introduction to *network analysis* (*ina*)

Lovro Šubelj

University of Ljubljana
Faculty of Computer and Information Science
spring 2019/20

announcements *3rd week*

- *homework #1* out *next Monday*
- *homework #1* due in *three weeks*
- *course project* details in *two weeks*

- *four knights* challenge *review*
- again *Kahoot! quiz* in *lectures*
- *homework #0* review in *labs*
- bring *your laptops* to *labs*

- *feedback box* from *last week*
- posts to *feedback box* get *candy* =)

Erdős-Rényi *random graph*

introduction to *network analysis* (*ina*)

Lovro Šubelj
University of Ljubljana
spring 2019/20

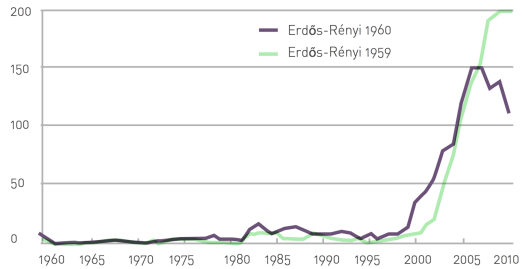
graph *models*

- *graph model* is *ensemble* of random graphs
- *algorithm* for random graphs of given parameters
  - *baseline* for *network structure* statistics
  - for *reasoning* about *network evolution*
  - for *generating* large *random graphs*
- *random graph* refers to *Erdős-Rényi model* [ER59]

assume *undirected* G from now on

Pál Erdős

Alfréd Rényi



Erdős-Rényi model

graph $G(n, m)$ *model*

— $G(n, m)$ *random graph* model [ER59]

— *randomly place* m *links* between $\binom{n}{2}$ node pairs

— *computationally convenient* but *analytically hard*

n, m given $\quad\quad \langle k \rangle = 2m/n$

input parameters n, m
output graph G

1: $G \leftarrow n$ isolated nodes
2: **while not** G has m links **do**
3: add link for random node pair
4: **end while**
5: **return** G

graph $G(n, p)$ *model*

— $G(n, p)$ *random graph* model [SR51]

— *place links* between $\binom{n}{2}$ node pairs *with probability* p

— *computationally hard* but *analytically convenient*

n, p given $\quad$ $m, \langle k \rangle$ *unknown*

input parameters n, p
output graph G

1: $G \leftarrow n$ isolated nodes
2: **for all** $\binom{n}{2}$ node pairs in G **do**
3: add link with probability p
4: **end for**
5: **return** G

graph *density & degree*

— *number of links* m follows *binomial distribution* $\mathrm{B}(\binom{n}{2}, p)$

$$x \sim \mathrm{B}(n,p) \text{ then } p_x = \binom{n}{x} p^x (1-p)^{n-x} \text{ and } \langle x \rangle = np$$

$$\langle m \rangle = \sum_{m=0}^{\binom{n}{2}} m\mathrm{P}(m) = \sum_{m=0}^{\binom{n}{2}} m \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m} = \binom{n}{2} p$$

— then *density* $\rho = p$ and *average degree* $\langle k \rangle = (n-1)p$

graph *degree distribution*

— *degree distribution* p_k is *binomial distribution* $\mathrm{B}(n-1,p)$

$$x \sim \mathrm{B}(n,p) \text{ then } p_x = \binom{n}{x} p^x (1-p)^{n-x} \text{ and } \langle x \rangle = np$$

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

— p_k approximately *Poisson distribution* $\mathrm{Pois}(\langle k \rangle)$ for $n \gg \langle k \rangle$

$$x \sim \mathrm{Pois}(\lambda) \text{ then } p_x = \frac{\lambda^x e^{-\lambda}}{x!} \text{ and } \langle x \rangle = \lambda$$

$$\ln\left[(1-p)^{n-1-k}\right] = (n-1-k)\ln\left(1-\tfrac{\langle k \rangle}{n-1}\right) \simeq -(n-1-k)\tfrac{\langle k \rangle}{n-1} \simeq -\langle k \rangle$$

