Aalto University
School of Science

CS-E4600 – Algorithmic methods of data mining

Slide set 10 : Introduction to graph mining

Aristides Gionis

Aalto University

fall 2018

introduction to graphs and networks

reading material

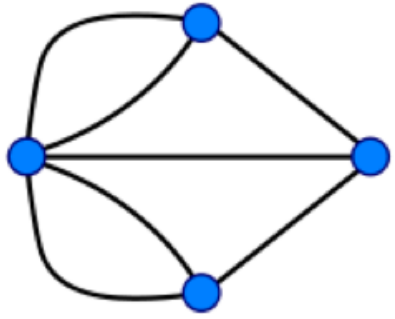
- "*Graph mining : laws, generators, and algorithms*"

  survey paper by Chakrabarti and Faloutsos

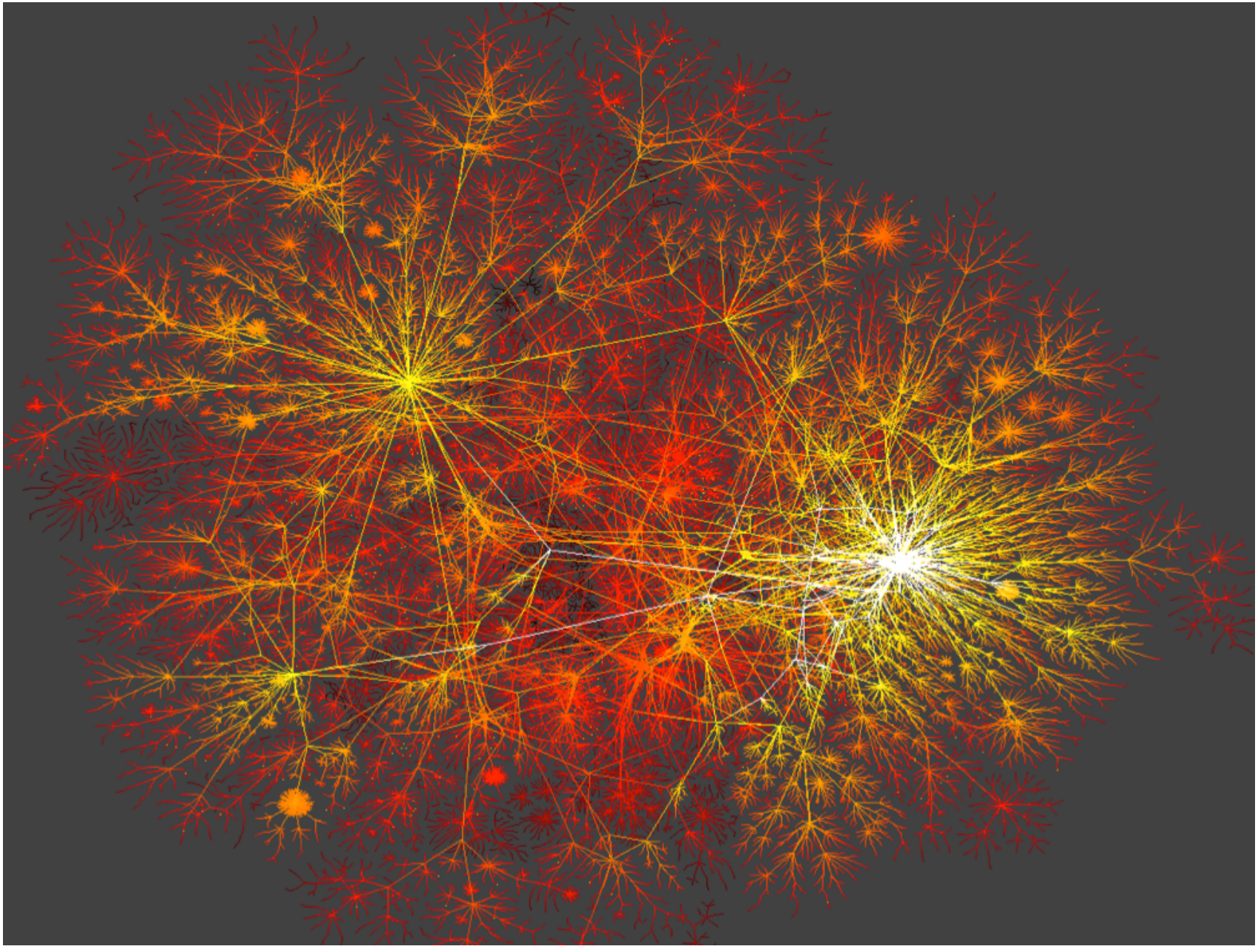
  long paper – skim through and identify the relevant parts

