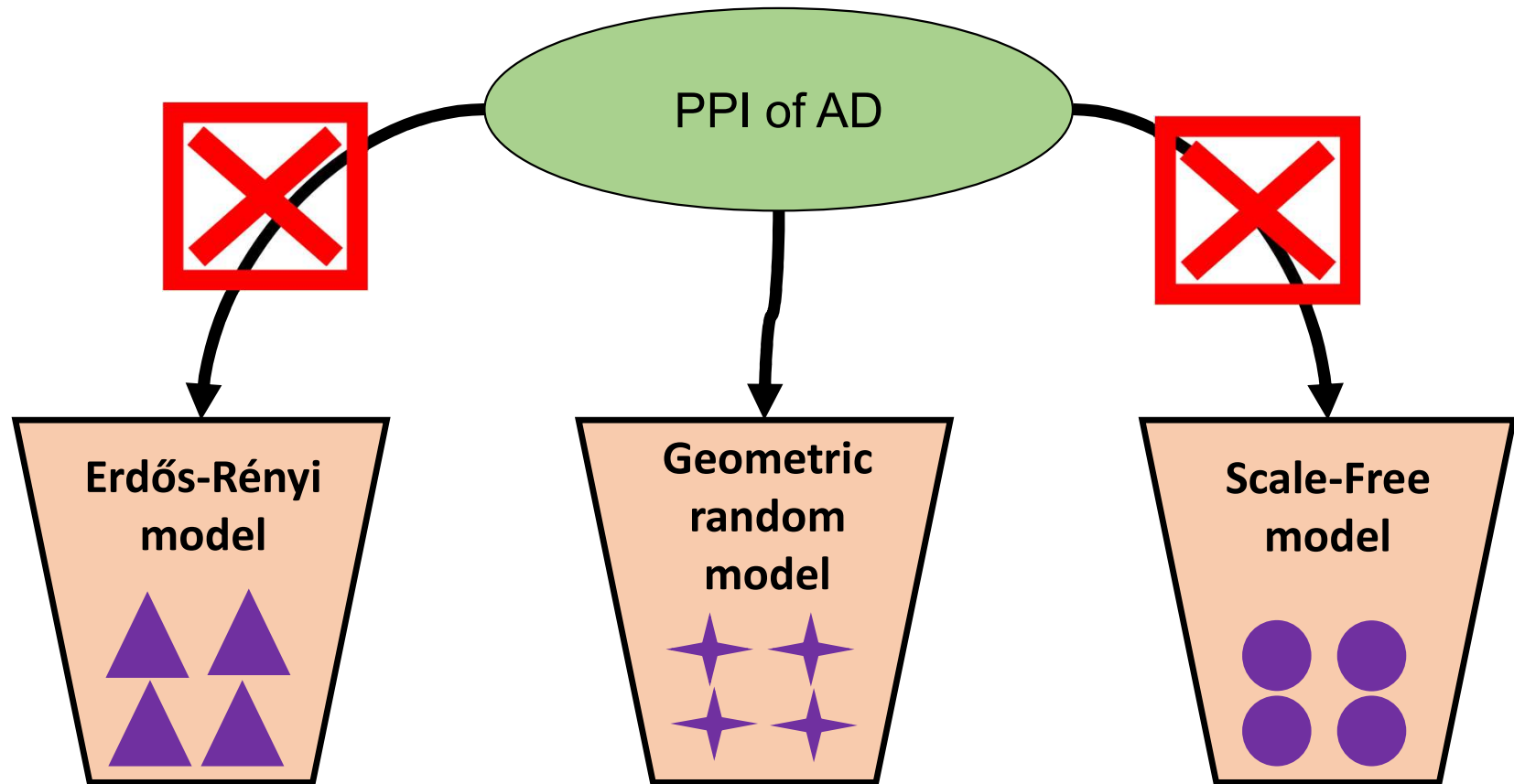
... and GCD numbers agree with the visualization.

GCD-11 numbers

Comparing PPI of AD against 100 samples of each model

	Mean	Std.Dev
Geometric Random model (3D)	0.47	0.01
Scale-Free model	1.09	0.02
Erdos-Renyi model	2.49	0.03

**So, the AD-specific subgraph
is more Geometric Random
than Scale-Free or Erdos-Renyi.**



Caveats

- We are not claiming that it is more Geo3D than Geo2D since the MDS visualization showed some overlap between the two.
- Assumed, Scale-Free = Barabasi-Albert.
- Some criticism has been levied, regarding threshold behaviour of GDDA but it has been defended and in addition GCD is not known to be prone to such problems since we are only seeing correlation numbers.

Take home message

It may not be enough to just study the distribution of 2-node subgraph counts (i.e.) conventional degree distribution.

One should start studying the higher distributions also to fully realize the intricate large scale structure of a network.

References

1. Watts, D. and Strogatz, S. 1998. "Collective dynamics of 'small-world' networks". *Nature* **393**, 440–442
2. Barabási, Albert-László. 2016. "Network science". Cambridge university press.
3. Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 2009 51:4, 661-703
4. Pržulj, N., D. G. Corneil, and I. Jurisica. 2004. "Modeling Interactome: Scale-Free or Geometric?" *Bioinformatics* 20 (18): 3508–15. <https://doi.org/10.1093/bioinformatics/bth436>.
5. Pržulj, Nataša. 2007. "Biological Network Comparison Using Graphlet Degree Distribution." *Bioinformatics* 23 (2): 177–83.
6. Yaveroğlu, Ömer Nebil, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. 2015. "Revealing the Hidden Language of Complex Networks." *Scientific Reports* 4 (1): 4547.