

Complex Networks Clustering and Edges Correlation

Igor Kanovsky

Max Stern Academic College of Emek Yezreel

Israel

Clustering Problem and Links Correlation

Clustering problem : How to recognize communities within a graph.

- There is no exact definition of the cluster. There are a lot of methods.
- Links correlation: links from a node are not distributed randomly.

Our aims was:

- to design an effective algorithm for community recognition in Small World networks.
- to check a linkage between links correlations and clustering in real world networks.

Main Idea

- If two linked nodes have a big number of common neighbours it means that they are member of the same community. (If they have not it does not mean that they belong to different community).
 - What it is “a big number” ?
 - How can the idea be applied?
 - What is a measure for a local links correlation in a network?

Outline

- ❑ Small World Network definition.
- ❑ Link weighting.
- ❑ Iterated Cluster Recognition Algorithm (ICRA).
- ❑ Extended Small World Graph Model.
- ❑ Zachary's Karate Club Study.
- ❑ Spidering the Web.
- ❑ Conclusion.

Small World Network (1/3)

- Def.1. The characteristic path length $L(G)$ of a graph $G = (V, E)$ is the average length of the shortest path between two vertices in G
- Def.2. The clustering coefficient $C(G) = \langle C(v) \rangle$ of a graph $G = (V, E)$ is the average clustering coefficient of its vertices $C(v)$;

Small World Network (2/3)

- Def. 3. Clustering coefficient for vertex v :

$$C(v) = \frac{\text{number of edges in } G[N(v)] \cdot 2}{k(v) \cdot (k(v) - 1)}$$

where $N(v)$ - neighbourhood of v , $k(v) = |N(v)|$
(degree of v).

The clustering coefficient $C(v)$ of a vertex v is the edge density of the graph $G[N(v)]$ induced by $N(v)$.

Small World Network (3/3)

- Def. 4. Small World network is a graph $G(V, E)$ with $L \sim L_R \sim \log|V|$ and $C \gg C_R$, where $G_R(V_R, E_R)$ - a random graph with $|V_R| = |V|$, $|E_R| = |E|$.

A lot of real world graphs are Small World graphs:

1. Social relationships.
2. Business (organization) collaborations.
3. The Web. The Internet.
4. Biological data (DNA structure, cells metabolism etc.).

Link Weighting In Small World

Definition: For the link connecting nodes v_1, v_2 we define the edge weighting parameter β as:

$$\beta(v_1, v_2) = \frac{|N(v_1) \cap N(v_2)|}{|N(v_1) \cup N(v_2)|}$$

Where $N(v)$ is the neighborhood of the vertex v .

Iterated Cluster Recognition Algorithm (ICRA) (1/3)

Definition: Two adjacent nodes v_1, v_2 belong to the same cluster if

$$\beta(v_1, v_2) > \alpha$$

Where α is the level of the cluster separation for a Small World network (threshold value which is a parameter of the algorithm).

By simple iterated procedure a Small World network can be divided into m clusters, where m is between 1 to $|V|$ depends on α and the graph link structure.

ICRA (2/3)

ICRA Flow :

Input:

graph $G=(V,E)$, level of cluster separation α ;

Output: clustering $\{V_1, V_2, \dots, V_k\}$;

1. $i=0$;

2. loop while V is not empty

a. find an arbitrary cluster C in $G[V]$;

b. $i++$;

c. $V_i=C$; $V=V-C$;

3. $k=i$;

I CRA (3/3)