graphs: a simple model

- entities — set of vertices
- pairwise relations among entities
  — set of edges
- can add directions, weights,. . .
- used to model many real-world datasets



the internet map



types of networks

- social networks
- knowledge and information networks
- technology networks
- biological networks

network science

- the world is full with networks
- what do we do with them?
  - understand their topology and measure their properties
  - study their evolution and dynamics
  - create realistic models
  - create algorithms to make sense of network data

properties of real-world networks

properties of real-world networks

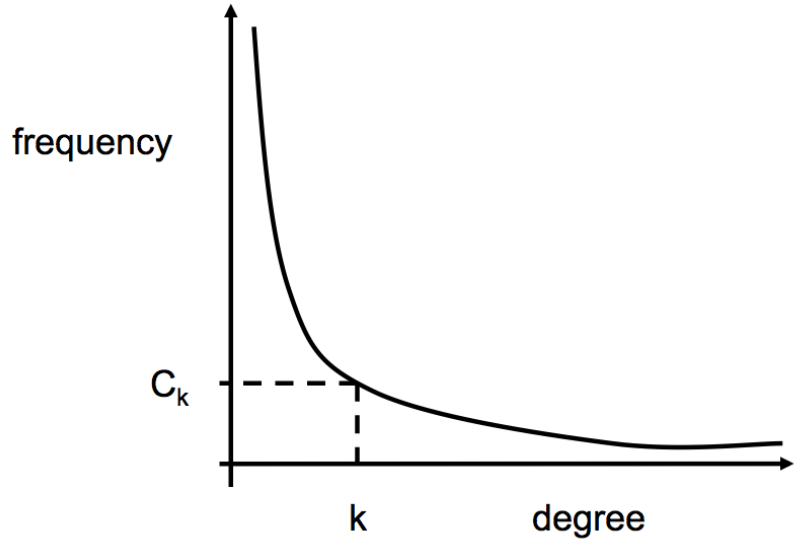
diverse collections of graphs arising from different phenomena

— are there typical patterns ? yes !

- static networks
  - heavy tails
  - clustering coefficients
  - communities
  - small diameters
- time-evolving networks
  - densification
  - shrinking diameters