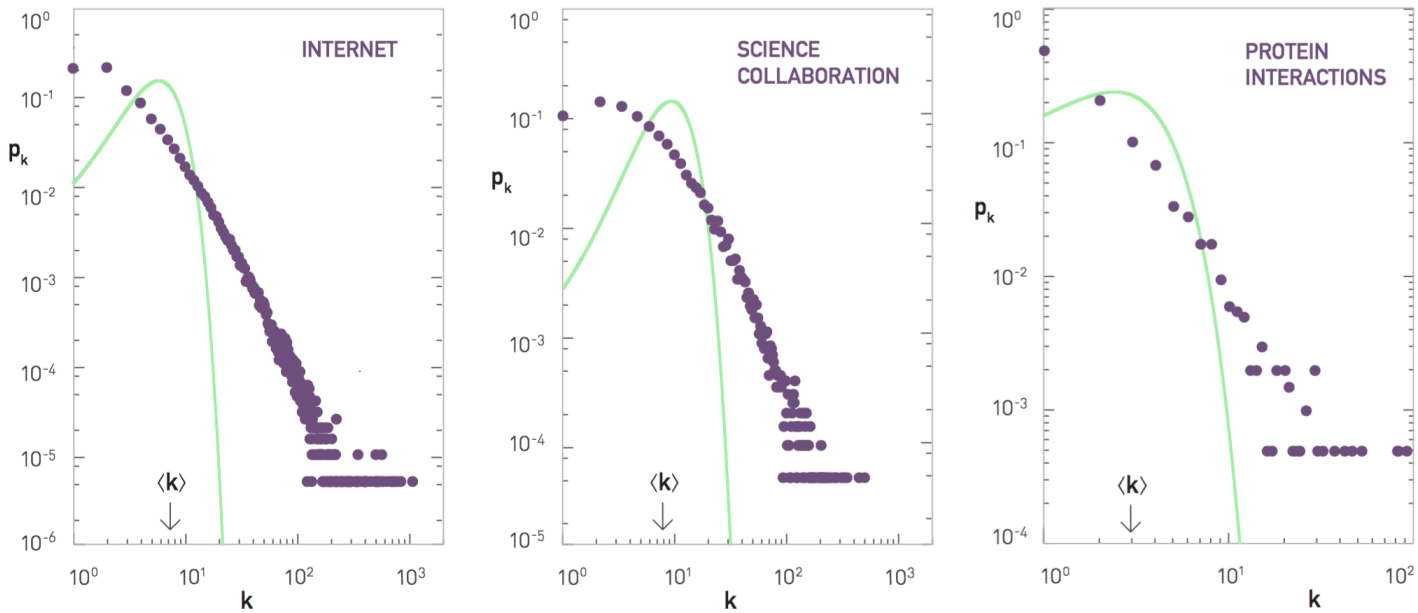
$$p_k \simeq \frac{(n-1)^k}{k!}\left(\frac{\langle k \rangle}{n-1}\right)^k e^{-\langle k \rangle} = \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!}$$

network *degree distribution*

— *scale-free* p_k of real networks [Bar16]

— real networks are *not Poisson graphs*

— random graphs *lack hubs* with $k \gg \langle k \rangle$

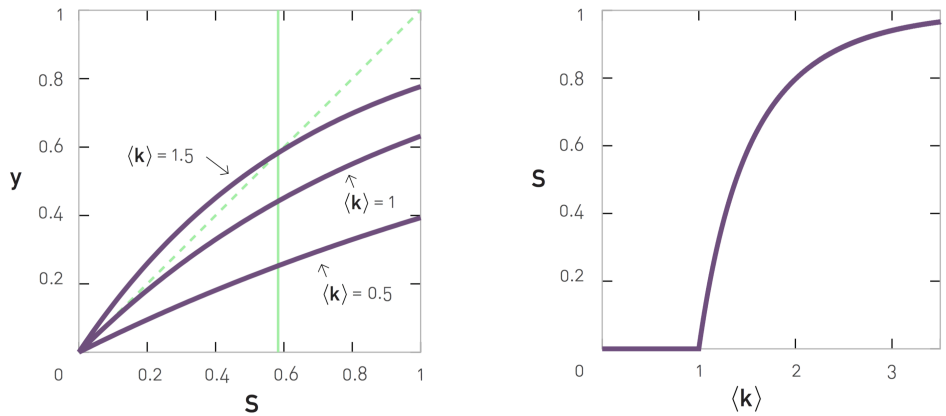


graph *connectivity*

— *fraction of nodes* in *giant component* S for $n \gg \langle k \rangle$

$$\ln(1-S) = (n-1)\ln(1-pS) \simeq -(n-1)pS = -(n-1)\frac{\langle k \rangle}{n-1}S = -\langle k \rangle S$$

