Τυχαία Γραφήματα - Μικρόκοσμοι και Αυτοόμοια Δίκτυα

Χρόνης Μωυσιάδης Ομότ. Καθηγητής ΑΠΘ
Βασίλης Καραγιάννης ΕΔΙΠ Τμ. Μαθηματικών

Τυχαία Δίκτυα

Ένα τυχαίο δίκτυο σχηματίζεται από ένα σύνολο $V = \{v_1, v_2, \ldots, v_n\}$, n κορυφών, στο οποίο προσθέτουμε ακμές με τυχαίο τρόπο.

Υπάρχουν διάφοροι τρόποι να επιλέξουμε τυχαία τις ακμές που οδηγούν σε διαφορετικά μοντέλα και σε διαφορετικές κατανομές πιθανοτήτων στα γραφήματα που προκύπτουν.

Τα συνήθη μοντέλα που μελετήθηκαν είναι:

Το μοντέλο Erdős–Rényi ($G(n,M)$)

Στο μοντέλο $G(n,M)$ αντιστοιχίζεται ίση πιθανότητα σε όλα τα γραφήματα που έχουν ακριβώς M ακμές ($0 \leq M \leq N$), όπου $N = \binom{n}{2}$. Αφού το ζητούμενο είναι να σχηματιστεί γράφημα με M ακμές, οι οποίες είναι από τις N υπάρχουσες, άρα υπάρχουν $\binom{N}{M}$ τρόποι επιλογής αυτών των M ακμών, που είναι ισοπίθανες.

Έτσι ο δειγματοχώρος $G(n,M)$ περιέχει $\binom{N}{M}$ στοιχεία καθένα με πιθανότητα $\binom{N}{M}^{-1}$.

Το M, σχεδόν πάντα, είναι συνάρτηση του n, δηλαδή $\mathrm{M} = \mathrm{M}(n)$ και τότε το μοντέλο συμβολίζεται $\mathrm{G}(\mathrm{n}, \mathrm{M}(n))$.

Το μοντέλο αυτό ορίστηκε αρχικά στην εργασία *“On Random Graphs I”* (1959) των Erdős και Rényi (από όπου προέρχεται και η ονομασία)

Το μοντέλο Gilbert ($G(n,p)$)

Στο μοντέλο $G(n,p)$ θεωρείται ότι κάθε μία από τις $N = \binom{n}{2}$ δυνατές ακμές που μπορούν να συνδέσουν δύο από τις n κορυφές, επιλέγεται με πιθανότητα p, όπου $0 < p < 1$. Ισοδύναμα μπορούμε να θεωρήσουμε ότι έχουμε το πλήρες γράφημα K_n και από αυτό διαγράφουμε τις ακμές του με πιθανότητα $\mathrm{q} = 1 - p$.

Ένα στοιχείο G_0 στο δειγματοχώρο $G(n,p)$ που έχει n κορυφές και m ακμές έχει πιθανότητα εμφάνισης

$$P(\{G_0\}) = P(G = G_0) = p^m \cdot (1-p)^{N-m}$$

Αν $p = \frac{1}{2}$ τότε $G(n, \frac{1}{2})$ είναι ο δειγματοχώρος που περιέχει όλα τα γραφήματα n κορυφών G^n με την ίδια πιθανότητα

$$P(G = G_0) = \left(\frac{1}{2}\right)^{\mathrm{N}}$$

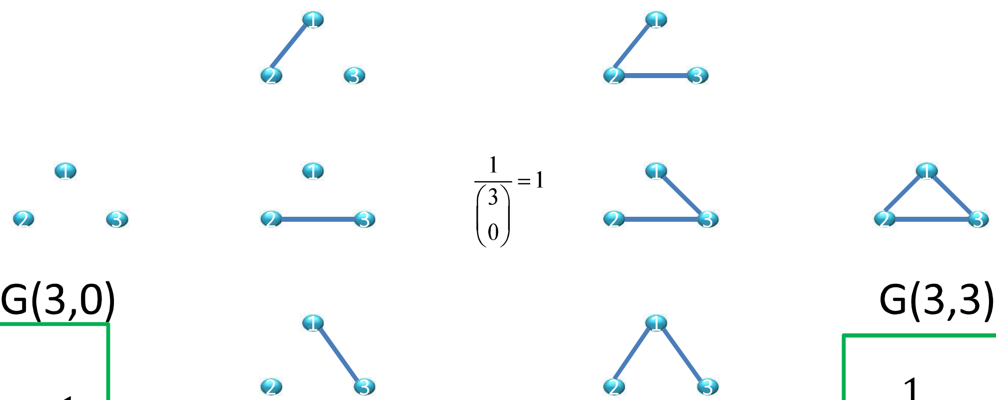
Το μοντέλο αυτό προτάθηκε από τον Edgar Gilbert στην εργασία του *“Random Graphs”* (1959).

Ο δειγματοχώρος G^n

- Θεωρούμε ένα τυχαίο γράφημα να σχηματίζεται ως εξής:
- Ξεκινούμε από το γράφημα n κορυφών χωρίς ακμές (G_0). Φέρουμε μία ακμή ενώνοντας δύο από τις κορυφές, άρα σχηματίζεται το G_1 (γράφημα με μία ακμή). Συνεχίζουμε, προσθέτοντας σε κάθε βήμα μία ακμή, σχηματίζοντας την ακολουθία γραφημάτων:
$$G_0 \subset G_1 \subset G_2 \subset \cdots \subset G_N$$
όπου G_t γράφημα με ακριβώς t ακμές.
- Υπάρχουν $\mathrm{N}!$ τέτοιες ακολουθίες που συνιστούν το δειγματοχώρο G^n.
- Ο Bollobas (1985) διαπίστωσε ότι οι τρεις χώροι που περιγράψαμε είναι περίπου ταυτόσημοι (closely related)

Παράδειγμα στο χώρο $\mathrm{G}(3,\mathrm{m})$

- Εδώ έχουμε 4 δειγματοχώρους για $m = 0, 1, 2, 3$, όλοι με ίσες πιθανότητες.



G(3,0)

$$\frac{1}{\binom{3}{0}} = 1$$

G(3,1)

$$\frac{1}{\binom{3}{1}} = \frac{1}{3}$$

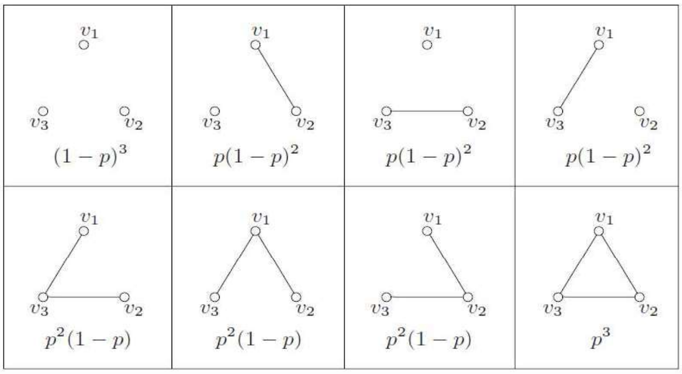
G(3,2)

$$\frac{1}{\binom{3}{2}} = \frac{1}{3}$$

G(3,3)

$$\frac{1}{\binom{3}{3}} = 1$$

Παράδειγμα στο χώρο G(3, p)



Ποια η πιθανότητα ότι ένα τυχαίο γράφημα G στο χώρο G(3, p) να είναι συνδετικό;

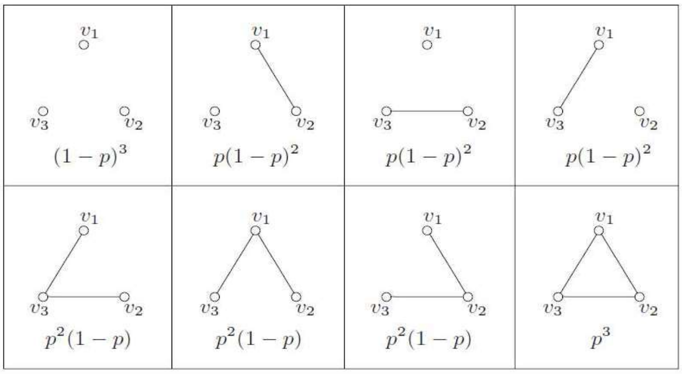
$$P(A) = 3p^2(1-p) + p^3 = p^2(3-2p)$$

Ποια η πιθανότητα ότι ένα τυχαίο γράφημα G στο χώρο G(3, p) να είναι διμερές;

$$P(B) = 1 - p^3$$

π.χ. αν $p = \frac{1}{2}$, τότε $P(A) = \frac{1}{2}$, $P(B) = \frac{7}{8}$ και $P(AB) = \frac{3}{8}$ δηλαδή τα γεγονότα το γράφημα είναι διμερές και είναι συνδετικό δεν είναι ανεξάρτητα.

Παράδειγμα στο χώρο G(3, p)



Ποια η μέση τιμή της τυχαίας μεταβλητής που δίνει το πλήθος των παραγόντων του γραφήματος;
Εδώ οι παράγοντες είναι είτε 1, είτε 2 είτε 3 με πιθανότητα:

k	1	2	3
P(X=k)	$p^3 + 3p^2(1-p)$	$3\,p\,(1-p)^2$	$(1-p)^3$

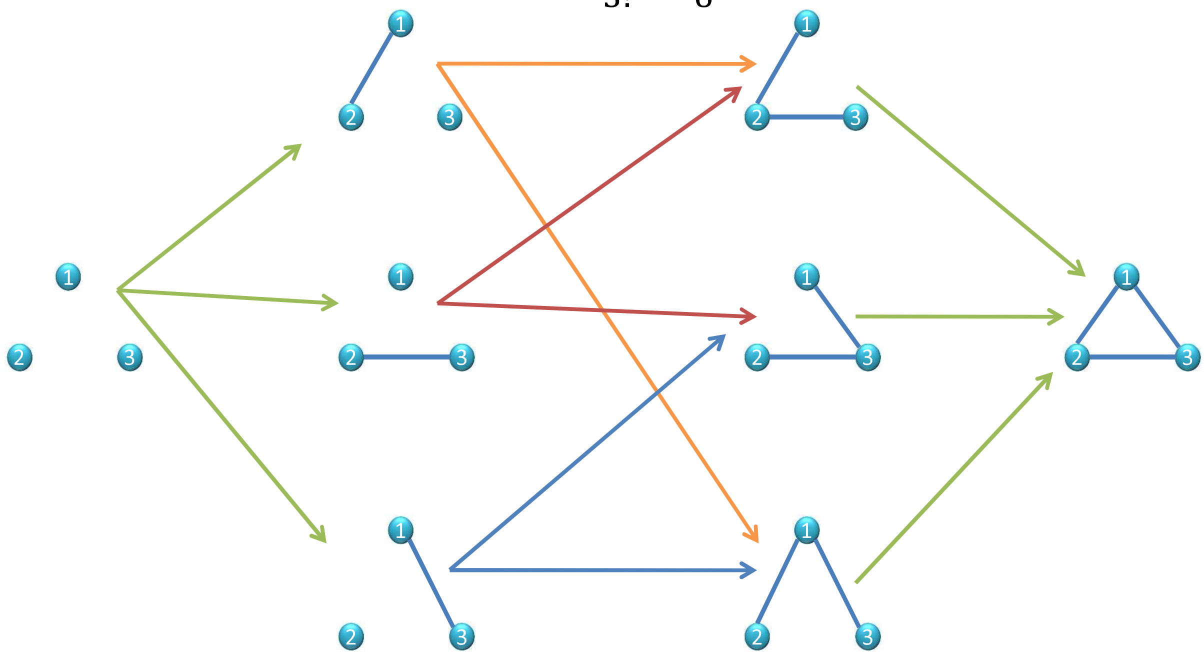
και άρα EX $= 1 \cdot P(X = 1 + 2 \cdot P(X = 2 + 3 \cdot P(X = 3) =$

$$= p^3 + 3p^2(1-p) + 2 \cdot 3\,p\,(1-p)^2 + 3 \cdot (1-p)^3 =$$
$$3 - 3p + p^3$$

Παράδειγμα

Ο δειγματοχώρος G^3 με $N! = \binom{3}{2}!$=6 στοιχεία,

και $P(G_0 \subset G_1 \subset G_2 \subset G_3) = \frac{1}{3!} = \frac{1}{6}$



Η κατανομή των βαθμών στο $G(n,p)$

Αν μία κορυφή έχει βαθμό d τότε σημαίνει ότι d από τους n-1 δυνατούς γείτονες, που επιλέγονται με $\binom{n-1}{d}$ τρόπους, συνδέθηκαν με πιθανότητα p^d και οι υπόλοιποι δεν συνδέθηκαν με πιθανότητα $(1-p)^{n-1-d}$. Άρα η πιθανότητα να συμβαίνει αυτό είναι:

$$P(d) = \binom{n-1}{d} p^d (1-p)^{n-1-d}$$

δηλαδή η τ.μ. που μετρά το βαθμό μιας κορυφής ακολουθεί διωνυμική κατανομή με μέση τιμή $\bar{d} = (n-1)p$ και διασπορά $\sigma^2 = (n-1)p(1-p)$.