degree distribution

- C_k = number of vertices with degree k



- problem : find the probability distribution that fits best the observed data

power-law degree distribution

- C_k = number of vertices with degree k, then

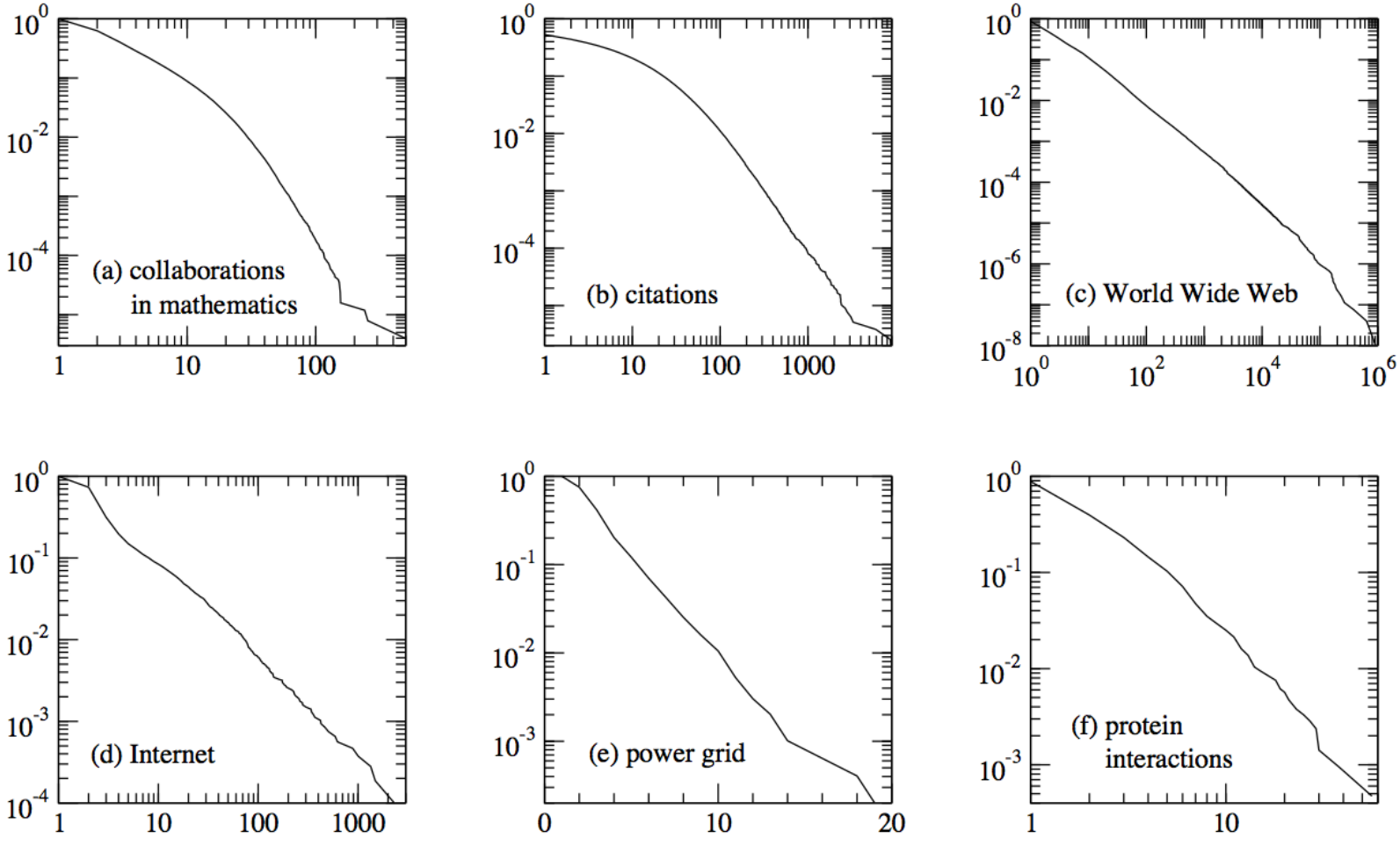
$$C_k = c\,k^{-\gamma}$$

with $\gamma > 1$, or

$$\ln C_k = \ln c - \gamma \ln k$$

- plotting $\ln C_k$ versus $\ln k$ gives a straight line with slope $-\gamma$
- heavy-tail distribution : there is a non-negligible fraction of nodes that has very high degree (hubs)
- scale free : average is not informative

power-law degree distribution



power-laws in a wide variety of networks [Newman, 2003]

sheer contrast with Erdős-Rényi random graphs

clustering coefficients

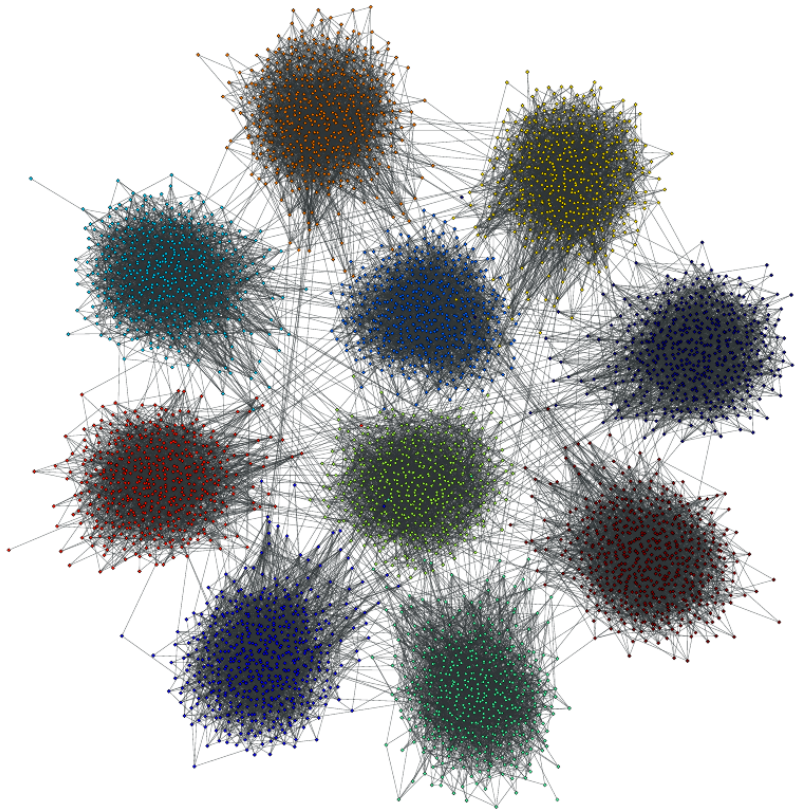
- a proposed measure to capture local clustering is graph transitivity