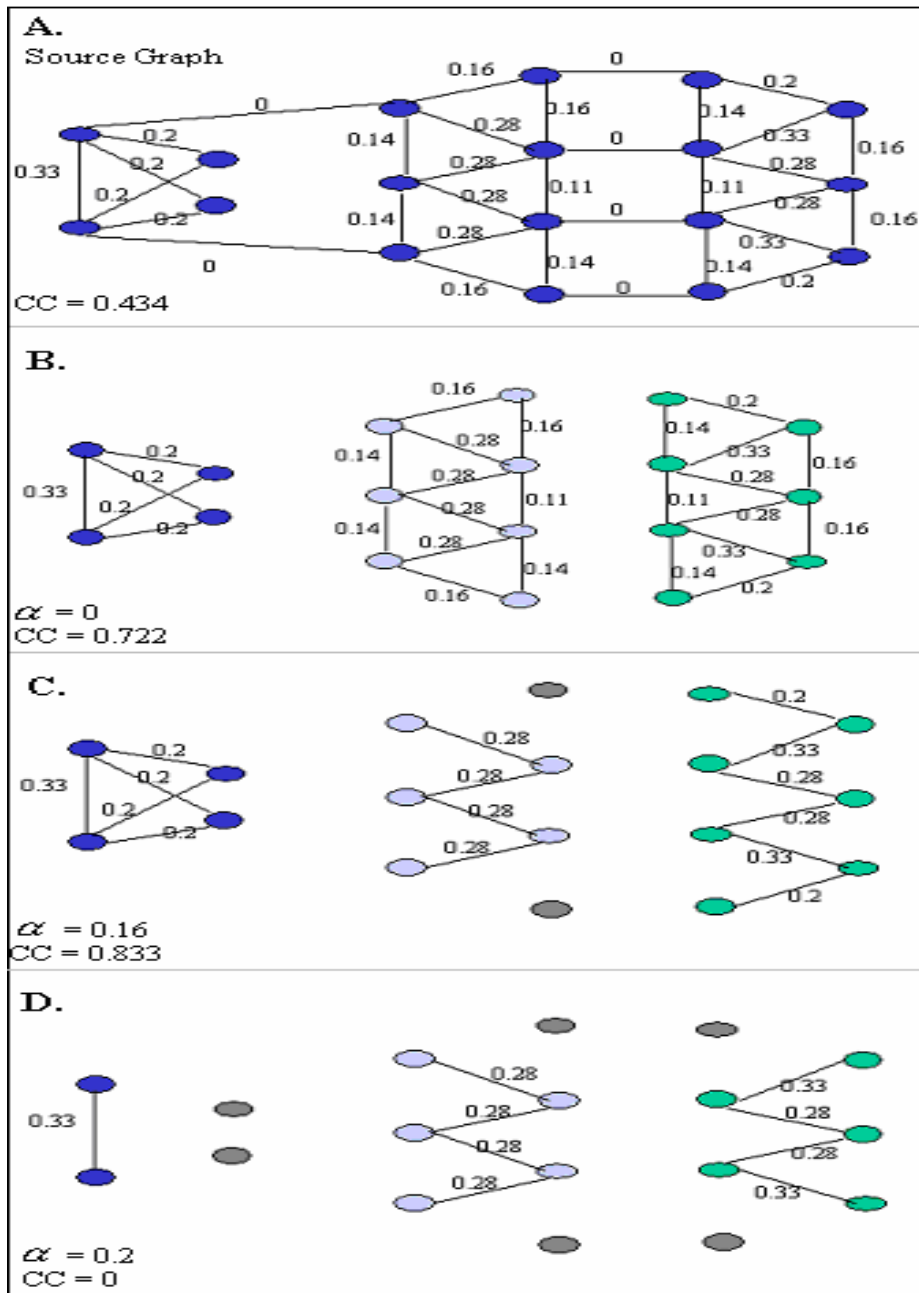
Finding an arbitrary cluster flow :

Input: graph $G=(V,E)$;

Output: A cluster $C \subset G$.

1. put arbitrary vertex v into queue Q ; $C=\emptyset$;
2. loop while Q is not empty
 - a. get vertex u from Q ; add u to C ;
 - b. loop for each $w \in N(u)$
 - if $\beta(u, w) > \alpha$ and $w \notin Q \cup C$
put w into Q ;
 - end if.