Αν τώρα αφήσουμε το n να τείνει στο άπειρο, τότε η κανονική τείνει σε Poisson,

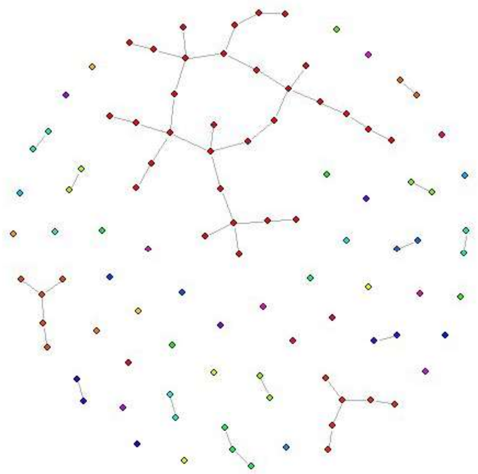
$$P(d) = e^{-(n-1)p} \frac{\left((n-1)p\right)^d}{d!}$$

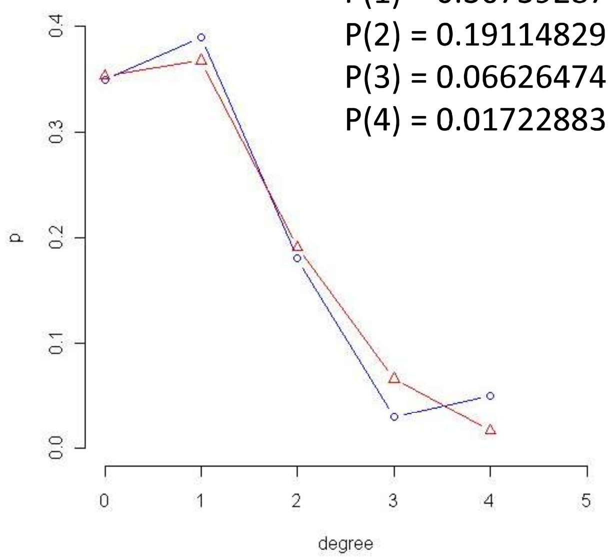
Αντίστροφα, αν θεωρήσουμε ένα τυχαίο δίκτυο στο οποίο οι βαθμοί των κορυφών είναι ισόνομες και ανεξάρτητες κατανομές Poisson(λ) με λ=n(p-1) (i.i.d. Poisson(λ) distributions), τότε διαπιστώνεται ότι αυτό είναι ισοδύναμο με το Erdős–Rényi μοντέλο, του n τείνοντος στο άπειρο.

Για το λόγο αυτό τα μοντέλα αυτά λέγονται επίσης μοντέλα Poisson.

Παράδειγμα G(100, 0.01)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
0.00	0.00	1.00	1.04	2.00	4.00	1,02





Κατανομή Βαθμών

P(0) = 0.35345468
P(1) = 0.36759287
P(2) = 0.19114829
P(3) = 0.06626474
P(4) = 0.01722883

Άσκηση

Έστω G τυχαίο γράφημα στο G(50, 0.02)
α) Ποιο μέρος των κορυφών του G αναμένεται να έχουν βαθμό 1;
β) Ποιο μέρος των κορυφών του G αναμένεται να έχουν βαθμό τουλάχιστον ίσο με 1;
γ) Ποια η πιθανότητα να υπάρχει κορυφή στο G με βαθμό 15;

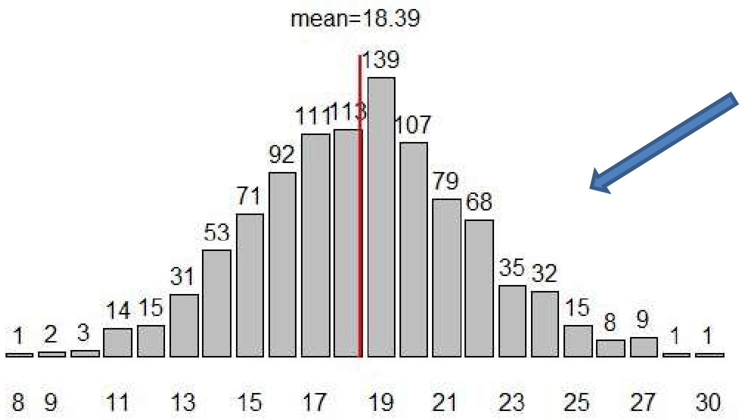
$P(1) = \binom{49}{1} 0.02^1 0.98^{48}$=0.37, $50 * 0.37 = 18.5,$

δηλ. 18-19 κορυφές αναμένεται να έχουν βαθμό 1

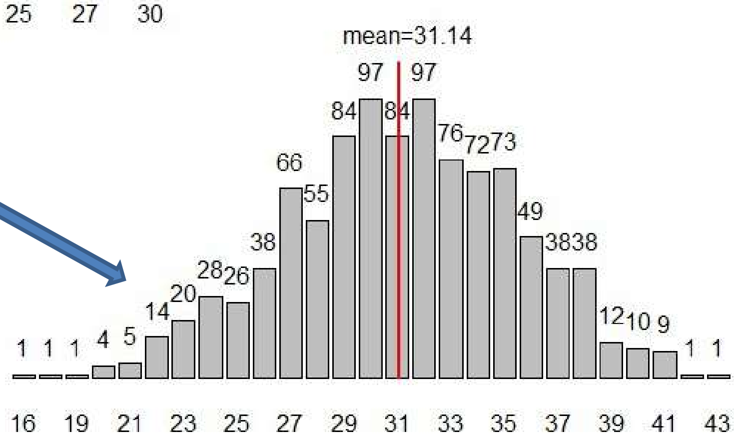
$1 - P(0) = 1 - \binom{49}{0} 0.02^0 0.98^{49}$=1-0.37, $50 * 0.63 = 31.5$

Προσομοίωση για 1000 επιλογές τυχαίων δικτύων

Πλήθος κορυφών βαθμού 1



Πλήθος κορυφών βαθμού τουλάχιστον ίσο με 1



Ιδιότητες του μοντέλου $G(n,p)$

Ποιες είναι οι ιδιότητες (τοπολογικά χαρακτηριστικά) του μοντέλου $G(n,p)$;

- Είναι συνδετικό;
- Έχει κύκλους ή είναι δέντρο-δάσος;
- Υπάρχει συσχέτιση των βαθμών των γειτόνων;
- Ποιες κορυφές παίζουν κεντρικό ρόλο;
- Υπάρχουν κλίκες;
- Υπάρχουν κοινότητες;

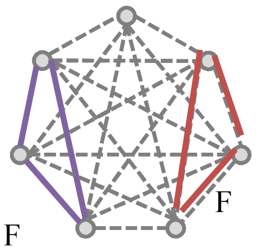
Πολλές από τις ιδιότητες αυτές εμφανίζονται ξαφνικά καθώς η πιθανότητα p αυξάνει από 0 έως 1.
Υπάρχει μία χαρακτηριστική τιμή, ένα κατώφλι, που από κει και μετά εμφανίζεται κάποια ιδιότητα.

Συνάρτηση κατωφλίου για υπογραφήματα

Αν Q η ιδιότητα ένα τυχαίο γράφημα περιέχει υπογράφημα με k κορυφές και l ακμές, αποδεικνύεται ότι η συνάρτηση

$$p_Q(n) = cn^{-\frac{k}{l}}$$

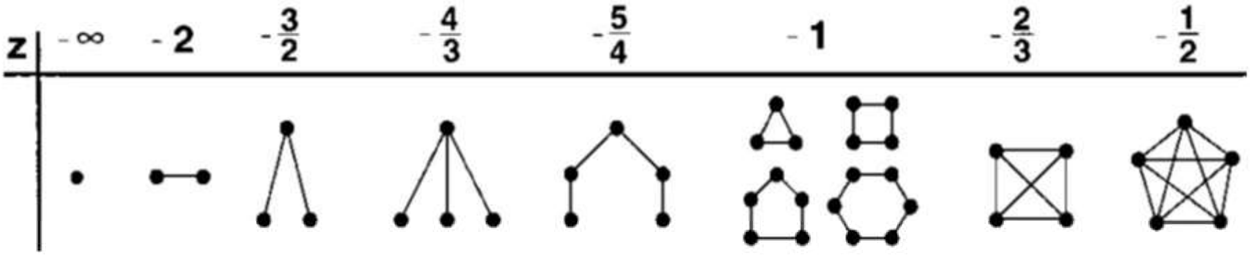
είναι συνάρτηση κατωφλίου.