$$T(G) = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}}$$

- captures "transitivity of clustering"
- if u is connected to v and v is connected to w, it is also likely that u is connected to w

community structure

loose definition of community: a set of vertices densely connected to each other and sparsely connected to the rest of the graph



artificial communities:

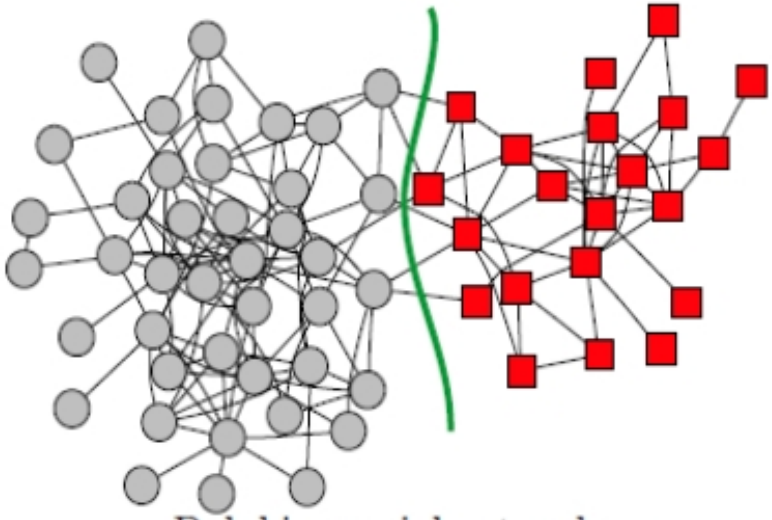
http://projects.skewed.de/graph-tool/

community structure

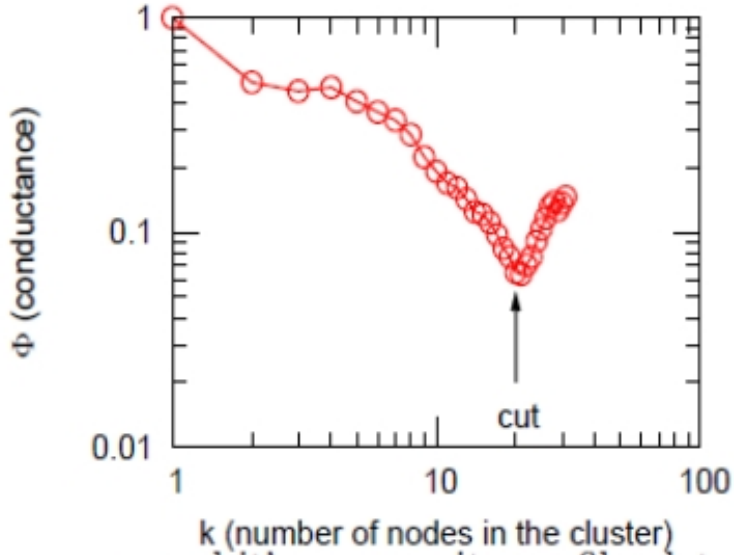
[Leskovec et al., 2009]

- study community structure in an extensive collection of real-world networks
- introduce the network community profile (NCP) plot
- characterizes the best possible community over a range of scales

community structure



Dolphins social network . . .