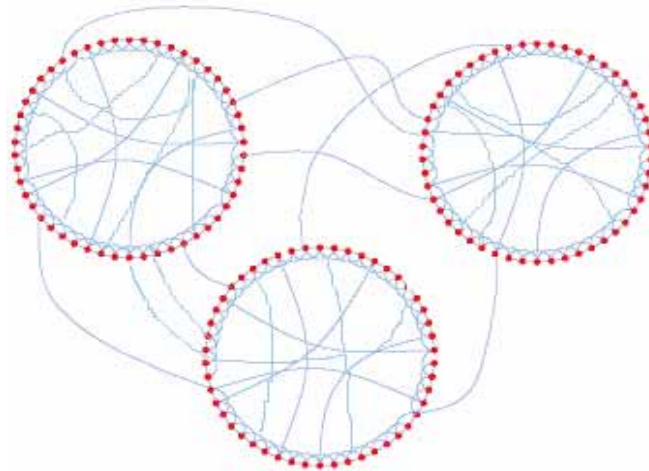
ICRA Example



Small World Simulation

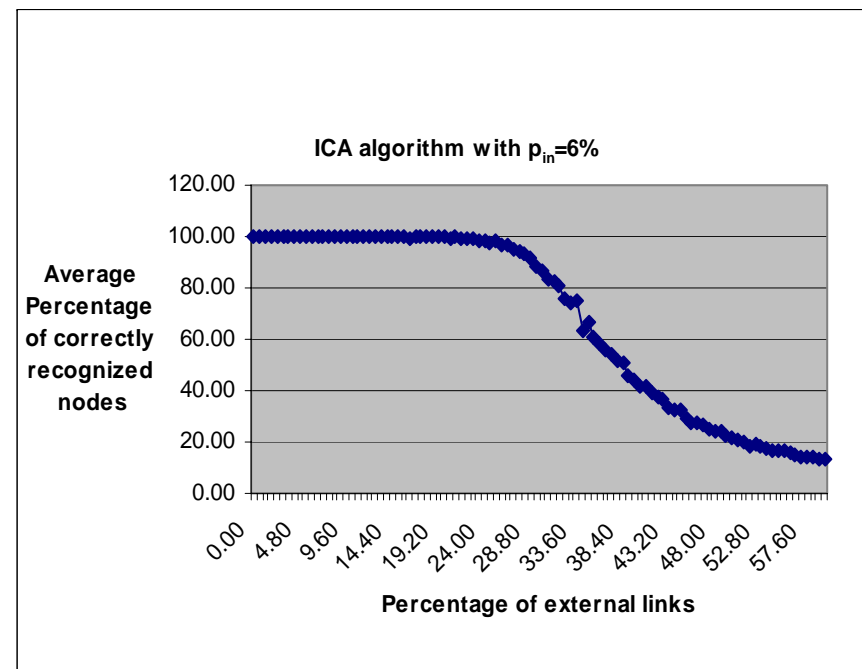
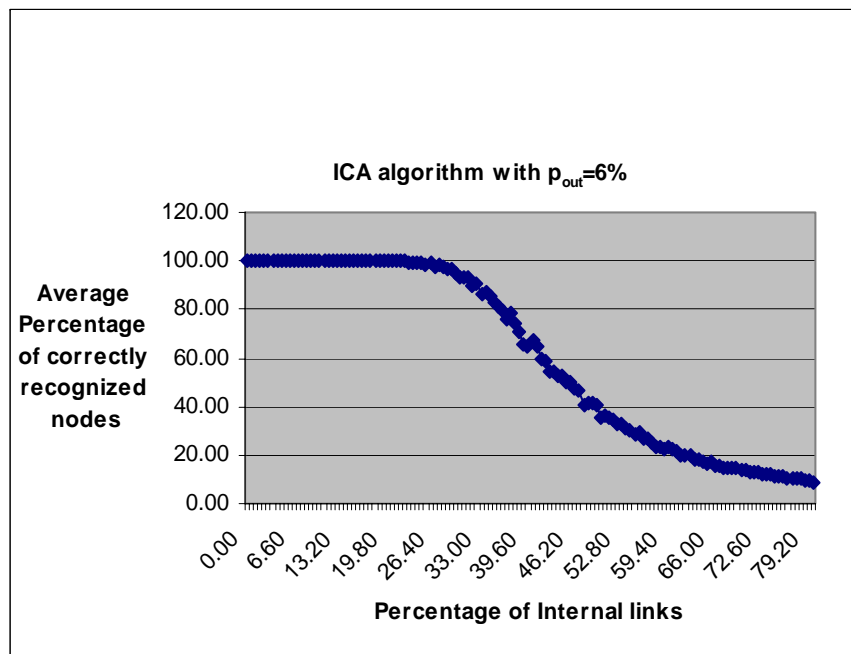
Extension of Watts and Strogatz Small World Model :

Collection of ring lattices with randomly reconnected P_{in} links inside the ring lattice and P_{out} reconnected links between the rings.



Small World simulation Results

Typical behavior of the ICA ability to recognize clusters under increasing link randomization.



Small World simulation Results

- We focus on ICRA recognition of at least 80% of the nodes. We define $P_{critical}$ as the value of p_{out} when the percentage of correctly recognized nodes starts to go down below 80%.
- $p_{in}=0.05\%$. (For different p_{in} tested changes are not significant).

p_{in} (%)	$P_{critical}$ (%)	Standard Deviation (%)
0.05	37.8	5.78
0.15	37.8	5.67
0.35	37.2	4.82
0.75	37.2	5.68
1.15	36.6	5.54
2	36.6	3.93
6	32.4	3.47

Small World simulation Results

Exists a cluster separation level α_c that provides best results of ICA cluster recognition. For our simulation model $\alpha_c = 0.1$.

α	$P_{critical}$ (%)	Standard Deviation (%)
0	7.8	9.36
0.1	37.8	5.78
0.2	21.6	6.53
0.3	9.6	5.27

Percentage of Correctly recognized nodes . α - $P_{critical}$ - Percentage of Correctly recognized nodes for $N=3200$, $z=8$, $M=16$, $N_c=200$, $p_{in}=0.05\%$. (average of 10 measurements).