Π.χ. για να εμφανίζονται τρίγωνα στο τυχαίο γράφημα πρέπει $p = \frac{c}{n}$, (διότι στο τρίγωνο $k = 3$ και $l = 3$)

Συνάρτηση κατωφλίου

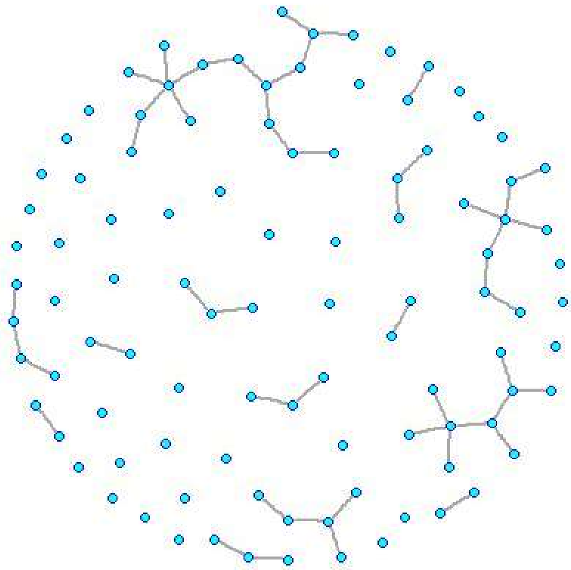
Οι R.Albert and A-L. Barabasi σε εργασία τους του 2002, βρήκαν ότι αν η πιθανότητα p είναι ανάλογη της δύναμης n^z , όπου το z παίρνει τιμές που φαίνονται στην πρώτη γραμμή του παρακάτω σχήματος :



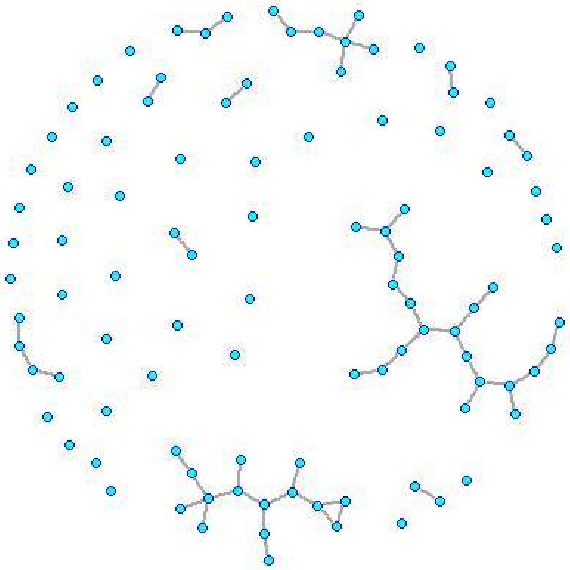
Τότε στο τυχαίο γράφημα θα αρχίσουν να εμφανίζονται δομές όπως αυτές που φαίνονται στη δεύτερη γραμμή του σχήματος.

$n = 100, M = 50$ $\qquad$ $n = 100, p = 0.01$

$$\binom{n}{2} p \approx M \text{ άρα } p \approx \frac{2M}{n(n-1)}$$

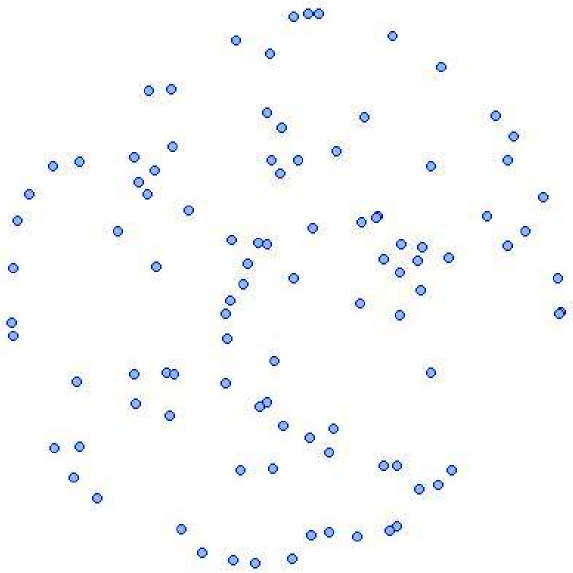


βαθμοί	0	1	2	3
	34	38	22	6

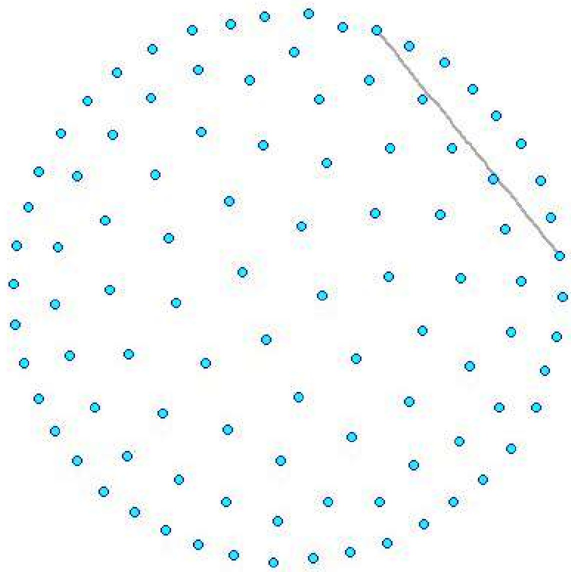


βαθμοί	0	1	2	3	4
	38	33	18	9	2

$n = 100, \qquad p = 100^{-2}$ $\qquad$ $n = 100, \qquad p = 5 \cdot 100^{-2}$

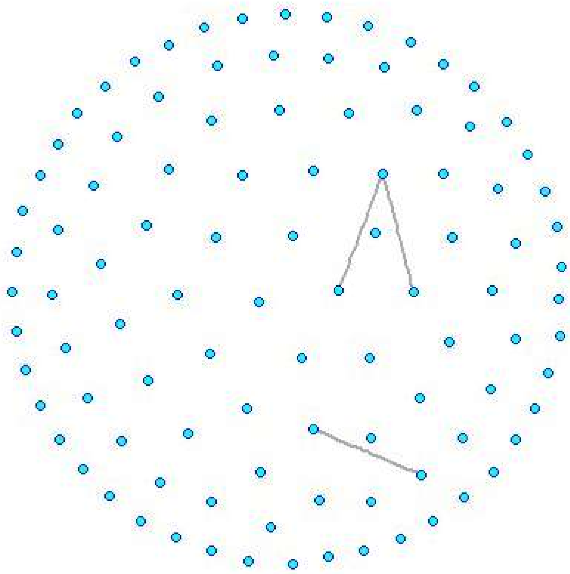


βαθμοί	0
	100



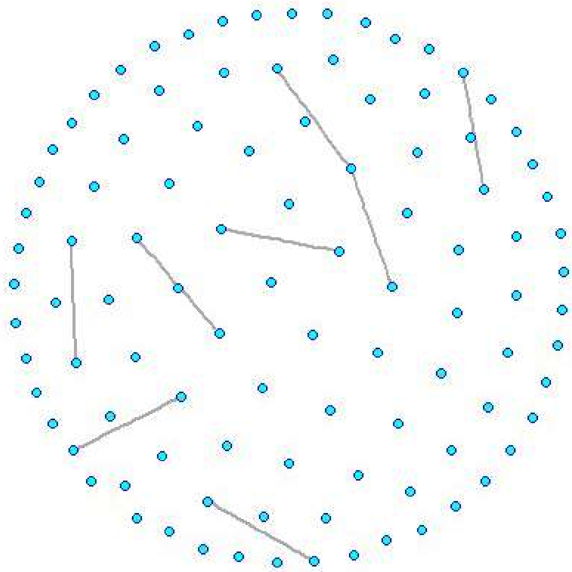
βαθμοί	0	1
	98	2

$n = 100, \qquad p = 100^{-3/2}$



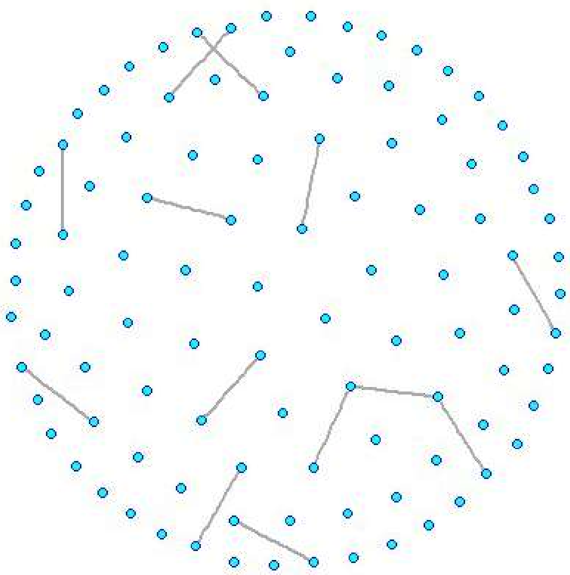
	0	1	2
βαθμοί	95	4	1

$n = 100, \qquad p = 100^{-4/3}$



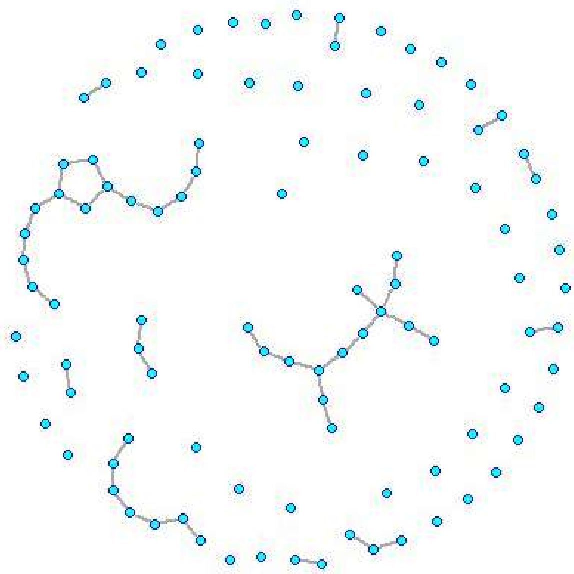
	0	1	2
βαθμοί	85	14	1

$n = 100, \qquad p = 100^{-5/4}$



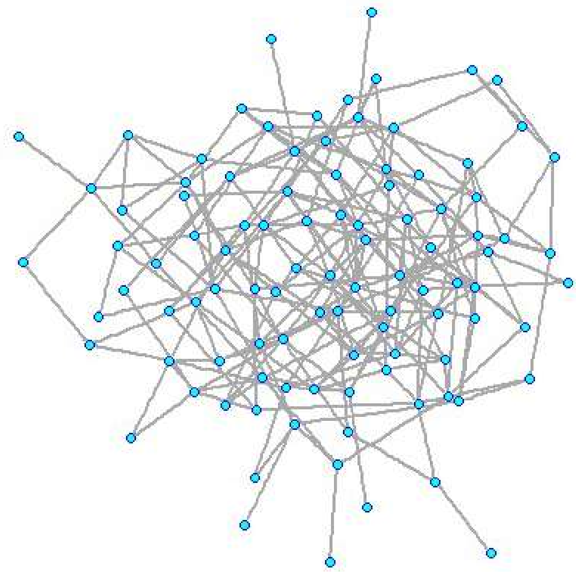
	0	1	2
βαθμοί	76	22	2

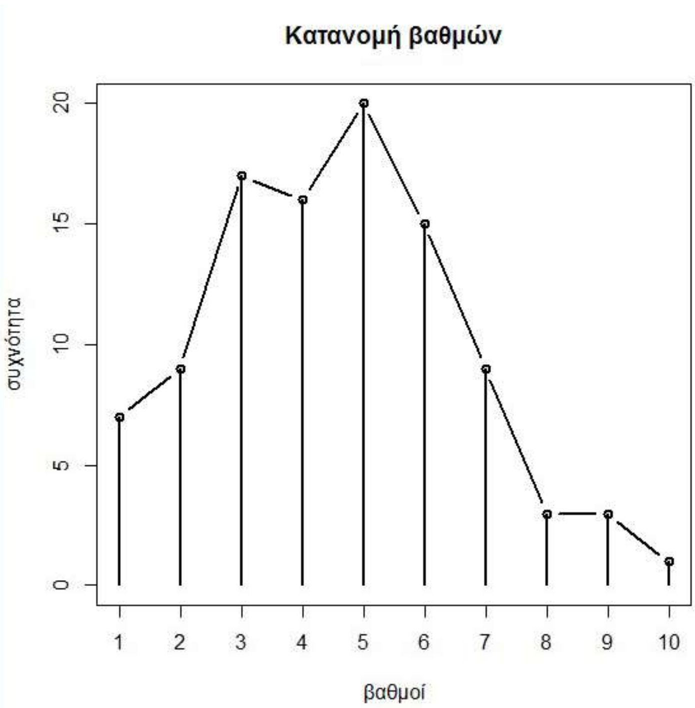
$n = 100, \qquad p = 1.1 \cdot 100^{-1}$