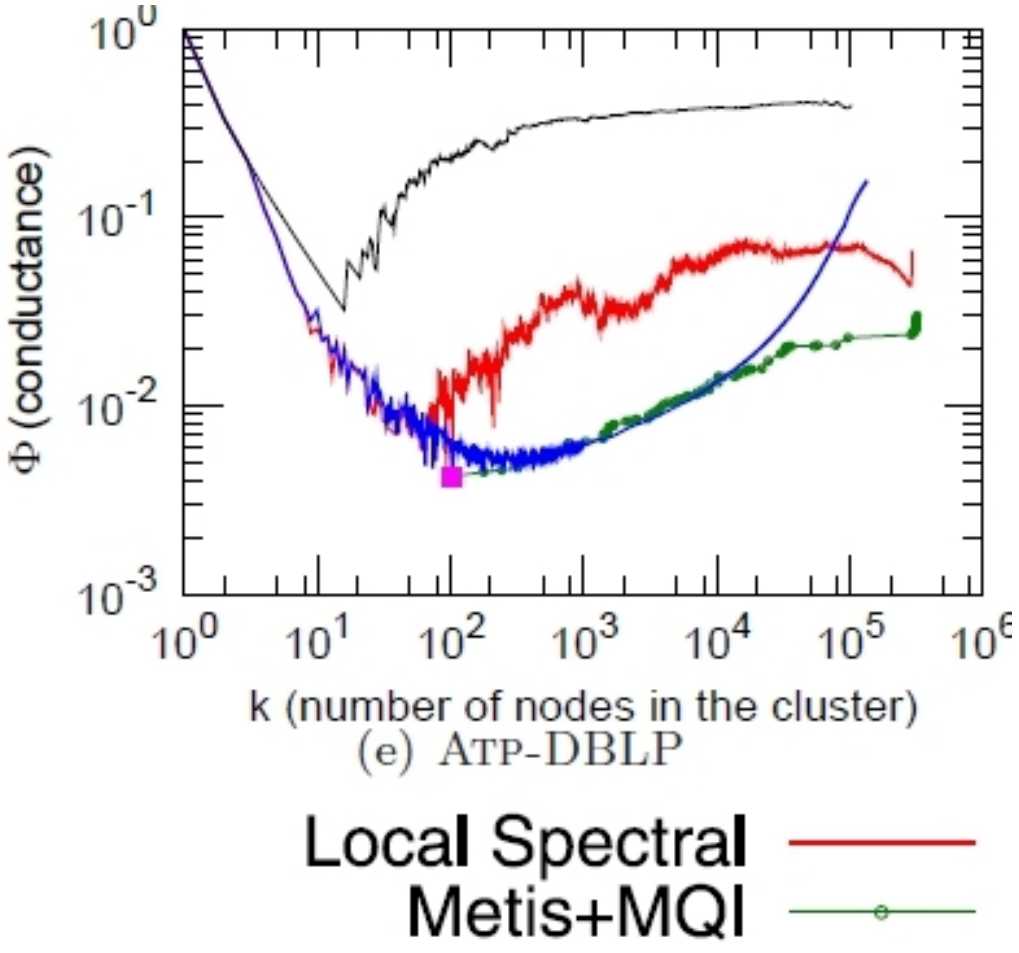


. . . and it's community profile plot

`dolphins` network and its NCP [Leskovec et al., 2009]

community structure

- do large real networks have such nice structure ? NO !



NCP of a DBLP graph (source [Leskovec et al., 2009])

community structure

important findings of [Leskovec et al., 2009]

1. up to certain size k ($\sim$ 100 vertices) there are good cuts
   - as the size increases so does the quality of the community
2. at the size k we observe the best possible community
   - such communities are typically connected to the remainder with a single edge
3. above the size k the community quality decreases
   - this is because they blend in and gradually disappear

small-world phenomena

small worlds : graphs with short paths

- Stanley Milgram (1933-1984)
  “The man who shocked the world”
- obedience to authority (1963)
- small-World experiment (1967)

  — we live in a small-world

for criticism on the small-world experiment, see *“Could It Be a Big World After All? What the Milgram Papers in the Yale Archives Reveal About the Original Small World Study”* by Judith Kleinfeld

small-world experiments

- letters were handed out to people in Nebraska to be sent to a target in Boston
- people were instructed to pass on the letters to someone they knew on first-name basis
- the letters that reached the destination (64 / 296) followed paths of length around 6
- *Six degrees of separation* : (play of John Guare)