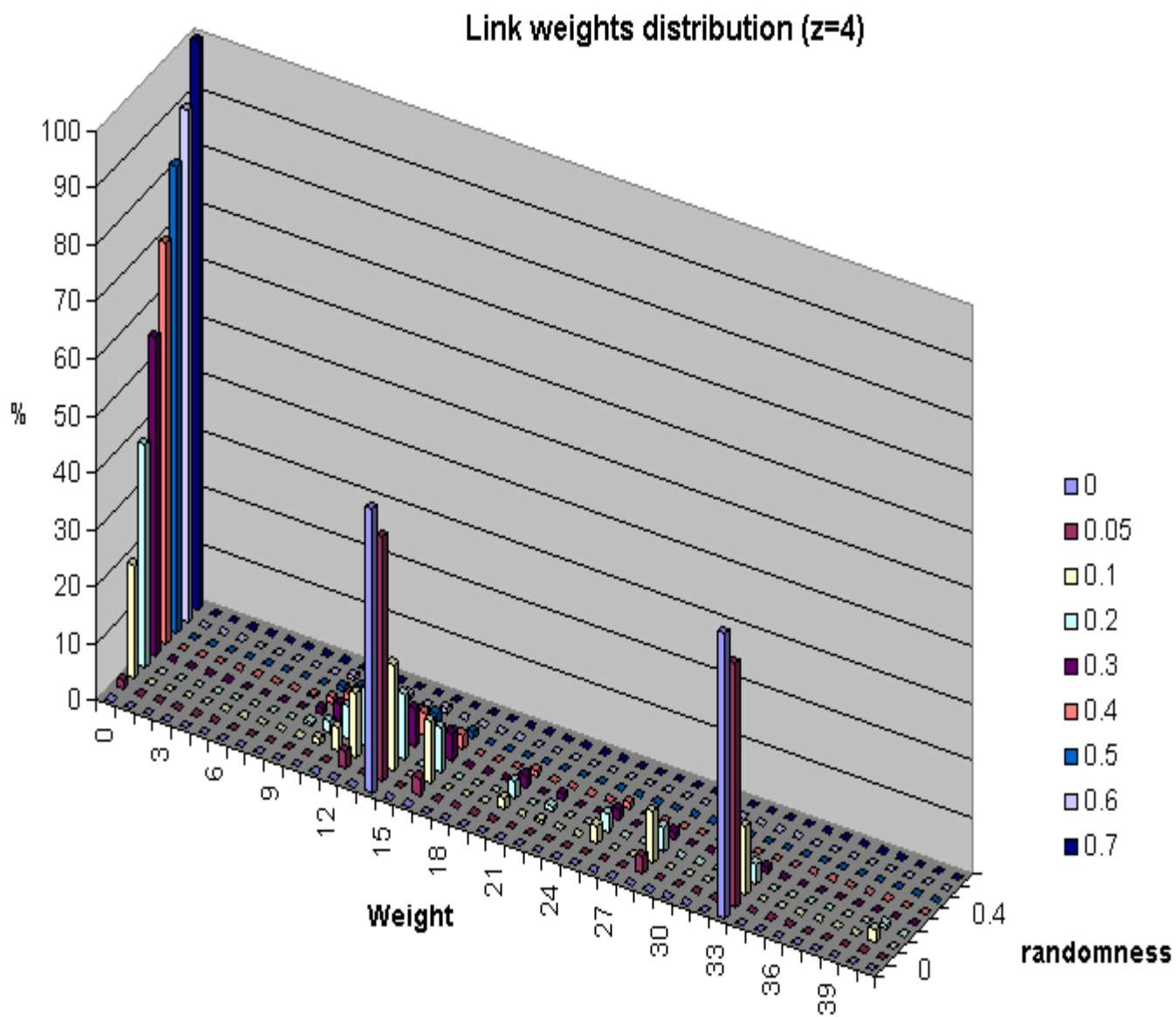
Small World simulation Results

α	$P_{critical}$ (%)	Standard Deviation (%)
0	7.2	6.78
0.1	53.4	5.07
0.2	23.4	3.37
0.3	13.8	1.7