	0	1	2	3	4
βαθμοί	44	27	25	3	1

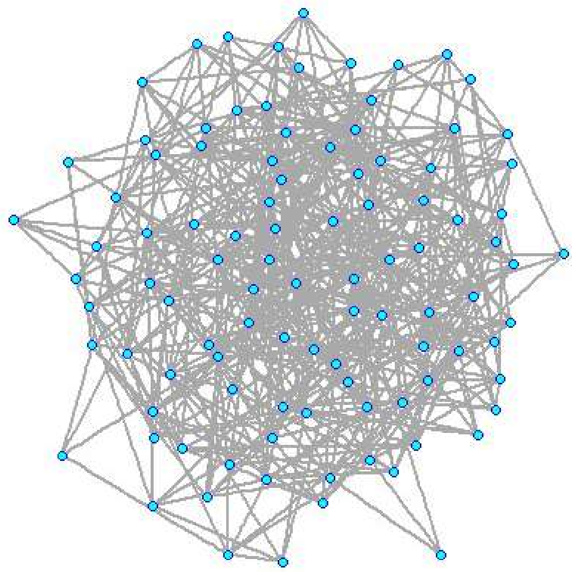
$n = 100, \qquad p = 100^{-2/3}$

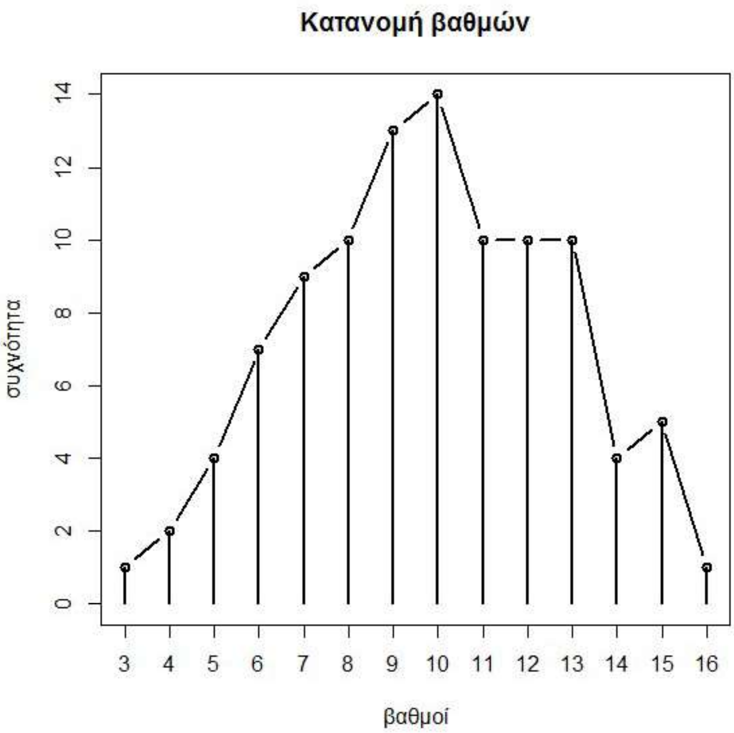




βαθμοί	1	2	3	4	5	6	7	8	9	10
	7	9	17	16	20	15	9	3	3	1

$n = 100, \qquad p = 100^{-1/2}$





βαθμοί	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	1	2	4	7	9	10	13	14	10	10	10	4	5	1

Μέσος βαθμός, διάμετρος κτλ

- Σχέση μεταξύ $G(n,p)$ και $G(n,M)$

  $M = \binom{n}{2} p$ άρα $p = \frac{2M}{n(n-1)}$ και $\overline{d} = \frac{2M}{n} = p(n-1)$

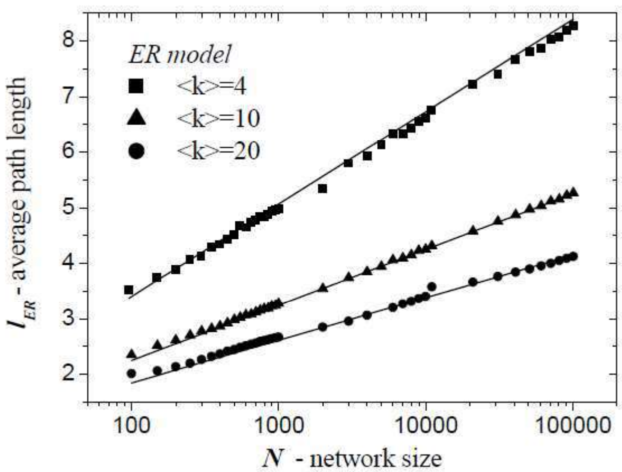
- Λόγω της ξαφνικής εμφάνισης ιδιοτήτων, η διάμετρος εκτιμάται διαφορετικά για διάφορες τιμές της p:
  - Αν $p < \frac{1}{n}$, τότε η διάμετρος είναι αυτή ενός παράγοντα δέντρου και είναι $L_G = \log(n)\,/\log(p(n-1))$
  - Αν $p > \frac{1}{n}$, τότε εμφανίζεται γιγάντια συνιστώσα και αν $p \geq \frac{3.5}{n}$ τότε η διάμετρος είναι αυτή της γιγάντιας συνιστώσας και είναι $L_G = \log(n)\,/\log(p(n-1))$
  - Αν $p > \frac{log(n)}{n}$, τότε το γράφημα είναι σχεδόν βέβαια συνδετικό και η διάμετρος εκτιμάται ίση με $L_G = (1+\varepsilon)\log(n)\,/\log(p(n-1))$

Μέσο μήκος μονοπατιών

Ο Fronczak κ.α. το 2004 έδειξε ότι για μεγάλα τυχαία γραφήματα το μέσο μήκος μονοπατιού (ή η μέση απόσταση μεταξύ των κορυφών) ισούται με:

$$\bar{l} = \frac{logn-\gamma}{logpn} + 0.5,$$

όπου $\gamma \approx 0.5772$ η σταθερά Euler



Δηλαδή αυξανομένου του n και ανάλογα με το p, το μέσο μήκος μονοπατιού αυξάνει, όμως η αύξηση είναι πολύ μικρή όπως φαίνεται στο σχήμα.

Πλήθος τριγώνων και τριάδων σε γράφημα

Έστω N_v συμβολίζει το σύνολο των γειτόνων της κορυφής v, και k_v τον πληθικό αριθμό του N_v (επομένως k_v ισούται με τον βαθμό της κορυφής v). Τότε το σύνολο των τριγώνων με κορυφή v, έστω $\tau(v)$, δίνεται από τον πληθικό αριθμό

$$\tau(v) = \left|\left\{\{u, w\} \in E:\ u, w \in N_v\right\}\right|,$$

δηλ. με το πλήθος των συνδέσεων μεταξύ των γειτόνων μεταξύ τους. Αθροίζοντας για όλες τις κορυφές βρίσκουμε το σύνολο των τριγώνων του G, ίσο με:

$$\tau(G) = \frac{1}{3}\sum_{v\in V} \tau(v)\ ,$$

διότι κάθε τρίγωνο το μετρούμε τρεις φορές, μία για κάθε κορυφή του.

Το πλήθος των τριάδων $\rho(v)$ (δηλαδή μονοπατιών μήκους 2) με κέντρο μια κορυφή v ισούται με το πλήθος των συνδυασμών των γειτόνων ανά δύο, δηλαδή:

$$\rho(v) = \binom{k_v}{2} = \frac{k_v(k_v-1)}{2}\ ,$$

και αθροίζοντας βρίσκουμε όλες τις τριάδες του G, που είναι:

$$\rho(G) = \sum_{v\in V} \rho(v) = \sum_{v\in V} \frac{k_v(k_v-1)}{2}\ .$$

Συντελεστές και μέτρα

- Συντελεστής Σύμπλεξης
- Λόγος μεταβατικότητας
- Μέτρα Κεντρικότητας
  - Βαθμική κεντρικότητα
  - Ιδιοκεντρικότητα
  - Κατά Katz κεντρικότητα
  - Βαθμική κεντρικότητα Page
  - Κεντρικότητα Εγγύτητας
  - Ενδιάμεση Κεντρικότητα

Συντελεστής σύμπλεξης (clustering coefficient)

Ο συντελεστής σύμπλεξης μιας κορυφής v μετρά κατά πόσον οι γείτονές της απέχουν από το να αποτελούν κλίκα. Ορίζεται από τη σχέση:

$$C(v) = \frac{\tau(v)}{\rho(v)} = \frac{\pi\lambda\acute{\eta}\theta o\varsigma\ \tau\rho\iota\gamma\acute{\omega}\nu\omega\nu\ \mu\varepsilon\ \kappa o\rho\upsilon\varphi\acute{\eta}\ \tau\eta\nu\ v}{\pi\lambda\acute{\eta}\theta o\varsigma\ \tau\rho\iota\acute{\alpha}\delta\omega\nu\ \mu\varepsilon\ \mu\varepsilon\sigma\alpha\acute{\iota}\alpha\ \kappa o\rho\upsilon\varphi\acute{\eta}\ \tau\eta\nu\ v},$$

στην περίπτωση που ο βαθμός της v είναι $\mathrm{k}_v \geq 2$, ενώ θεωρείται ίσος με 0, αν ο βαθμός είναι το πολύ ίσος με 1.

Η ελάχιστη τιμή του $C(v)$ είναι 0 και πετυχαίνεται όταν το G είναι αστέρι με κορυφή v (τότε οι γείτονες της v σχηματίζουν πλήρως ασυνδετικό γράφημα).

Η μέγιστη τιμή του $C(v)$ είναι 1 και πετυχαίνεται όταν το G είναι κλίκα (τότε και οι γείτονες της v σχηματίζουν πλήρες γράφημα δηλ. κλίκα).

Ο μέσος όρος των συντελεστών σύμπλεξης όλων των κορυφών ονομάζεται συντελεστής σύμπλεξης το γραφήματος, δηλαδή:

$$C(\mathrm{G}) = \frac{1}{n} \cdot \sum_{v \in V} C(v) \quad .$$

Η ελάχιστη τιμή του $C(G)$ είναι 0 και πετυχαίνεται όταν το G είναι πλήρως ασυνδετικό γράφημα και η μέγιστη τιμή του είναι 1 που πετυχαίνεται όταν το G είναι πλήρες γράφημα).

Μεταβατικότητα

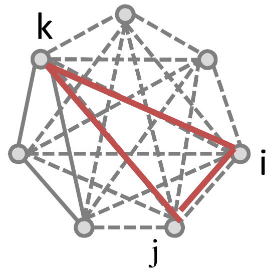
Σε ένα τυχαίο κοινωνικό δίκτυο Poisson ένα από τα ζητούμενα είναι αν, στην περίπτωση που ο k έχει φίλους τον i και τον j, οι i και j είναι φίλοι μεταξύ τους. Αν δηλαδή οι φίλοι μου είναι και μεταξύ τους φίλοι. Μια τέτοια τριάδα θεωρείται μεταβατική τριάδα. Προφανώς κάθε τρίγωνο δίνει τρεις μεταβατικές τριάδες.