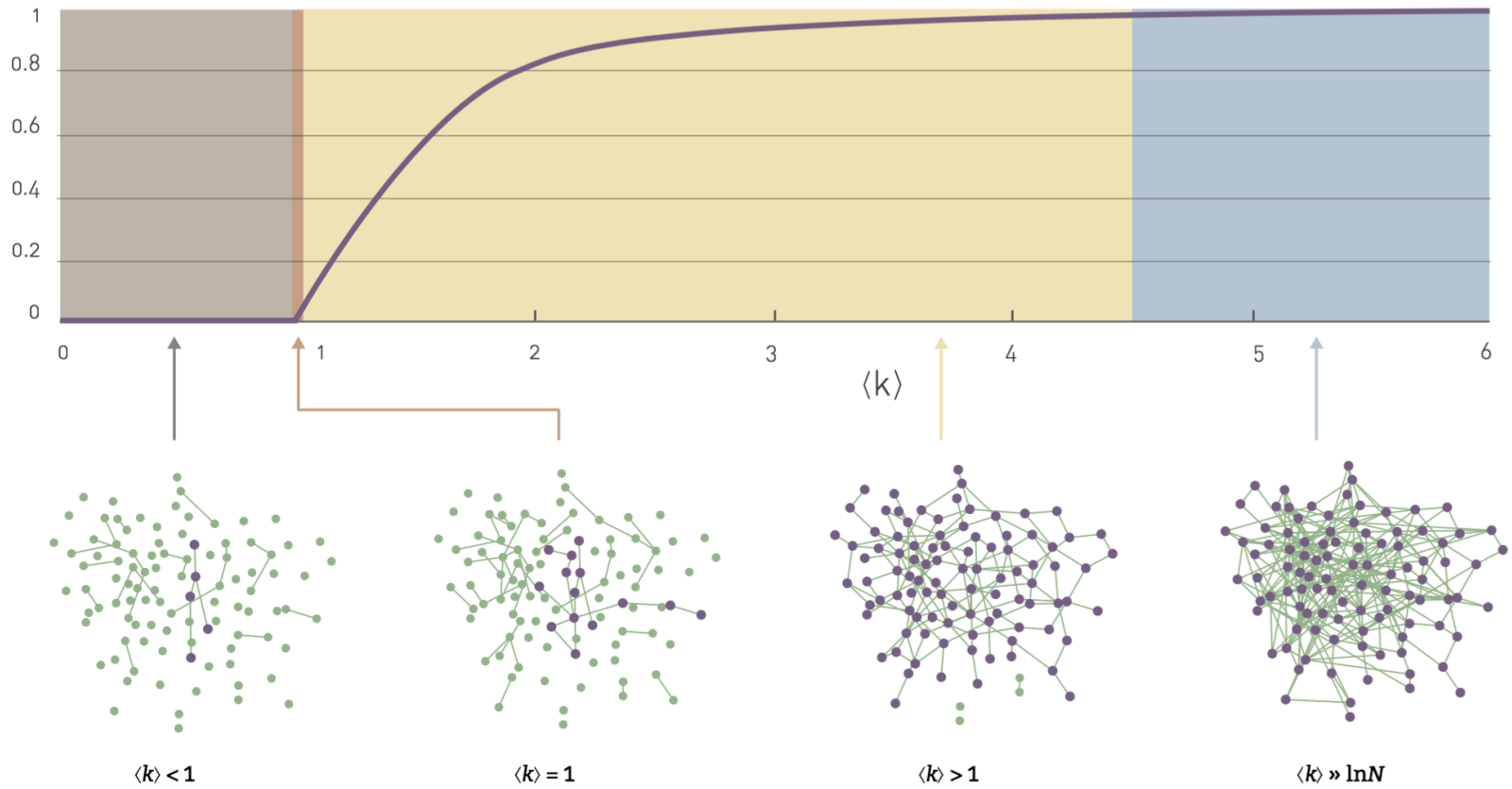
$$1 - S = (1 - p + p(1-S))^{n-1} \qquad S = 1 - e^{-\langle k \rangle S}$$



— *emergence* of *giant component* or *phase transition* at $\langle k \rangle = 1$

$$\frac{\mathrm{d}}{\mathrm{d}S}(1 - e^{-\langle k \rangle S})\Big|_{S=0} = \langle k \rangle e^{-\langle k \rangle S}\Big|_{S=0} = \langle k \rangle > 1$$

graph *evolution*



subcritical $n_S \sim \ln n$ *critical point* $n_S \sim n^{2/3}$ *supercritical* $n_S \sim n\frac{\langle k\rangle - 1}{n-1}$ *fully connected* $n_S \approx n$

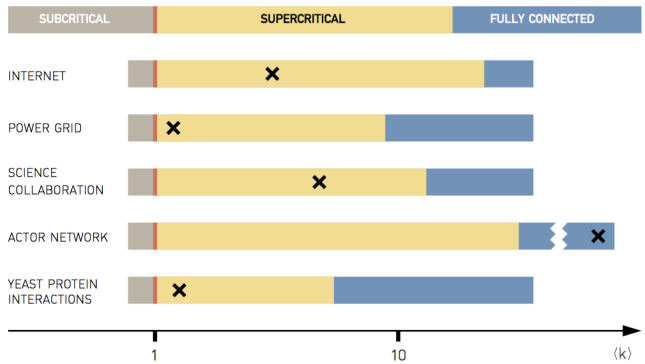
see random graph evolution NetLogo demo

network *connectivity*

— *connectivity* of real networks [Bar16]

— networks *supercritical* with $1 < \langle k \rangle < \ln n$

NETWORK	N	L	⟨k⟩	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,439	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61