  also :
  — the Kevin Bacon game
  — the Erdős number

small diameter

proposed measures

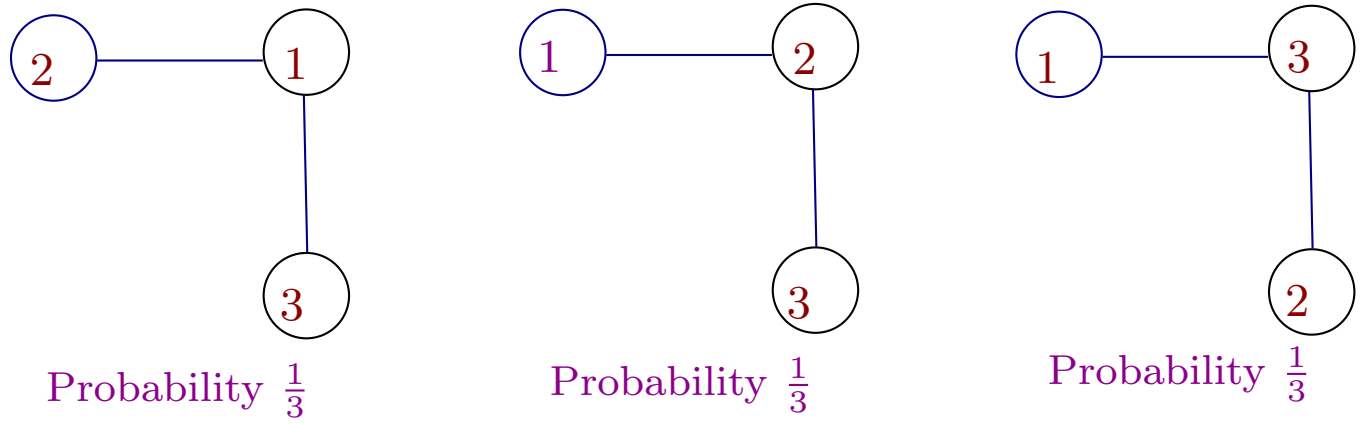
- diameter : largest shortest-path over all pairs
- effective diameter : upper bound of the shortest path of 90% of the pairs of vertices
- average shortest path : average of the shortest paths over all pairs of vertices
- characteristic path length : median of the shortest paths over all pairs of vertices
- hop-plots : plot of $|N_h(u)|$, the number of neighbors of u at distance at most h, as a function of h

  [Faloutsos et al., 1999].

Erdős-Rényi graphs

random graphs

- a random graph is a set of graphs together with a probability distribution on that set
- example



a random graph on $\{1, 2, 3\}$ with 2 edges with the uniform distribution

random graphs

- the $G(n, p)$ model:
- n : the number of vertices
- $0 \leq p \leq 1$: probability
- for each pair (u, v), independently generate the edge (u, v) with probability p
- $G(n, p)$ a family of graphs, in which a graph with m edges appears with probability $p^m(1-p)^{\binom{n}{2}-m}$

- the $G(n, m)$ model: related, but not identical

degree distribution

- degree distribution : binomial

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

- the limit distribution of the normalized binomial distribution $\mathrm{Bin}(n,p)$ is the normal distribution provided that $np(1-p) \to +\infty$ as $n \to +\infty$
- if $p = \frac{\lambda}{n}$ the limit distribution of $Bin(n,p)$ is the Poisson distribution

random graphs and real datasets

- a beautiful and elegant theory studied exhaustively
- have been used as idealized generative models
- unfortunately, they don't always capture reality. . .

models of real-world networks

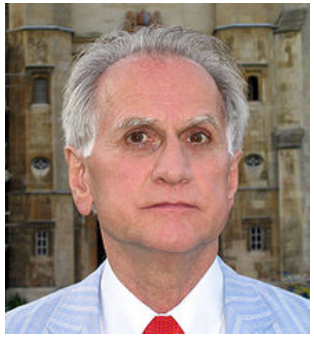
models

- growth with preferential attachment
- structure $+$ randomness $\rightarrow$ small-world networks
- forest-fire model

preferential attachment

R. Albert

L. Barabási



B. Bollobás

O. Riordan

growth model:

- at time n, vertex n is added to the graph
- one edge is attached to the new vertex
- the other vertex is selected at random with probability proportional to its degree
- obtain a sequence of graphs $\{G_1^{(n)}\}$
- power law distribution arises!