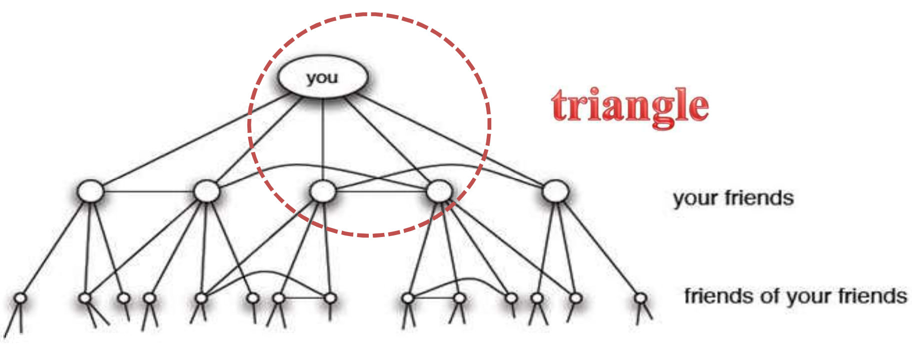
Αν η πιθανότητα στο δίκτυο είναι p τότε:

$$P(i\ \varepsilon\acute{\iota}\nu\alpha\iota\ \varphi\acute{\iota}\lambda o\varsigma\ \tau o\upsilon\ j) = P(\upsilon\pi\acute{\alpha}\rho\chi\varepsilon\iota\ \eta\ \alpha\kappa\mu\acute{\eta}\ \{i,j\}) = p$$

Οπότε:

$$P(i, j\ \varepsilon\acute{\iota}\nu\alpha\iota\ \varphi\acute{\iota}\lambda o\iota\ \delta o\theta\acute{\varepsilon}\nu\tau o\varsigma\ \acute{o}\tau\iota\ \kappa\alpha\iota\ o\iota\ \delta\acute{\upsilon}o\ \varepsilon\acute{\iota}\nu\alpha\iota\ \varphi\acute{\iota}\lambda o\iota\ \tau o\upsilon\ k\ ) =$$
$$P(i, j, k\ \varepsilon\acute{\iota}\nu\alpha\iota\ \acute{o}\lambda o\iota\ \varphi\acute{\iota}\lambda o\iota\ )/P(i, j\ \varepsilon\acute{\iota}\nu\alpha\iota\ \varphi\acute{\iota}\lambda o\iota\ \tau o\upsilon\ k)$$
$$= p^3/p^2 = p$$





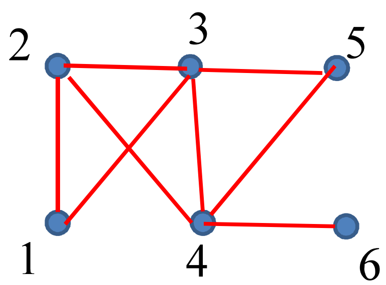
Συντελεστής μεταβατικότητας (transitivity ratio)

Ο λόγος μεταβατικότητας (transitivity ratio) ενός γραφήματος G ορίζεται ως ο λόγος:

$$T(G) = \frac{\pi\lambda\acute{\eta}\theta o\varsigma\ \mu\varepsilon\tau\alpha\beta\alpha\tau\iota\kappa\acute{\omega}\nu\ \tau\rho\iota\acute{\alpha}\delta\omega\nu}{\pi\lambda\acute{\eta}\theta o\varsigma\ \delta\upsilon\nu\alpha\tau\acute{\omega}\nu\ \sigma\upsilon\nu\delta\varepsilon\tau\iota\kappa\acute{\omega}\nu\ \tau\rho\iota\acute{\alpha}\delta\omega\nu}$$
$$= \frac{3\ \cdot \pi\lambda\acute{\eta}\theta o\varsigma\ \tau\rho\iota\gamma\acute{\omega}\nu\omega\nu\ \tau o\upsilon\ G}{\pi\lambda\acute{\eta}\theta o\varsigma\ \delta\upsilon\nu\alpha\tau\acute{\omega}\nu\ \sigma\upsilon\nu\delta\varepsilon\tau\iota\kappa\acute{\omega}\nu\ \tau\rho\iota\acute{\alpha}\delta\omega\nu\ \tau o\upsilon\ G}$$
$$T(G) = \frac{3 \cdot \tau(G)}{\rho(G)} = \frac{\sum_{v \in V} \tau(v)}{\sum_{v \in V} \rho(v)}$$

Ο λόγος μεταβατικότητας (Transitivity ratio) δεν ταυτίζεται με το συντελεστή σύμπλεξης όπως εύκολα προκύπτει από παραδείγματα

Παράδειγμα



v	τ(v)	ρ(v)	C(v)
1	1	1	1
2	2	3	2/3
3	3	6	1/2
4	2	6	1/3
5	1	1	1
6	0	0	0

Ο συντελεστής σύμπλεξης κάθε κορυφής δίνεται στον πίνακα (τελευταία στήλη).
Ο συντελεστής σύμπλεξης του γραφήματος είναι:

$$C(\mathrm{G}) = \frac{1}{n} \cdot \sum_{v \in V} C(v) = \frac{1 + \frac{2}{3} + \frac{1}{2} + \frac{1}{3} + 1 + 0}{6} = \frac{7}{12}$$

Ο λόγος μεταβατικότητας του γραφήματος είναι:

$$T(G) = \frac{\sum_{v \in V} \tau(v)}{\sum_{v \in V} \rho(v)} = \frac{1+2+3+2+1+0}{1+3+6+6+1+0} = \frac{9}{17}$$

Centrality Measures
Μέτρα κεντρικότητας

Ποιες είναι οι «σημαντικότερες» κορυφές;
Ως προς ποια κριτήρια;
- Πλήθος Συνδέσεων
- Σημαντικότητα συνδέσεων
- Μικρότερες αποστάσεις
- Κομβική σπουδαιότητα

Έχουν οριστεί διάφορα μέτρα κεντρικότητας (centrality measures)

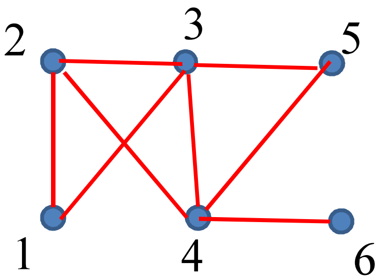
Βαθμική Κεντρικότητα (Degree Centrality)

- Ο βαθμός κορυφής $\deg(v)$, δηλ. το πλήθος ακμών που συνδέονται με την κορυφή, είναι ένας δείκτης κεντρικότητας.
- Η Βαθμική Κεντρικότητα ορίζεται ως το πηλίκο

$$\frac{\deg(v)}{n-1}$$

- Αντικαθιστώντας το βαθμό με τον έσω- (ή έξω-) βαθμό (σε κατευθυνόμενα γραφήματα) έχουμε ανάλογα εσω-(έξω-) βαθμική κεντρικότητα.
- Σε κοινωνικά δίκτυα σημαίνει όσο περισσότερες φιλίες (γνωριμίες, σχέσεις) έχει ένα άτομο (κορυφή) τόσο σημαντικότερο είναι στο δίκτυο.

Παράδειγμα



Βαθμοί κορυφών

1	2	3	4	5	6
2	3	4	4	2	1

Βαθμική κεντρικότητα

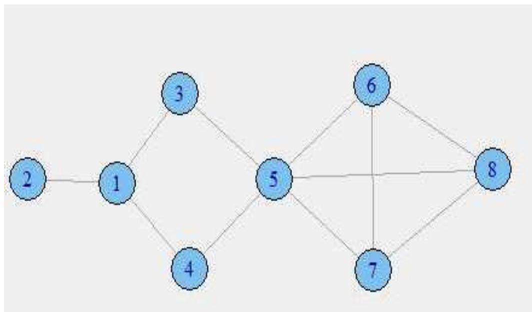
1	2	3	4	5	6
$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

Βαθμική κεντρικότητα (κανονικοποιημένη)

1	2	3	4	5	6
.50	.75	1	1	.50	.25

Ιδιοκεντρικότητα (Eigenvector Centrality)

Είναι φυσική γενίκευση της βαθμικής κεντρικότητας.



Στο Σχήμα.
Οι κορυφές 1, 6 έχουν βαθμό 3.

Όμως οι γείτονες της 1 έχουν συνολικό άθροισμα βαθμών 5,
ενώ οι γείτονες της 6 έχουν συνολικό άθροισμα βαθμών 11.

Αυτό μπορεί να ερμηνευτεί ότι η κορυφή 6 έχει μεγαλύτερη σημαντικότητα από την κορυφή 1 (έχει βαθμικά σημαντικότερους γείτονες).

Ιδιοκεντρικότητα

Έστω ο πίνακας αντιστοίχισης του G

$$A = (A_{i,j}),\text{ όπου } \mathrm{A_{i,j}} = \begin{cases} 1, & (i\,,j) \in \mathrm{E(G)} \\ 0, & (i\,,j) \notin \mathrm{E(G)} \end{cases}$$

Αν $\mathbf{x_0} = \mathbf{1} = (1,1,\dots 1,)'$ τότε

$$\mathbf{x_1} = \mathrm{A} \cdot \mathbf{x_0} \qquad \text{(οι βαθμοί κορυφών)}$$

Διαπιστώνουμε

$$\mathbf{x_2} = A \cdot \mathbf{x_1} = A^2 \cdot \mathbf{x_0} \quad \text{(είναι το άθροισμα βαθμών άμεσων γειτόνων)}$$

Συνεχίζοντας βρίσκουμε

$$\mathbf{x}_{\boldsymbol{n}} = A^n \cdot \mathbf{x_0} \qquad \text{(αθροίσματα n-τάξης)}$$

Ιδιοκεντρικότητα

- Αποδεικνύεται ότι οριακά ισχύει

$$\mathbf{x_n} \to k\, \boldsymbol{v_1},$$

  όπου $\boldsymbol{v_1}$ το ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή του πίνακα Α.

- Η κανονικοποιημένη μορφή (normalize) του ιδιοδιανύσματος $\boldsymbol{v_1}$ του πίνακα Α, δηλαδή αυτή που προκύπτει με διαίρεση των συνιστωσών του με το μέτρο του $|\boldsymbol{v_1}|$, είναι η ιδιοκεντρικότητα των κορυφών του γραφήματος που παριστάνει το Α.

Ύπαρξη Ιδιοκεντρικότητας

- Η ύπαρξή του διανύσματος $\boldsymbol{v_1}$ και το ότι είναι πραγματικό και θετικό εξασφαλίζεται από το θεώρημα Peron-Frobenius (P-F).
- Πράγματι σε συνδετικά απλά γραφήματα ο πίνακας αντιστοίχισης είναι συμμετρικός και πρωταρχικός (δηλ. υπάρχει δύναμη του A, έστω η A^k, με όλα τα στοιχεία διάφορα του 0), προϋποθέσεις για την ισχύ του θεωρήματος P-F.
- Στα κατευθυνόμενα γραφήματα:
  - δεν ορίζεται μονοσήμαντα το ιδιοδιάνυσμα της μέγιστης ιδιοτιμής.
  - ο πίνακας A είναι μη-συμμετρικός.

Έτσι η ιδιοκεντρικότητα δεν λειτουργεί καλά.

Η ιδιοκεντρικότητα ορίστηκε αρχικά από τον Bonacich το 1987.

Κεντρικότητα Katz

- Είναι γενίκευση της ιδιοκεντρικότητας. Αν θεωρήσουμε την σχέση

$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

για θετικές ποσότητες α, β, τότε είναι σαν να δίνουμε μία «ποσότητα κεντρικότητας» σε όλους τους κόμβους (να μην είναι 0).