— *Facebook* friendships [BBR+12] *connected* with $S > 0.997$

graph *diameter & distance*

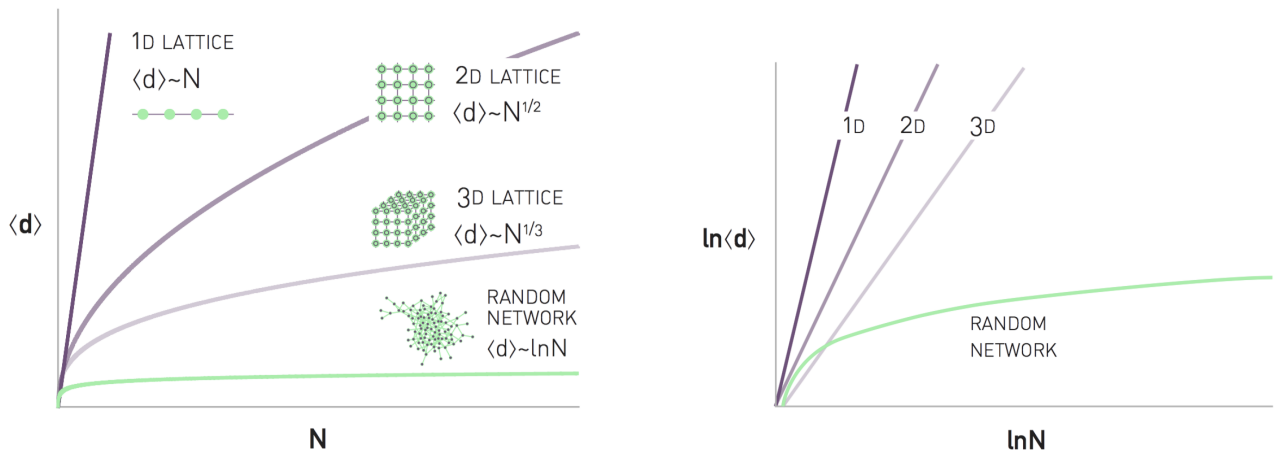
— *diameter* d_{max} and *average distance* $\langle d \rangle$ for $n \gg \langle k \rangle$

$$1 + \langle k \rangle + \langle k \rangle^2 + \cdots + \langle k \rangle^{d_{max}} = \frac{\langle k \rangle^{d_{max}+1} - 1}{\langle k \rangle - 1} \approx \langle k \rangle^{d_{max}} \simeq n$$

$$d_{max} \simeq \frac{\ln n}{\ln \langle k \rangle} \qquad \langle d \rangle \approx \frac{\ln n}{\ln \langle k \rangle}$$

— $\langle d \rangle = 4.74$ for *Facebook* [BBR+12] while $\frac{\ln n}{\ln \langle k \rangle} = 3.98$

— random graphs *small-world* opposed to regular *lattices*



network *diameter & distance*

— *diameter* d_{max} and *distance* $\langle d \rangle$ of real networks [Bar16]

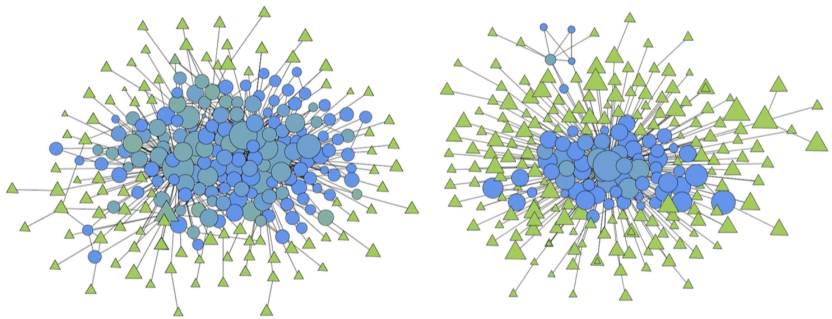
— $\langle d \rangle$ *well estimated* by $\frac{\ln n}{\ln \langle k \rangle}$ whereas $d_{max} \gg \frac{\ln n}{\ln \langle k \rangle}$

NETWORK	N	L	$\langle k \rangle$	$\langle d \rangle$	d_{max}	$\frac{\ln N}{\ln \langle k \rangle}$
Internet	192,244	609,066	6.34	6.98	26	6.58
WWW	325,729	1,497,134	4.60	11.27	93	8.31
Power Grid	4,941	6,594	2.67	18.99	46	8.66
Mobile Phone Calls	36,595	91,826	2.51	11.72	39	11.42
Email	57,194	103,731	1.81	5.88	18	18.4
Science Collaboration	23,133	93,439	8.08	5.35	15	4.81
Actor Network	702,388	29,397,908	83,71	3,91	14	3,04
Citation Network	449,673	4,707,958	10.43	11,21	42	5.55
E. Coli Metabolism	1,039	5,802	5.58	2.98	8	4.04
Protein Interactions	2,018	2,930	2.90	5.61	14	7.14

graph *clustering*

— *clustering coefficients* $\langle C \rangle$ [WS98] and C [NSW01]

$$C = \langle C \rangle = \langle C_i \rangle = \frac{2\langle t_i \rangle}{k_i(k_i-1)} = \frac{2p\binom{k_i}{2}}{k_i(k_i-1)} = p$$