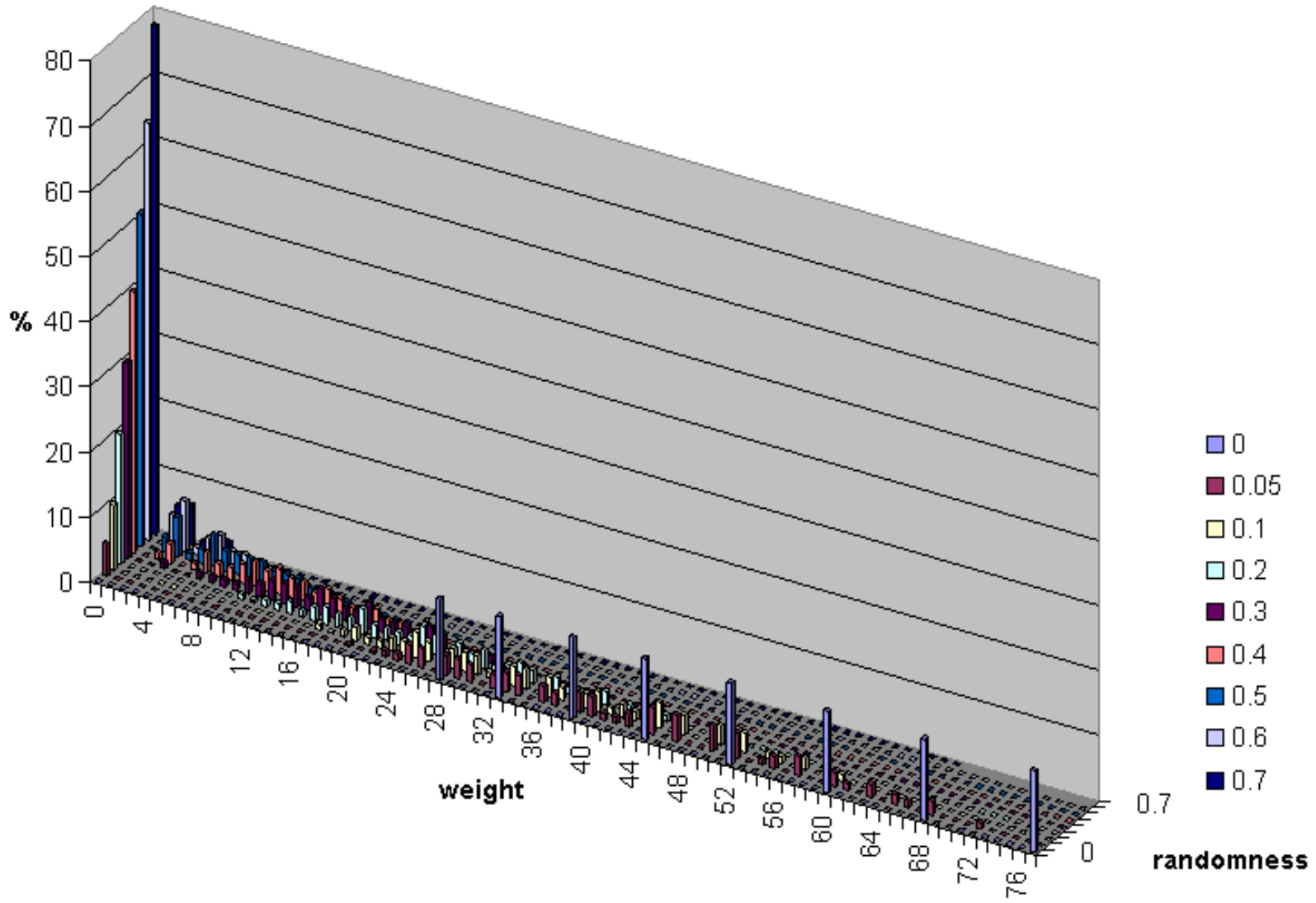
Percentage of Correctly recognized nodes . α - $P_{critical}$ - Percentage of Correctly recognized nodes for $N=6000$, $z=10$, $M=20$, $N_c=300$, $p_{in}=0.05\%$. (average of 10 measurements).

α	$P_{critical}$ (%)	Standard Deviation (%)
0	12	5.75
0.1	37.8	3.3
0.2	21.6	1.91
0.3	9.6	3.99

Percentage of Correctly recognized nodes . α - $P_{critical}$ - Percentage of Correctly recognized nodes for $N=6400$, $z=8$, $M=32$, $N_c=200$, $p_{in}=0.05\%$. (average of 10 measurements).



Link weight distribution (z=12)