- Προτάθηκε το 1957 από τον Katz και υπολογίζεται από τη σχέση:

$$x = \beta (I - \alpha A)^{-1} \mathbf{1}$$

Ως β μπορούμε να πάρουμε το 1, ενώ ως α μία τιμή μικρότερη του $\frac{1}{\lambda_1}$ (στην οποία μηδενίζεται το διάνυσμα).

Βαθμική κεντρικότητα Page (PageRank centrality)

Η κατά Katz κεντρικότητα έχει το εξής μειονέκτημα: αν ένα κόμβος έχει μεγάλη κεντρικότητα, τότε και όσοι συνδέονται με αυτόν έχουν επίσης μεγάλη κεντρικότητα. Για παράδειγμα κάποιος κόμβος που συνδέεται με τη Google, ή την Yahoo που έχουν μεγάλη κεντρικότητα θα έχει και αυτός μεγάλη κεντρικότητα.

Οι Larry Page και Sergey Brin που δημιούργησαν την Google ανέπτυξαν έναν νέο αλγόριθμο για αξιολόγηση της σημαντικότητας των σελίδων που συνδέονται με την Google, που ήταν βελτίωση της κατά Katz κεντρικότητας, που τον ονόμασαν PageRank centrality (Βαθμική κεντρικότητα Page).

Αυτό που διορθώθηκε είναι ότι έγινε διαίρεση με τον εξω-βαθμό, δηλαδή:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

Κεντρικότητα Εγγύτητας (closeness centrality)

- Αν $d_{ij}, j = 1, 2, \dots, d_i$ οι αποστάσεις της κορυφής i από τις άλλες κορυφές του γραφήματος, τότε:

$$l_i = \frac{1}{n-1} \sum_{j \neq i} d_{ij}$$

η μέση απόσταση, είναι μικρή για τις κεντρικές κορυφές και μεγαλύτερη για τις απομακρυσμένες κορυφές.

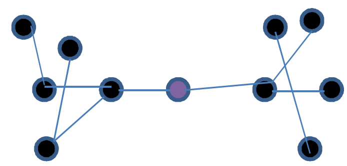
Αν δεν υπάρχει μονοπάτι μεταξύ δύο κορυφών θέτουμε ως απόσταση το πλήθος όλων των κορυφών (αυθαίρετο).

Η αντίστροφη ποσότητα $C_i = \frac{1}{l_i}$, είναι η κεντρικότητα εγγύτητας.

Κεντρικότητα Εγγύτητας **(συν.)**

- Η κεντρικότητα αυτή χρησιμοποιείται πολύ σε κοινωνικά δίκτυα, έχει όμως κάποια μειονεκτήματα.
  - Το εύρος των τιμών είναι πολύ μικρό και έτσι δεν μπορεί να γίνει καλός διαχωρισμός και ταξινόμηση των κορυφών ιδιαίτερα σε μεγάλα δίκτυα.
  - Μεταβάλλεται εύκολα με την προσθήκη ή διαγραφή κορυφών.
- Έχουν προταθεί και άλλοι ορισμοί, όπως το να χρησιμοποιηθεί ο αρμονικός μέσος των αποστάσεων ο οποίος έχει ωραίες ιδιότητες και δεν έχει πρόβλημα με κορυφές που δεν συνδέονται, αφού το αντίστροφό τους είναι 0 και δεν μεταβάλλει την κεντρικότητα.

Ενδιάμεση Κεντρικότητα (betweenness centrality)



Η ενδιάμεση κορυφή έχει μικρή βαθμική κεντρικότητα αλλά μεγάλη ενδιάμεση κεντρικότητα

- Η ενδιάμεση κεντρικότητα μετρά το κατά πόσον μια κορυφή βρίσκεται σε μονοπάτια (γεωδαισιακές) μεταξύ άλλων κορυφών. Προτάθηκε από τον Freeman το 1977, αλλά είχε ήδη προταθεί από τον Anthonisse σε αδημοσίευτη εργασία του.
- Αν n_{st}^{i} =1 όταν η κορυφή i βρίσκεται στη γεωδαισιακή που συνδέει τις κορυφές s, t και 0 σε άλλη περίπτωση, τότε η ενδιάμεση κεντρικότητα ορίζεται ως

$$x = (x_i)\text{ , όπου } x_i = \sum_{s,t} n_{s,t}^{i}$$

Αν n_{st}^{i} το πλήθος των γεωδαισιακών από το s προς το t που περνούν από το i και g_{st} το πλήθος των γεωδαισιακών από το s προς το t, τότε η ενδιάμεση κεντρικότητα της κορυφής i γράφεται: $x_i = \sum_{s,t} \frac{n_{s,t}^{i}}{g_{st}}$

Το πείραμα του Stanley Milgram (1967)

Ο ψυχολόγος Stanley Milgram θέλοντας να αποδείξει ότι οι άνθρωποι αλληλοσυνδέονται πολύ περισσότερο από όσο νομίζουμε, διάλεξε τυχαία 96 άτομα από τον τηλεφωνικό κατάλογο και τους έστειλα από ένα δέμα που είχε το λογότυπο του Harvard (όπου δίδασκε ο Milgram) που έπρεπε να το παραδώσουν σε κάποιον (που ήταν φίλος του Milgram) και για τον οποίο ξέρανε μόνο το όνομά του, τη διεύθυνσή του και το ότι ήταν χρηματιστής. Κάθε παραλήπτης διάβαζε στο δέμα ότι θα έπρεπε να στείλει το δέμα σε κάποιον γνωστό του που κατά τη γνώμη του να ήταν κοινωνικά πλησιέστερα στον τελικό παραλήπτη.

Από τα 96 δέματα τα 18 πήγαν στον τελικό παραλήπτη, κάτι που ήταν περισσότερο από το αναμενόμενο. Σε μια μοντέρνα επανάληψη του πειράματος που έκανε ο Dodds το 2003, με αποστολή e-mails σε γνωστούς, η επιτυχία ήταν πολύ μικρότερη. Ο Milgram είχε ζητήσει από τους παραλήπτες των δεμάτων να σημειώνουν επάνω στο δέμα τα άτομα που συμμετείχαν στην αναζήτηση του τελικού παραλήπτη. Ο μέσος όρος των βημάτων που χρειάστηκαν ήταν 5.9

Real-World Networks - Small World Model

While random graphs exhibit some futures of real-world networks (small diameters or average distances **relative** to growing average degree), they lack other characteristics.

Stanley Milgram (1967), Harvard, what is the probability that two randomly selected people would know each other?

On average **5.5** hops (*in a sparse network*)

Six degrees of separation

Why?

In social networks (*which are sparse networks*) it has observed that the friends of us are usually also friends between them, or in other words people tend to group into relative small (or large clusters) in a way that in some cases does not depend to the place of living.

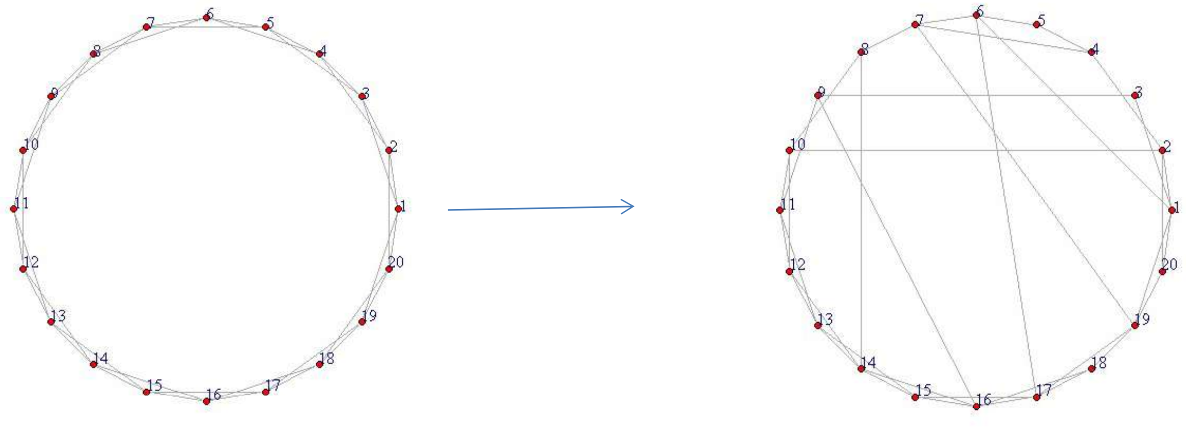
Small World Model

- Ένα δίκτυο λέγεται μικρόκοσμος όταν η μέγιστη απόσταση μεταξύ των κορυφών του είναι σχετικά «μικρή». Αν δηλαδή το πλήθος των βημάτων ή “hops” μεταξύ τυχαία επιλεγμένων κόμβων είναι κατά κάποια έννοια μικρό. Ειδικότερα θα λέμε ότι ένα δίκτυο έχει την ιδιότητα του μικρόκοσμου (the small-world property) αν η μέση γεωδαισιακή απόσταση ℓ μεταξύ των κόμβων του δικτύου είναι μικρή σε σχέση με το πλήθος N των κόμβων του δικτύου. Αυτή η σχέση συνήθως λαμβάνεται με το λογάριθμο του N. Δηλαδή θέλουμε η απόσταση ℓ να μην αυξάνεται γρηγορότερα από το logN, ή όπως γράφουμε με το κεφαλαίο «O» να ισχύει ℓ=O(logN) καθώς N→∞. Η βάση των λογαρίθμων δεν ενδιαφέρει.

Πιθανότητα επανασύνδεσης (rewiring probability)

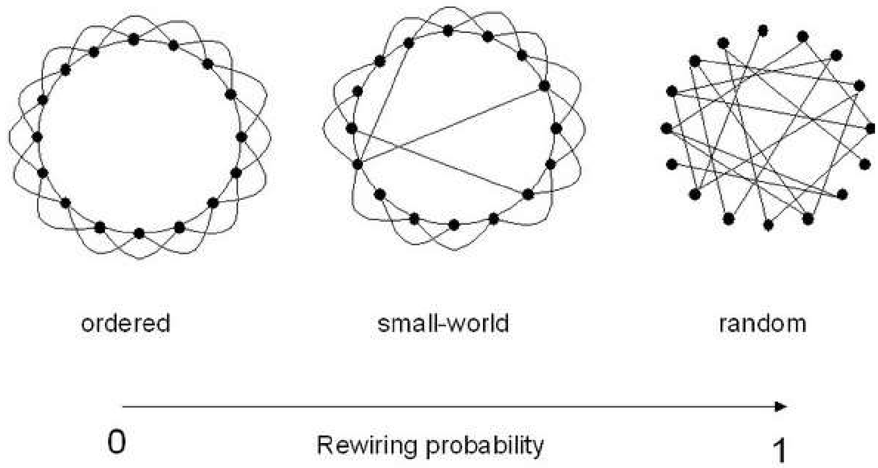
Αν έχουμε ένα κανονικό γράφημα με βαθμό κορυφών 2k, τότε μπορούμε να τοποθετήσουμε τις κορυφές του σε κυκλικό σχηματισμό, με τρόπο ώστε κάθε κορυφή να συνδέεται με τις k κοντινότερες κορυφές από δεξιά του και τις k κοντινότερες κορυφές από αριστερά του.