— $\langle C \rangle = 0.61$ for *Facebook* social circles [NL12] while $\rho < 10^{-6}$

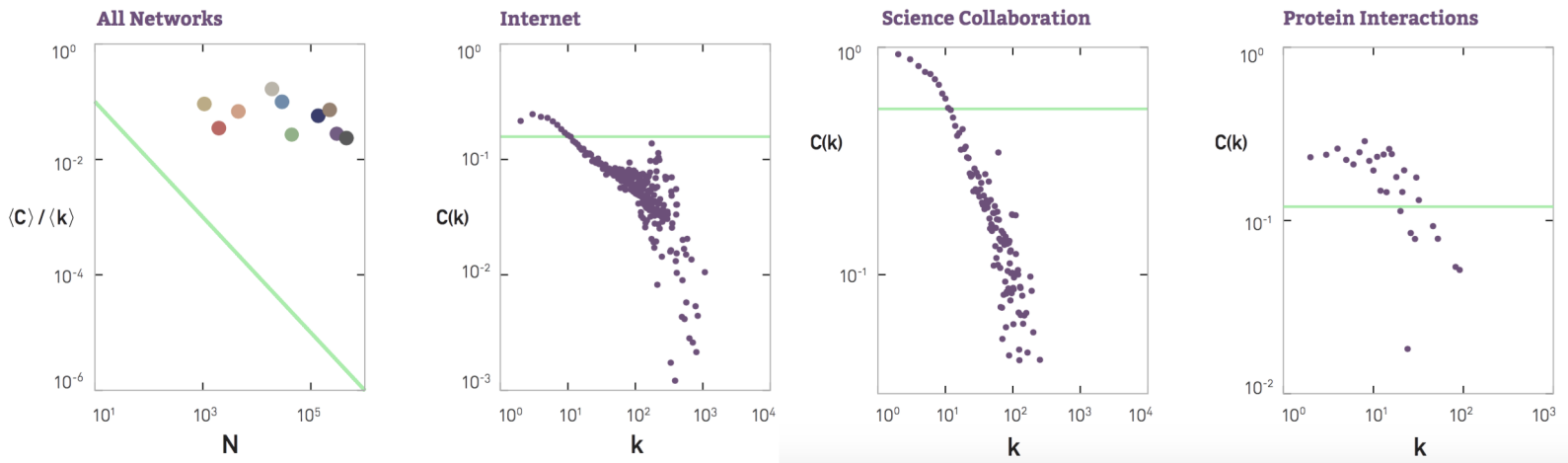
— random graphs *lack clustering* for $n \gg \langle k \rangle$ opposed to *lattices*

network *clustering*

— *clustering* $\langle C \rangle$ and $C_i(k)$ of real networks [Bar16]

— C_i *under-*/*overestimated* for *low-*/*high-k* nodes

— random graphs *substantially underestimate* $\langle C \rangle$



graph *references*

A.-L. Barabási.

Network Science.
Cambridge University Press, Cambridge, 2016.

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna.

Four degrees of separation.
In *Proceedings of the ACM International Conference on Web Science*, pages 45–54, Evanston, IL, USA, 2012.

David Easley and Jon Kleinberg.

Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, Cambridge, 2010.

P. Erdős and A. Rényi.

On random graphs I.
Publ. Math. Debrecen, 6:290–297, 1959.

Mark E. J. Newman.

Networks: An Introduction.
Oxford University Press, Oxford, 2010.

Azree Nazri and Pietro Lio.

Investigating meta-approaches for reconstructing gene networks in a mammalian cellular context.
PLoS ONE, 7(1):e28713, 2012.

M. E. J. Newman, S. H. Strogatz, and D. J. Watts.

Random graphs with arbitrary degree distributions and their applications.
Phys. Rev. E, 64(2):026118, 2001.

Ray Solomonoff and Anatol Rapoport.

Connectivity of random nets.
Bulletin of Mathematical Biophysics, 13(2):107–117, 1951.

graph *references*

D. J. Watts and S. H. Strogatz.
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442, 1998.

configuration graph model

introduction to *network analysis* (*ina*)

Lovro Šubelj
University of Ljubljana
spring 2019/20

configuration *model*

- random graphs *Poisson distribution* $p_k \simeq \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!}$ [ER59]
- real networks *power-law degree distribution* $p_k \sim k^{-\gamma}$ [BA99]
- *configuration model* random graph for *arbitrary* $\{k\}$ [NSW01]

assume *undirected* G from now on

Mark Newman

Steven Strogatz

Duncan Watts

configuration $G(\{k\})$ *model*

— $G(\{k\})$ *configuration model* [NSW01]