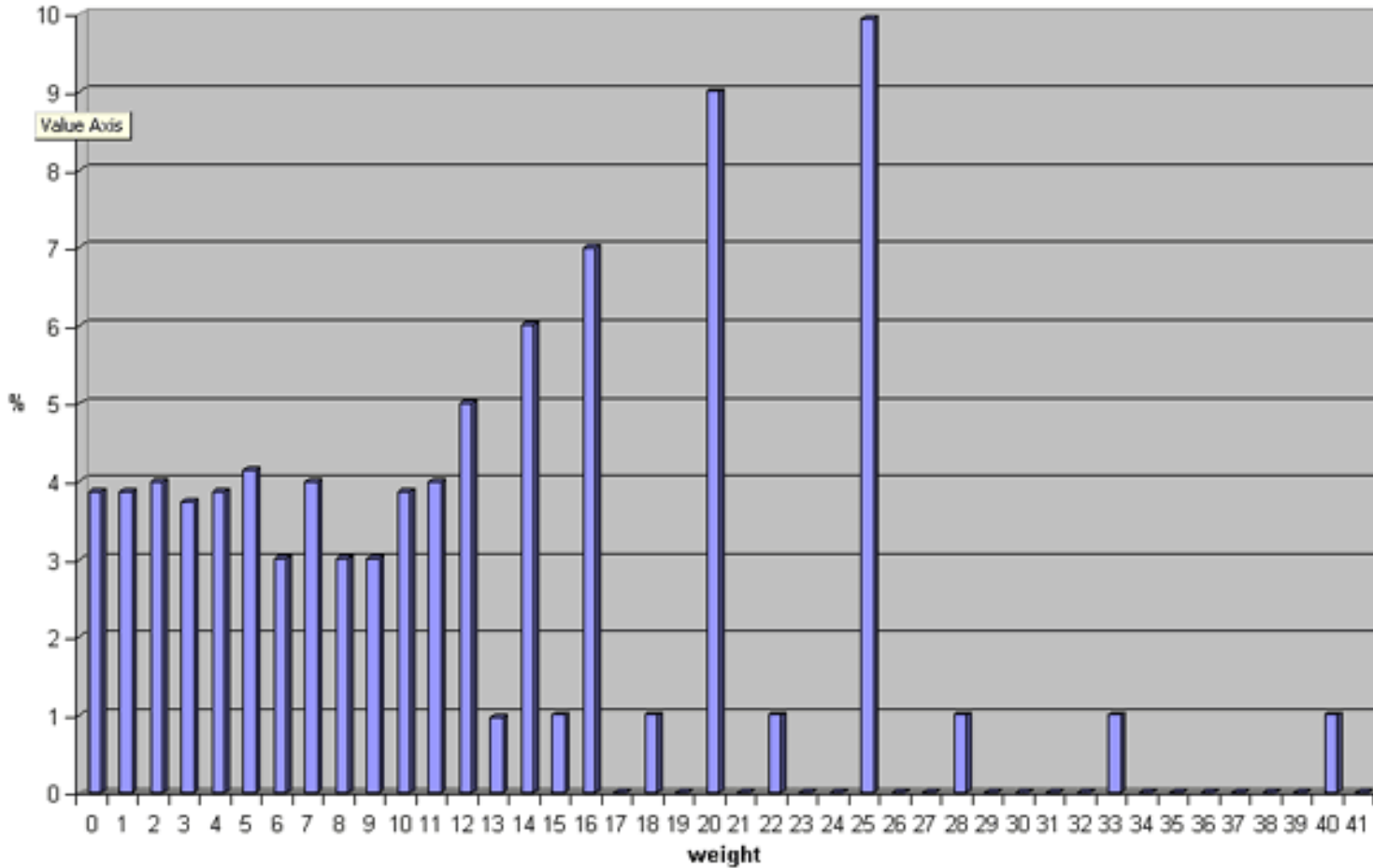


Small World simulation Results Cont.

As the value of z grows, there is cluster identification for higher values of p_{out}

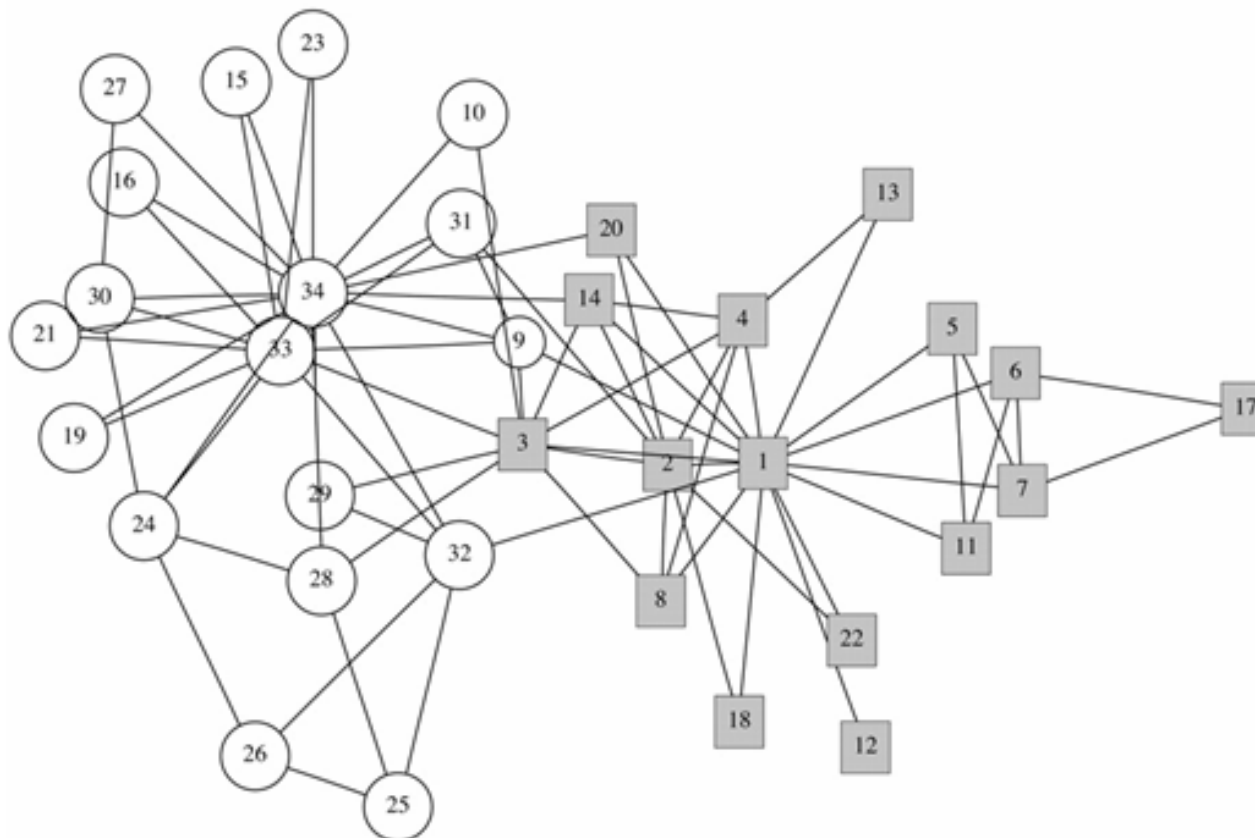
z	$P_{critical}$ (%)	Standard Deviation (%)
4	7.8	4.46
6	16.8	3.84
8	37.8	5.78
10	55.8	4.06
12	56	4.76
14	66	5.15
16	76.2	3.35
18	70.8	3.05
20	78.6	3.79