Π.χ. Για $n = 20, k = 2$, έχουμε το παρακάτω σχήμα αριστερά. Έστω πιθανότητα p, π.χ. $p = 0.2$, με την οποία επιλέγουμε κάθε μία από τις υπάρχουσες, ακμές και τις επανα-συνδέουμε (rewire) με πιθανότητα p, ή τις αφήνουμε όπως είναι με πιθανότητα $1 - p$. Στο παράδειγμα Υπάρχουν $\frac{20 \cdot 4}{2} = 40$ ακμές και άρα αναμένεται $40 \cdot p = 40 \cdot 0.2 = 8$ να επανασυνδεθούν (σχήμα δεξιά). Η διάμετρος από 5 που ήταν αρχικά, έγινε 4 (π.χ. d(1,13)=4), ενώ θεωρητικά αναμένονταν να είναι κοντά το 3, διότι $log20 = 2.996 = 3$.



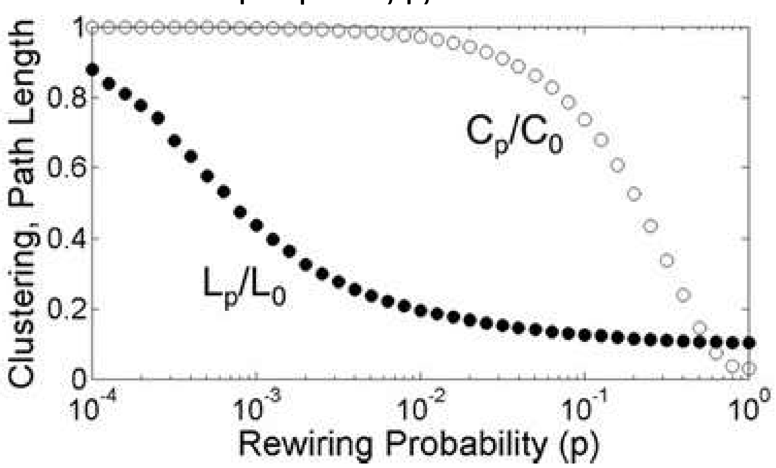
Watts and Strogatz - Small World Model

Η ιδέα του rewiring ξεκίνησε σε μια εργασία των **Watts and Strogatz (1998).** Ξεκινώντας από ένα 2k-κανονικό γράφημα και αυξάνοντας την πιθανότητα επανασύνδεσης από 0 μέχρι 1 μελέτησαν τα προκύπτοντα γραφήματα. Με p=0 έχουμε κανονικά γραφήματα, με p=1 έχουμε τυχαία δίκτυα ενώ με τιμές ενδιάμεσες παίρνουμε μικρόκοσμους. Αυτό φαίνεται στην παρακάτω εικόνα από την εργασία των Watts and Strogatz.



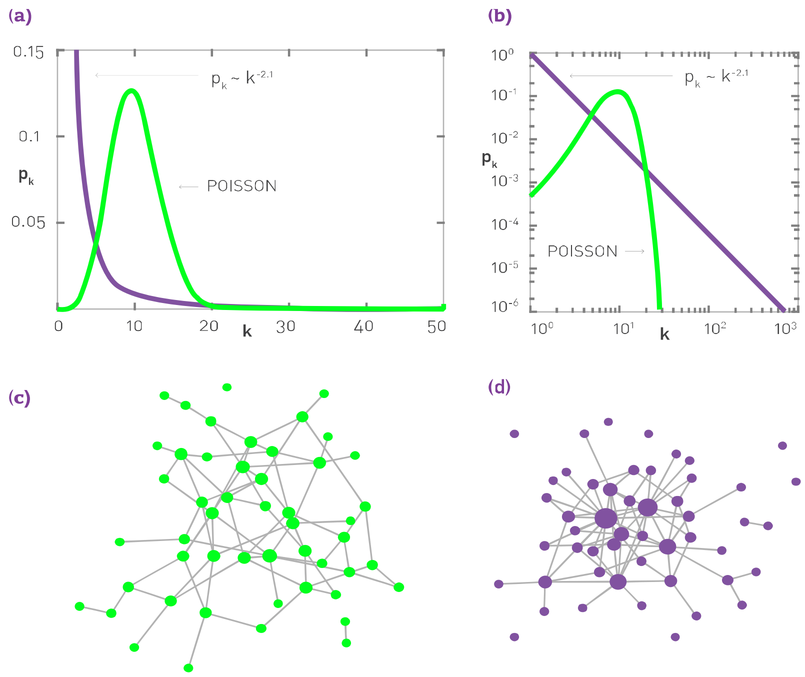
Watts and Strogatz - Small World Model

Από την ίδια εργασία των Watts and Strogatz διαπιστώνεται ότι όσο αυξάνει η Rewiring πιθανότητα τόσο μειώνεται το μέσο μήκος μονοπατιού, αρχικά απότομα και μετά αργά. Αντίθετα ο συντελεστής σύμπλεξης μειώνεται αρχικά αργά και για μεγάλα p απότομα. Αυτά φαίνονται στο παρακάτω σχήμα. Έτσι σε ένα μικρόκοσμο (π.χ. p=0.01) βλέπουμε ότι έχουμε σχεδόν ελάχιστη τιμή στο μέσο μήκος μονοπατιού, ενώ έχουμε σχετικά μεγάλο συντελεστή σύμπλεξης



Αυτοόμοια (scale-free) δίκτυα

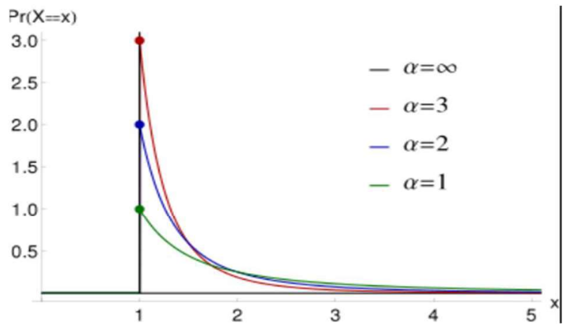
Έτσι ονομάζονται τα δίκτυα που η κατανομή των βαθμών τους είναι δυναμοκατανομή (power-low) αντί να είναι Poisson όπως είναι στα τυχαία δίκτυα Erdos-Renyi. Συμβαίνει στα δίκτυα αυτά η μορφή τους να είναι παρόμοια αν περιοριστούμε σε μέρη τους. Στο σχήμα το πράσινο δίκτυο είναι Poisson, το μώβ είναι αυτοόμοιο



Κατανομή **Pareto - Δυναμοκατανομή**

Ορισμός:

$$f(x) = \begin{cases} \dfrac{a\xi^a}{x^{a+1}}, x \geq \xi \\ \quad 0 \quad , x < \xi \end{cases}$$



$$1 - \mathrm{F}(x) = P(X > x) = \begin{cases} \left(\dfrac{\xi}{x}\right)^a, x \geq \xi \\ \quad 1 \quad , x < \xi \end{cases}$$

Ιδιότητες

Η αρχή 80-20 του Pareto. Το 80% του πλούτου κατέχετε από το 20% του πληθυσμού.

Οι περισσότερες πωλήσεις γίνονται προς λίγους πελάτες.

Οι περισσότερες πωλήσεις γίνονται από λίγους πωλητές.

Τα μεγαλύτερα προβλήματα δημιουργούνται απί λίγα άτομα, κλπ

Εμπειρικά παραδείγματα

- Η κατανομή μιας μεγάλης ποικιλίας φυσικών, βιολογικών, και ανθρωπογενών φαινομένων ακολουθούν προσεγγιστικά κάποια κατανομή νόμου δύναμης. Για παράδειγμα:
  - Τα μεγέθη των σεισμών, των κρατήρων στη Σελήνη και των ηλιακών εκλάμψεων,
  - η αναζήτηση τροφής διαφόρων ειδών,
  - τα μεγέθη των μοντέλων για νευρωνικούς πληθυσμούς,
  - η συχνότητα εμφάνισης των λέξεων στις περισσότερες γλώσσες,
  - η συχνότητα εμφάνισης των επωνύμων,
  - το πλήθος των ειδών σε διάφορες κατηγορίες οργανισμών,
  - τα μεγέθη της διακοπής ρεύματος,
  - η συχνότητα πολέμων, ποινικών διώξεων ανά κατάδικο

  και πολλές άλλες ποσότητες. Ελάχιστες κατανομές ταιριάζουν με κάποια κατανομή νόμου δύναμης σε όλες τις τιμές τους, αλλά οι περισσότερες ακολουθούν κάποια κατανομή νόμου δύναμης στην ουρά τους.
  - Η συχνότητα της ακουστικής απόσβεσης ακολουθεί κατανομή νόμου δύναμης για μεγάλης συχνότητας μπάντες για πολλά σύνθετα μέσα.
  - Οι νόμοι αλλομετρικής κλίμακας για σχέσεις μεταξύ βιολογικών μεταβλητών είναι ανάμεσα στις πιο γνωστές συναρτήσεις νόμου δύναμης στη φύση.

Παράδειγμα : World Wide Web

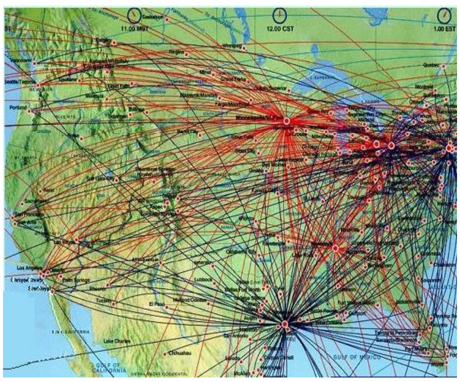
Nodes: WWW documents Links: URL links

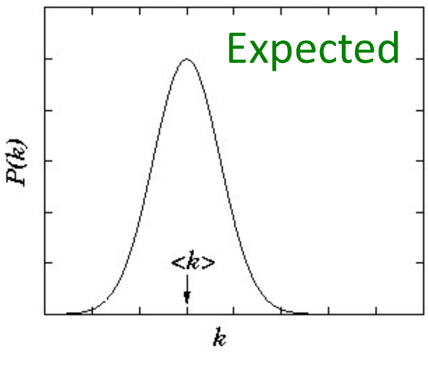
ROBOT: collects all URL's found in a document and follows them recursively. Over 3 billion documents

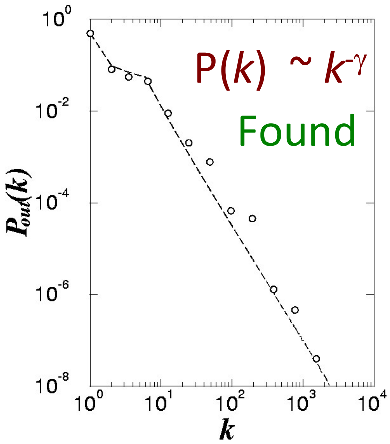
R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