small-world models

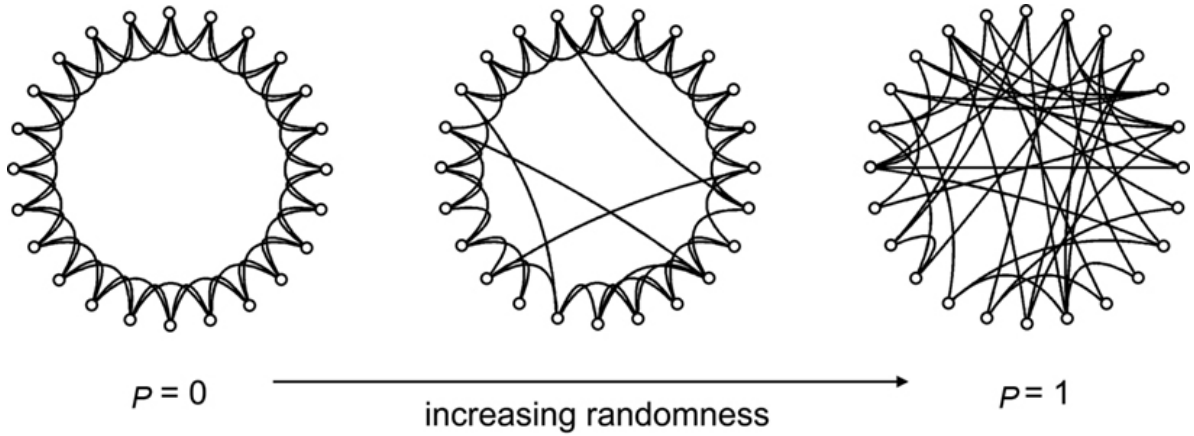
Duncan Watts

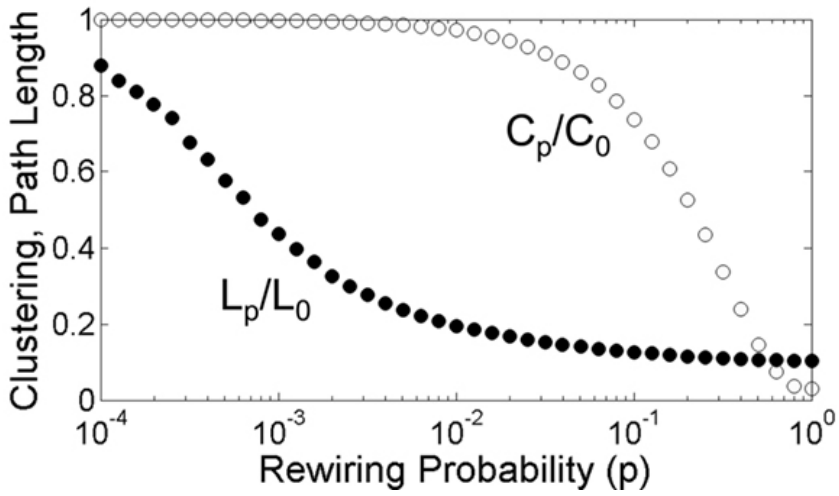
Steven Strogatz

construct a network with

- small diameter
- positive density of triangles

small-world models

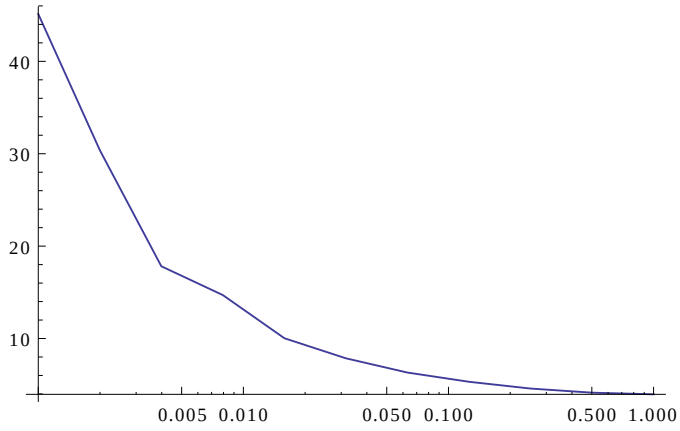




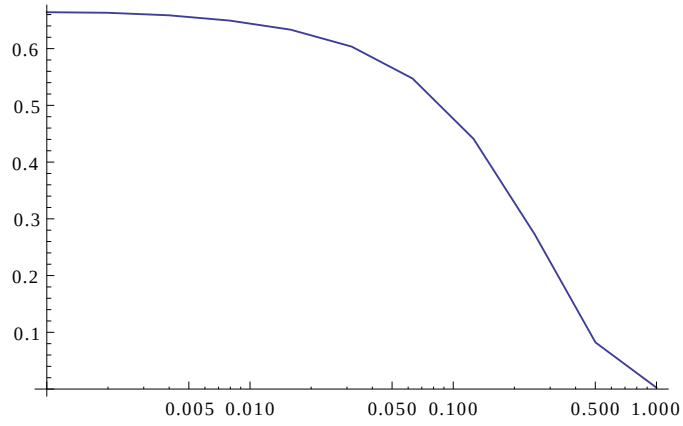
rewiring probability, p

even for a small value of p, $L(G(p))$ drops to $O(\log n)$, which $C(G(p)) \approx \frac{3}{4}$

small-world models



average distance



clustering coefficient

Watts-Strogatz graph on 4 000 vertices, starting from a 10-regular graph

- intuition: if you add a little bit of randomness to a structured graph, you get the small world effect
- related work: see [Bollobás and Chung, 1988]

navigation in a small world

how to find short paths using only local information?

- we will use a simple directed model [Kleinberg, 2000]
- underlying graph : 2-D grid $+$ long-range edges
- a local algorithm
  - can remember the source, the destination and its current location
  - can query the graph to find the long-distance edge at the current location
  - at each location, follow the edge whose end-point is the closest to the destination

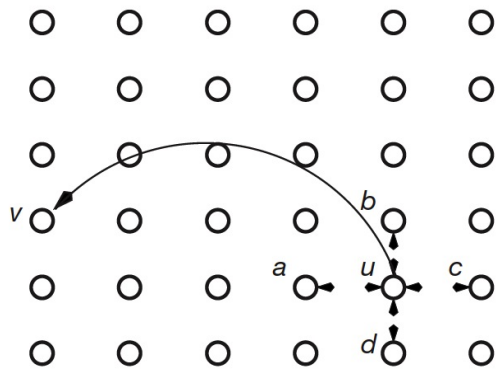
navigation in a small world

$d(u, v)$: shortest path distance using only original grid edges

directed graph model, parameter r :

- each vertex is connected to its four adjacent vertices
- for each vertex v we add an extra link (v, u) where u is chosen with probability proportional to $d(v, u)^{-r}$

notice: compared to the Watts-Strogatz model the long range edges are added in a biased way



(source [Kleinberg, 2000])

navigation in a small world

- $r = 0$: random edges, independent of distance
- as r increases the length of the long distance edges decreases in expectation

navigation in a small world

- $r = 0$: random edges, independent of distance
- as r increases the length of the long distance edges decreases in expectation

results

1. $r < 2$: the end points of the long distance edges tend to be uniformly distributed over the vertices of the grid