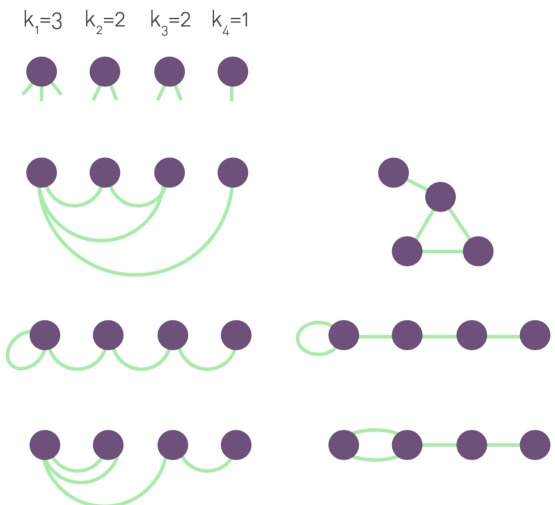
— *randomly link* m *stub pairs* between n nodes

— *computationally convenient* and *analytically tractable*

graphical $k_1, k_2 \ldots k_n$ $\qquad$ $m = \frac{1}{2}\sum_i k_i$

input sequence $\{k\}$
output graph G

1: $G \leftarrow n$ nodes with $\{k\}$ stubs
2: **while** G has node stubs **do**
3: link random node stub pair
4: **end while**
5: **return** G



configuration *probability*

— *probability of self-loop* p_i on i

$$p_i = m\frac{\binom{k_i}{2}}{\binom{2m}{2}} \approx \frac{k_i(k_i-1)}{4m}$$

— *probability of link* p_{ij} between i and j

$$p_{ij} = m\frac{k_i k_j}{\binom{2m}{2}} = k_i\frac{k_j}{2m-1} \approx \frac{k_i k_j}{2m}$$

— thus *number of multilinks* and *self-loops* is

$$\left[\frac{\langle k^2\rangle - \langle k\rangle}{\sqrt{2}\langle k\rangle}\right]^2 \qquad \sum_i p_i = \sum_i \frac{k_i(k_i-1)}{2n\langle k\rangle} = \frac{\langle k^2\rangle - \langle k\rangle}{2\langle k\rangle}$$

configuration *neighbors*

— *neighbor degree distribution* p_k is *not* p_k

– n_k is *number of degree-k* nodes thus $n_k = np_k$

$$\{\textit{neighbor } p_k\} = n_k \frac{k}{2m-1} \approx \frac{kp_k}{\langle k \rangle}$$

— *average neighbor degree* $\langle k \rangle$ is *not* $\langle k \rangle$

$$\frac{\langle k^2 \rangle}{\langle k \rangle} - \langle k \rangle = \frac{\langle k^2 \rangle - \langle k \rangle^2}{\langle k \rangle} = \frac{\sigma_k^2}{\langle k \rangle} > 0$$

$$\langle \textit{neighbor } k \rangle \approx \sum_k k \frac{kp_k}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle} > \langle k \rangle$$

— $\frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\langle k \rangle^2 + \langle k \rangle}{\langle k \rangle} = \langle k \rangle + 1$ even for *Poisson graph* [ER59]

network *neighbors*

— *friendship paradox* $\langle$*neighbor* $k\rangle > \langle k\rangle$ [Fel91] in real networks

— $\langle$*neighbor* $k\rangle$ *well estimated* by $\frac{\langle k^2\rangle}{\langle k\rangle}$ whereas $\langle k\rangle \ll \frac{\langle k^2\rangle}{\langle k\rangle}$

network	n	$\langle k\rangle \ll$	$\langle$**neighbor** $k\rangle$	$\approx \frac{\langle k^2\rangle}{\langle k\rangle}$
Southern women [DGG41]	32	5.56	7.57	7.02
Karate club [Zac77]	34	4.59	9.61	7.77
American football [GN02]	115	10.71	10.78	10.79
Java dependencies [ŠB11]	1368	16.20	207.52	140.53
Facebook circles [ML12]	4039	43.69	105.55	106.57
Physics collaboration [New01]	36 458	9.42	21.65	27.88
Enron e-mails [LLDM09]	36 692	20.04	472.86	280.16
Internet map [HJJ+03]	75 885	9.42	1853.73	1461.54
Actors collaboration [BA99]	382 219	78.69	282.72	417.69
Physics citation [ŠFB14]	438 943	21.56	78.38	77.72
Patent citation [HJT01]	3 774 768	8.75	17.15	21.33
Facebook snowball [Fer12]	8 217 272	3.06	308.52	157.06