Link weights distribution for triangular random graph
(on each step one node is added with two links to arbitrary chosen link)



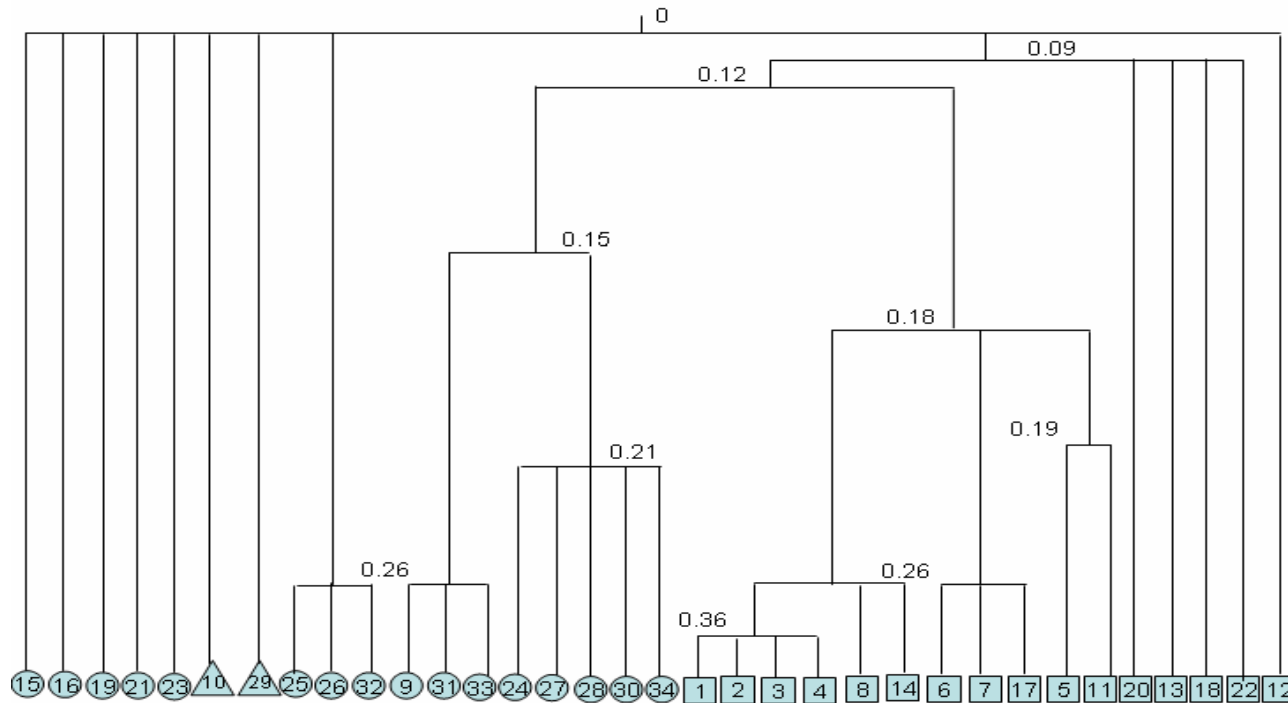
Zachary's Karate Club Study

The real-world friendship network from the well known Zachary's Karate Club :