Exponential Network

Scale-free Network

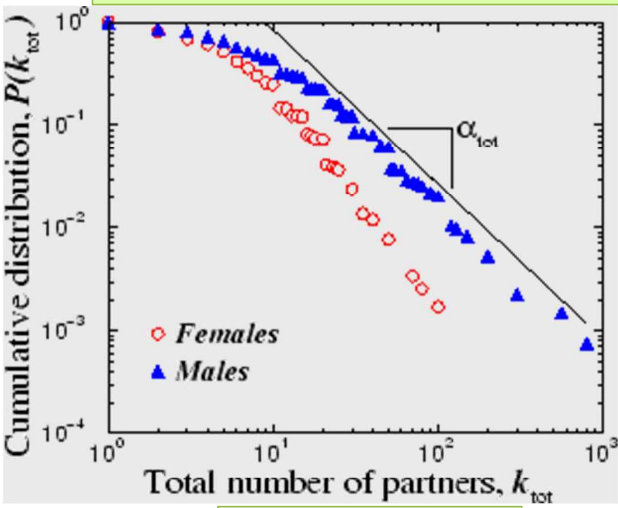






Παράδειγμα: Σεξουαλικές σχέσεις

Nodes: people (Females; Males)
Links: sexual relationships



4781 Swedes; 18-74;
59% response rate.

Liljeros et al. *Nature* 2001

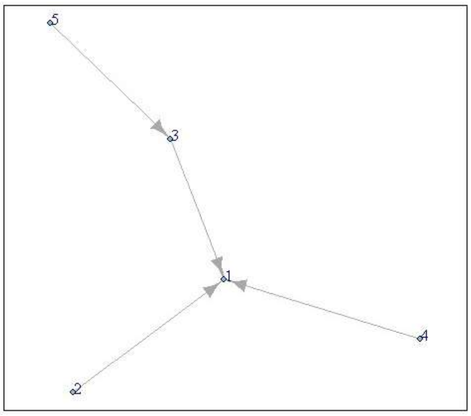
Το μοντέλο BA (Barabasi-Albert)

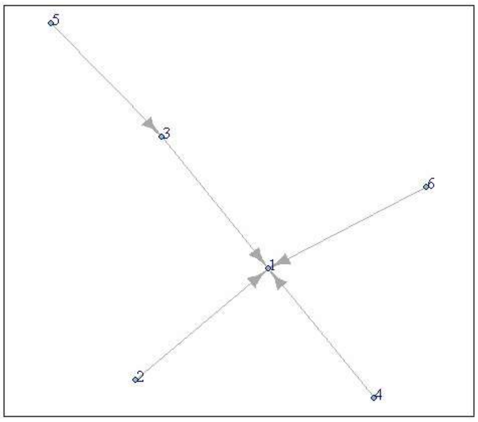
- Έστω αρχικά ένα δίκτυο με m_0 κορυφές. Αυξάνουμε τις κορυφές του δικτύου, ανά μία κάθε στιγμή ($t = l, t = 2, \dots$) και τις ακμές του με τον εξής αλγόριθμο. Κάθε νέα κορυφή συνδέεται με $m \leq m_0$ ακμές με τις προϋπάρχουσες κορυφές ανάλογα με τον έως τότε βαθμό των κορυφών. Έτσι η πιθανότητα p_j η νέα κορυφή που προστέθηκε σε κάποιο βήμα να συνδεθεί με την κορυφή i είναι:

$$p_j = \frac{k_i}{\sum_t k_t} = \frac{k_i}{2M} ,$$

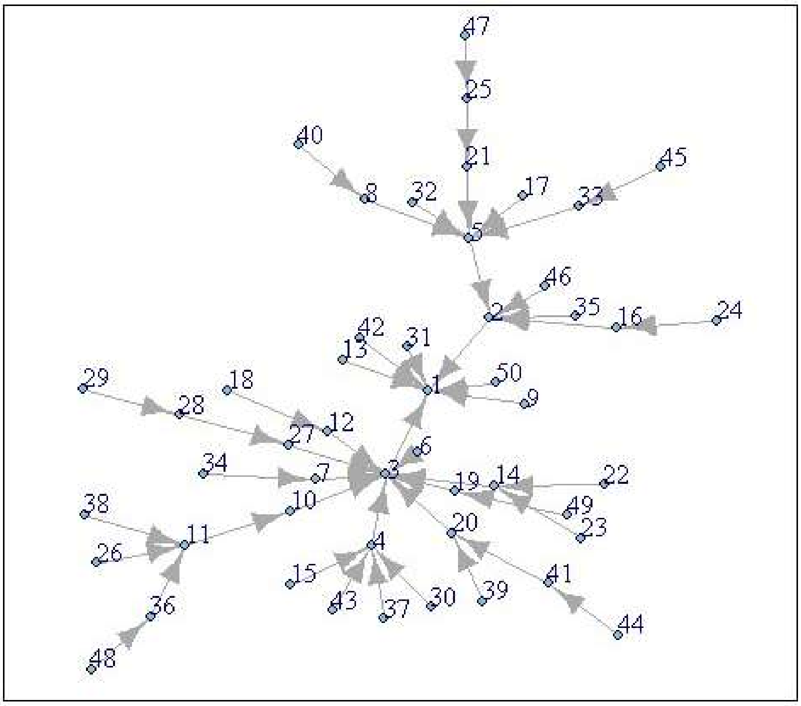
  όπου k_i είναι ο βαθμός της κορυφής i, το άθροισμα διατρέχει όλες τις προϋπάρχουσες κορυφές μέχρι τη στιγμή αυτή, και M το πλήθος των ακμών που είχε το δίκτυο πριν τη στιγμή αυτή. Ο τρόπος αυτός αρχίζει να ευνοεί κάποιους κόμβους οι οποίοι στη συνέχεια γίνονται όλο και περισσότερο πιθανό να συνδέονται με τους νέους κόμβους. Αυτοί οι κόμβοι λέγονται ομφαλοί (hubs).
- Η διαδικασία αυτή λέγεται **preferential attachement**. Θα μπορούσαμε εδώ να πούμε ότι στα δίκτυα αυτά οι «πλούσιοι γίνονται πλουσιότεροι»

Preferential Attachment BA model



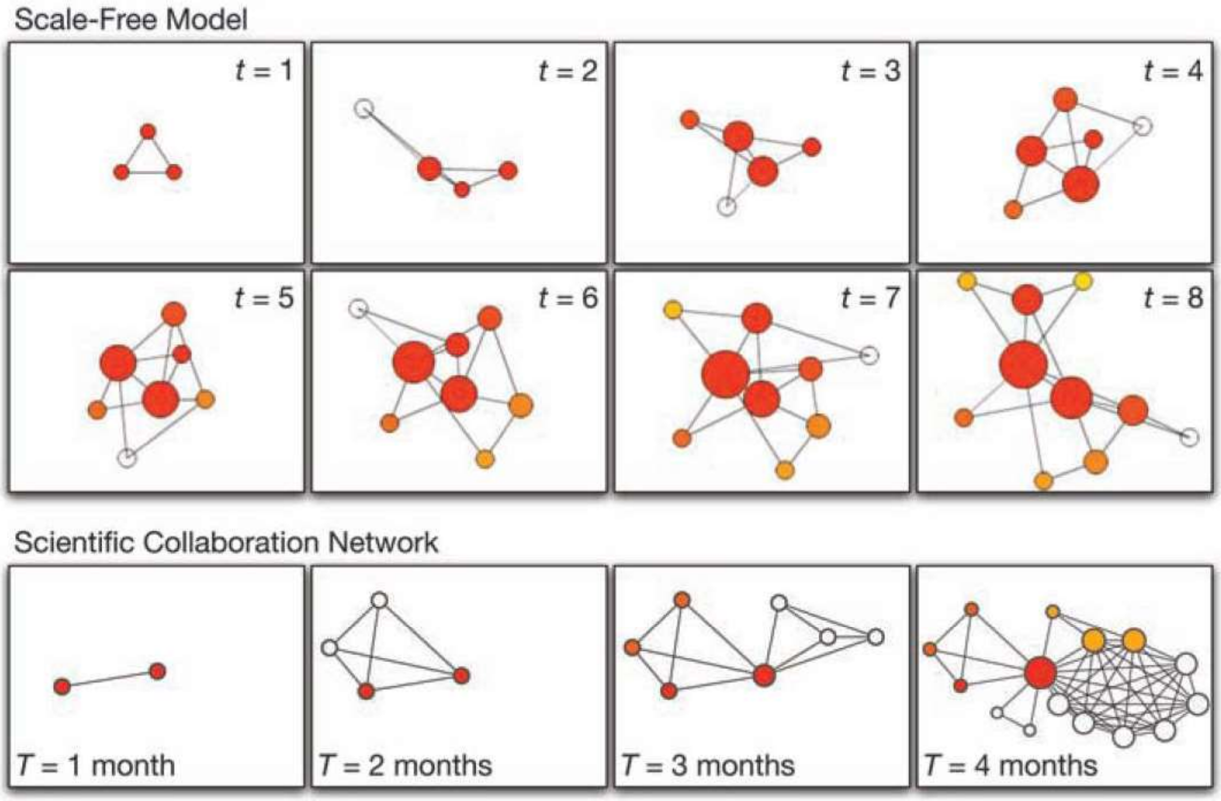


Preferential Attachment BA model



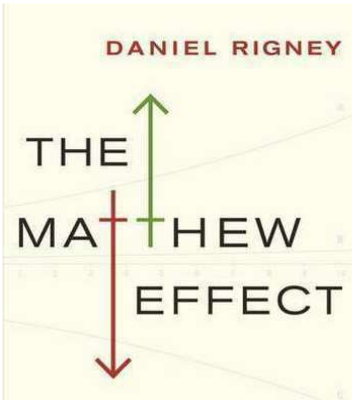
Δεν δημιουργούνται κύκλοι στο μοντέλο BA

The growth of BA network



Albert-László Barabási, 2008

The Matthew effect



Σε μια παραλλαγή, ιδιαίτερα σε οικονομικά δίκτυα, εκτός από το να αυξάνουν (πλουτίζουν) αυτοί που έχουν ήδη πολλά, αυτοί οι κόμβοι που έχουν λίγες συνδέσεις με κάποια πιθανότητα τις χάνουν (φτωχαίνουν).

Σε ένα τέτοιο μοντέλο θα λέγαμε ότι «οι πλούσιοι γίνονται πλουσιότεροι και οι φτωχοί φτωχότεροι». Αυτό το φαινόμενο που το παρατηρούμε και στην πραγματική ζωή λέγεται Matthew effect, από ένα απόσπασμα του «Κατά Ματθαίον» Ευαγγελίου, το 25:29, που λέει:

«τῷ γὰρ ἔχοντι παντὶ δοθήσεται καὶ περισσευθήσεται, ἀπὸ δὲ τοῦ μὴ ἔχοντος καὶ ὃ ἔχει ἀρθήσεται ἀπ᾽ αὐτοῦ.»

Ελεύθερη μετάφραση: «Στον καθένα που έχει θα δοθεί και θα περισσέψει, ενώ απ' αυτόν που δεν έχει και αυτό που έχει θα του αφαιρεθεί»

Ιδιότητες του μοντέλου BA

- Η κατανομή βαθμών είναι $P(k) \sim \mathrm{k}^{-3}$δηλαδή είναι «δυναμοκατανομή» (power low distribution) και το δίκτυο είναι ανεξάρτητο κλίμακος (scale free).
- Το «μέσο μήκος μονοπατιού» του μοντέλου BA είναι συστηματικά μικρότερο του αντίστοιχου μέσου μήκους μονοπατιού για τα τυχαία δίκτυα και εκφραζεται από τη σχέση $l = \frac{lnN}{\ln(lnN)}$.
- Έχει αποδειχθεί (Klemm and Eguiluz, Bollobas) ότι ο συντελεστής σύμπλεξης ακολουθεί ένα νόμο γινομένου $C \sim \mathrm{N}^{-0.75}$. Αυτό συνιστά σημαντική διαφορά από τους μικρόκοσμους, όπου ο συντελεστής σύμπλεξης δεν έχει σχέση με το μέγεθος του δικτύου, αλλά ακολουθεί ένα νόμο γινομένου ως προς το βαθμό κάποιας κορυφής $C(k) \sim k^{-1}$.

Δίκτυο συνεργασιών του ΕΣΙ