configuration *clustering*

— (*neighbor*) *excess degree distribution* q_k defined as

excess degree is remaining neighbor degree or neighbor degree-1

$$q_k = \frac{(k+1)p_{k+1}}{\langle k \rangle}$$

— then *network clustering coefficient* C [NSW01] is

$$\sum_{k_i,k_j} q_{k_i} q_{k_j} \frac{k_i k_j}{2m} = \frac{1}{2m} \left[\sum_k k q_k\right]^2 = \frac{1}{2m\langle k \rangle^2} \left[\sum_k k(k+1)p_{k+1}\right]^2 = \frac{1}{n\langle k \rangle^3} \left[\sum_k (k-1)k p_k\right]^2$$

$$C = \sum_{k_i,k_j} q_{k_i} q_{k_j} p_{ij} \approx \frac{\left[\langle k^2 \rangle - \langle k \rangle\right]^2}{n\langle k \rangle^3}$$

network *clustering*

— *average clustering coefficient* $\langle C \rangle$ [WS98] of real networks

— *neither* $G(n, p)$ [ER59] *nor* $G(\{k\})$ [NSW01] *explain* $\langle C \rangle \gg 0$

network	n	$\langle C \rangle$	$\gg \frac{[\langle k^2 \rangle - \langle k \rangle]^2}{n \langle k \rangle^3}$	$\gg \frac{\langle k \rangle}{n-1}$
Southern women [DGG41]	32	0.000	0.204	0.179
Karate club [Zac77]	34	0.571	0.294	0.139
American football [GN02]	115	0.403	0.078	0.094
Java dependencies [ŠB11]	1368	0.497	0.879	0.012
Facebook circles [ML12]	4039	0.606	0.063	0.011
Physics collaboration [New01]	36 458	0.657	0.002	0.000
Enron e-mails [LLDM09]	36 692	0.497	0.106	0.001
Internet map [HJJ+03]	75 885	0.160	2.985	0.000
Actors collaboration [BA99]	382 219	0.780	0.006	0.000
Physics citation [ŠFB14]	438 943	0.227	0.001	0.000
Patent citation [HJT01]	3 774 768	0.076	0.000	0.000
Facebook snowball [Fer12]	8 217 272	0.019	0.001	0.000

configuration *references*

A.-L. Barabási and R. Albert.

Emergence of scaling in random networks.
Science, 286(5439):509–512, 1999.

A.-L. Barabási.

Network Science.
Cambridge University Press, Cambridge, 2016.

A. Davis, B. B. Gardner, and M. R. Gardner.

Deep South.
Chicago University Press, Chicago, 1941.

David Easley and Jon Kleinberg.

Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, Cambridge, 2010.

P. Erdős and A. Rényi.

On random graphs I.
Publ. Math. Debrecen, 6:290–297, 1959.

Scott. L. Feld.

Why your friends have more friends than you do.
Am. J. Sociol., 96(6):1464–1477, 1991.

Stefano Ferretti.

On the degree distribution of faulty peer-to-peer overlays.
ICST Transactions on Complex Systems, 2012.

M. Girvan and M. E. J Newman.

Community structure in social and biological networks.
P. Natl. Acad. Sci. USA, 99(12):7821–7826, 2002.

configuration *references*

M Hoerdt, M Jaeger, A James, D Magoni, J Maillard, D Malka, and P Merindol.
Internet {IP}v4 overlay map produced by network cartographer (nec), 2003.

B. H. Hall, A. B. Jaffe, and M. Tratjenberg.
The NBER patent citation data file: Lessons, insights and methodological tools.
Technical report, National Bureau of Economic Research, 2001.

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney.
Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters.
Internet Math., 6(1):29–123, 2009.

Seth A. Myers and Jure Leskovec.
Clash of the contagions: Cooperation and competition in information diffusion.
In *Proceedings of the IEEE International Conference on Data Mining*, 2012.

M. E. J. Newman.
The structure of scientific collaboration networks.
P. Natl. Acad. Sci. USA, 98(2):404–409, 2001.

Mark E. J. Newman.
Networks: An Introduction.
Oxford University Press, Oxford, 2010.

M. E. J. Newman, S. H. Strogatz, and D. J. Watts.
Random graphs with arbitrary degree distributions and their applications.
Phys. Rev. E, 64(2):026118, 2001.

Lovro Šubelj and Marko Bajec.
Community structure of complex software systems: Analysis and applications.
Physica A, 390(16):2968–2975, 2011.

configuration *references*

Lovro Šubelj, Dalibor Fiala, and Marko Bajec.
Network-based statistical comparison of citation topology of bibliographic databases.
Sci. Rep., 4:6496, 2014.