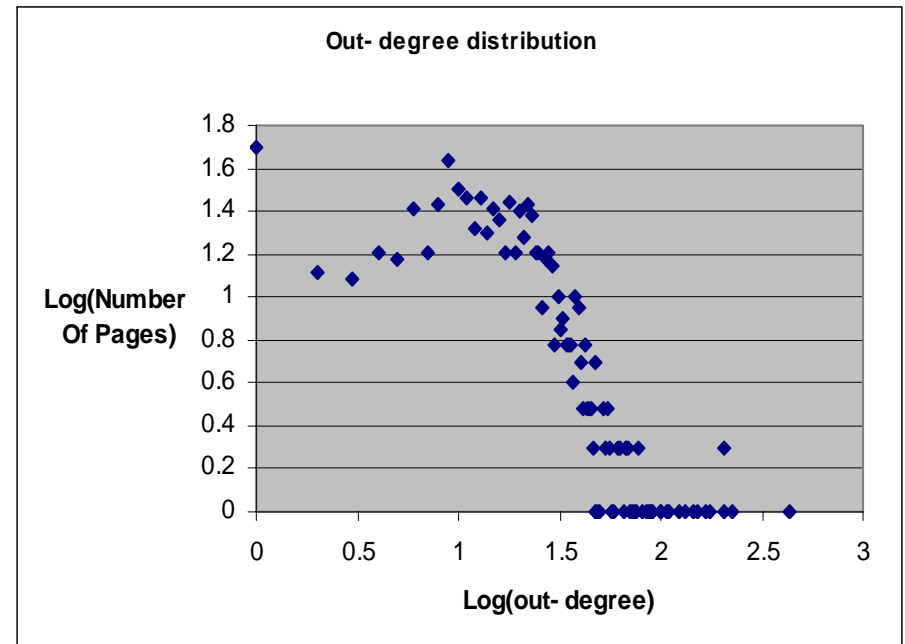
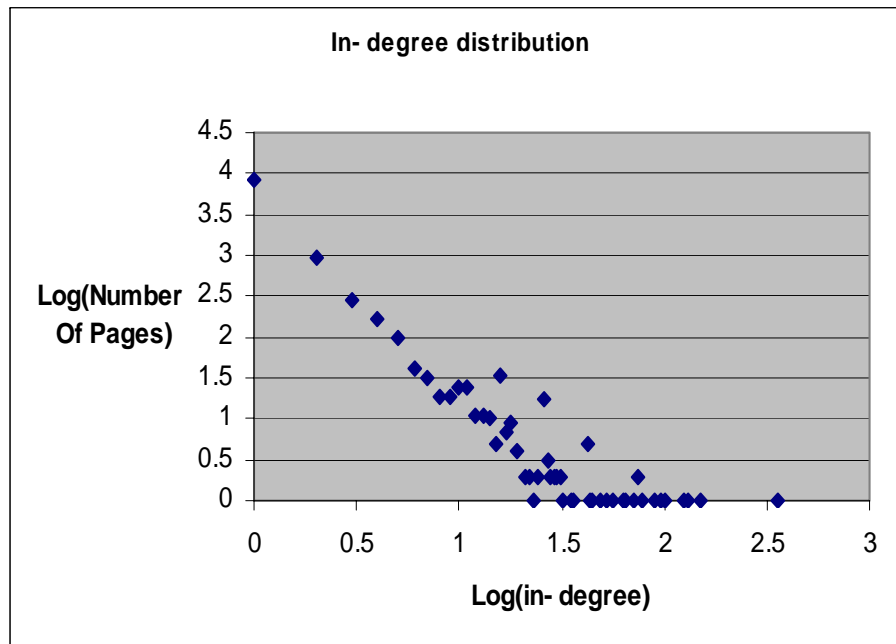
Zachary's Karate Club Study Cont.

The hierarchical tree of clustering produced
by our method :



Spidering the Web

Node degree distributions of Oxford university (www.ox.ac.uk) subgraph (consist of 10014 pages) that was spidered from the web.



Web Subgraphs Results

Example for spidering 4001 pages from Harvard University.
 Spidered graph $CC=0.007$. after the link removal for $\alpha = 0.2$
 the $CC=0.046$.

#	Domain name	Page title	Number Of pages
1	www.news.harvard.edu	Harvard University Office of News and Public Affairs	3
2	www.fas.harvard.edu	Faculty of Arts and Sciences, Harvard University	9
3	www.harvard.edu	Welcome to Harvard University	16
4	www.president.harvard.edu	Welcome to the Office of the President	10
5	www.atwork.harvard.edu	Harvard University Office of Human Resources	2
6	www.uos.harvard.edu	Harvard University, University Operations Services Home Page	5
7	lib.harvard.edu	Harvard Libraries	21
8	lib.harvard.edu	Harvard Libraries	2
9	www.admissions.college.harvard.edu	Harvard College Admissions Homepage	7
10	www.hsdm.harvard.edu	Harvard School of Dental Medicine	2
11	www.law.harvard.edu	Harvard Law School	2
12	atwork.harvard.edu	Harvard University Office of Human Resources	20

Advantages

1. Efficiency: ICA has a linear complexity $O(\langle E \rangle)$, or more exactly, an average complexity $O(\langle z^2 \rangle / V)$, where z is the average number of links per node in the graph.
2. Locality: No need to know full graph or number of clusters for local cluster recognition.
3. Applicability: Applicable for graphs with different nature ("big" and "small", power-law. etc.)
4. Tool ability: Helps to recognize a Small World subgraphs in not Small World graph.

Conclusion

- The success of this method demonstrate that the local link correlation is connected to the cluster structure of the graph.
- **Clustering weight of links distribution:** is it a new important metrics of complex networks ?
- ICRA has significant advantages compare to other clustering algorithms.
- Still, it must be emphasized that ICRA applicability is for special kind of graphs.

Thank you.

**For contacts:
igor kanovsky, igork@yvc.ac.il,
<http://www.yvc.ac.il/ik/>**