   - is unlikely on a short path to encounter a long distance edge whose end point is close to the destination
   - no local algorithm can find them
2. $r = 2$: there are short paths
   - a short path can be found be the simple algorithm that always selects the edge that takes closest to the destination

2. $r > 2$: there are no short paths, with high probability

acknowledgements

Paolo Boldi

Charalampos Tsourakakis

references


Bollobás, B. and Chung, F. R. K. (1988).
The diameter of a cycle plus a random matching.
SIAM Journal on discrete mathematics, 1(3):328–333.


Clauset, A., Shalizi, C. R., and Newman, M. E. (2009).
Power-law distributions in empirical data.
SIAM review, 51(4):661–703.


Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999).
On power-law relationships of the internet topology.
In *SIGCOMM*.


Kleinberg, J. M. (2000).
Navigation in a small world.
Nature, 406(6798):845–845.


Lakhina, A., Byers, J. W., Crovella, M., and Xie, P. (2003).
Sampling biases in ip topology measurements.
In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1, pages 332–341. IEEE.

references (cont.)

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005).

Graphs over time: densification laws, shrinking diameters and possible explanations.

In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA. ACM Press.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007).

Graph evolution: Densification and shrinking diameters.

ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):2.

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2009).

Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters.

Internet Mathematics, 6(1):29–123.

Li, L., Alderson, D., Doyle, J. C., and Willinger, W. (2005).

Towards a theory of scale-free graphs: Definition, properties, and implications.

Internet Mathematics, 2(4):431–523.

references (cont.)

Newman, M. E. J. (2003).

The structure and function of complex networks.

Tsourakakis, C. E. (2008).

Fast counting of triangles in large real networks without counting: Algorithms and laws.

In *ICDM*.