D. J. Watts and S. H. Strogatz.
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442, 1998.

Wayne W. Zachary.
An information flow model for conflict and fission in small groups.
J. Anthropol. Res., 33(4):452–473, 1977.

real networks models

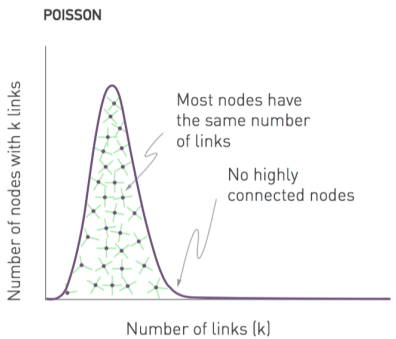
introduction to *network analysis* (*ina*)

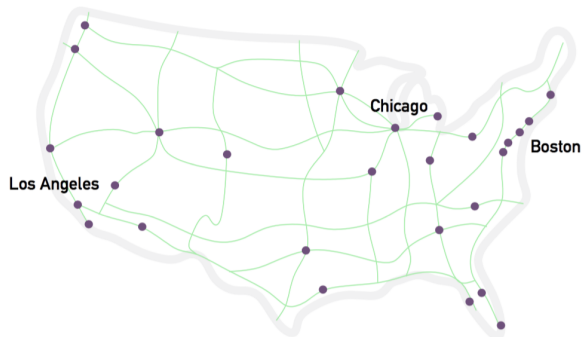
Lovro Šubelj
University of Ljubljana
spring 2019/20

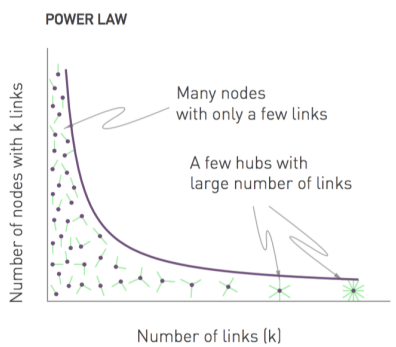
networks *scale-free*

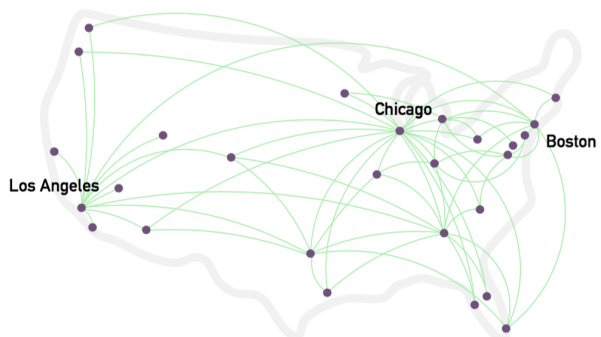
— *power-law* degree distribution $p_k \sim k^{-\gamma}$ [Pri65]

— *preferential attachment scale-free* model [BA99]





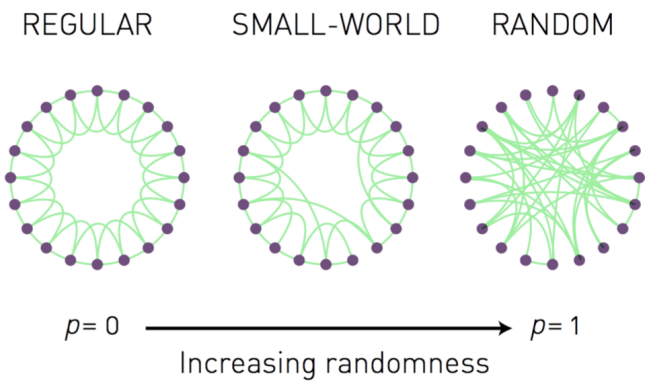


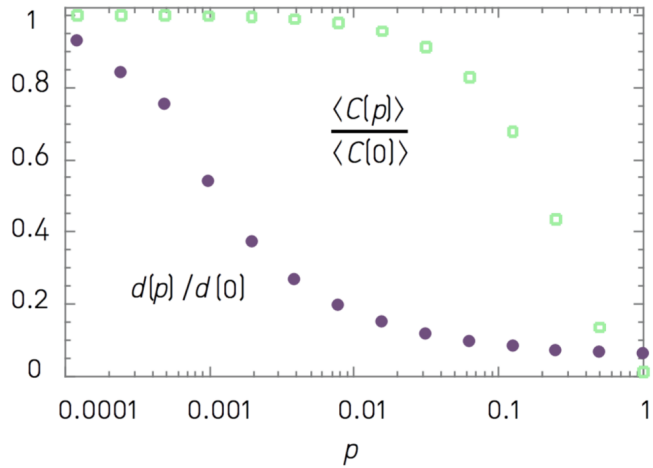


networks *small-world*

— *coexistence* of $\langle C \rangle \gg 0$ and $\langle d \rangle \simeq \frac{\ln n}{\ln \langle k \rangle}$

— *link rewiring small-world* model [WS98]





networks *references*

A.-L. Barabási and R. Albert.
Emergence of scaling in random networks.
Science, 286(5439):509–512, 1999.

A.-L. Barabási.
Network Science.
Cambridge University Press, Cambridge, 2016.

D. J. de Solla Price.
Networks of scientific papers.
Science, 149:510–515, 1965.

D. J. Watts and S. H. Strogatz.
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442, 1998.