



Aalto University
School of Science

Algorithmic methods for mining large graphs

Lecture 1 : Introduction to graph mining

Aristides Gionis

Aalto University

Bertinoro International Spring School 2016

March 7–11, 2016

tutorial website

slides available at

<https://users.ics.aalto.fi/gionis/biss2016.shtml>

lots of relevant papers are listed on the slides

you should attend this course if you ...

- are interested in algorithms and/or data mining
- want to dive deeper and do research in some of the course topics
- enhance your toolbox with techniques that may come handy in other research problems
- want to obtain a broad view on the area
- desperately need the credit units
- got stuck on a castle on a top of a hill on a rainy spring

about the course

- I will try to keep the lectures interactive
- questions and discussion are encouraged
- we do not have to cover all the material in the slides
- homework
 - will hand out at the last day
 - two weeks deadline
 - return to me by email

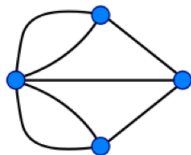
course agenda

- introduction to graph mining Tue afternoon
- computing basic graph statistics Tue afternoon, Wed morning
- finding dense subgraphs Wed afternoon, Thu morning
- spectral graph analysis Thu afternoon
- additional topics Fri morning
- inferring hierarchies in graphs
- mining dynamic graphs
- graph sparsifiers

introduction to graphs and networks

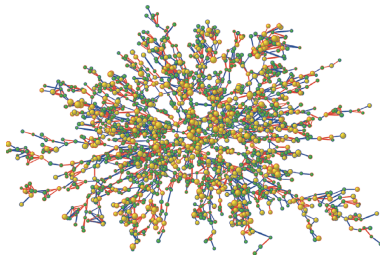
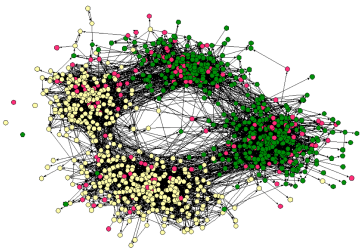
graphs: a simple model

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



analysis of graph datasets in the past

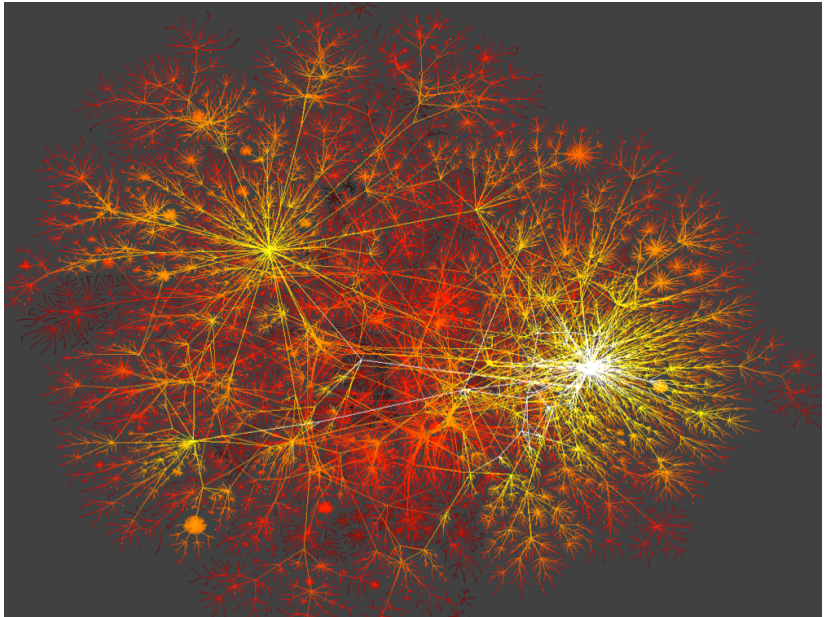
- graphs datasets have been studied in the past
e.g., networks of highways, social networks
- usually these datasets were **small**
- **visual inspection** can reveal useful information



analysis of graph datasets now

- more and larger networks are collected
- networks of thousands, millions, or billions of nodes
- impossible to visualize

the internet map

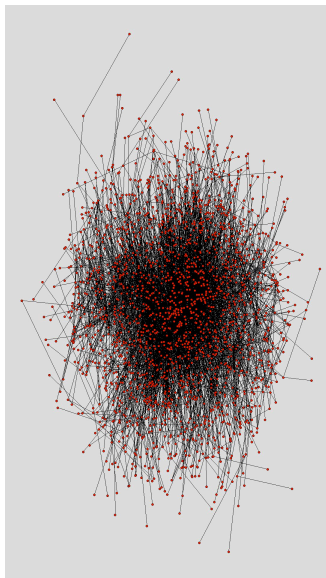


types of networks

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

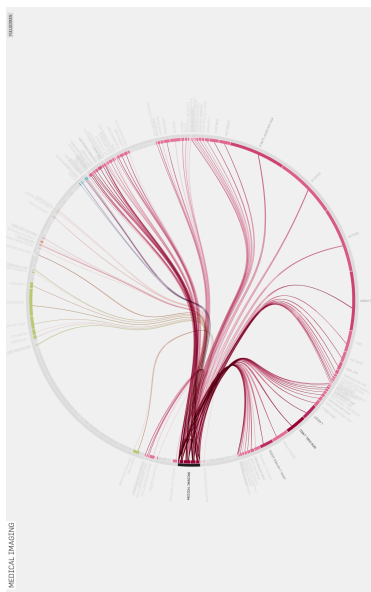
social networks

- links denote a **social interaction**
 - networks of acquaintances
 - collaboration networks
 - actor networks
 - co-authorship networks
 - director networks
 - phone-call networks
 - **e-mail** networks
 - IM networks
 - sexual networks



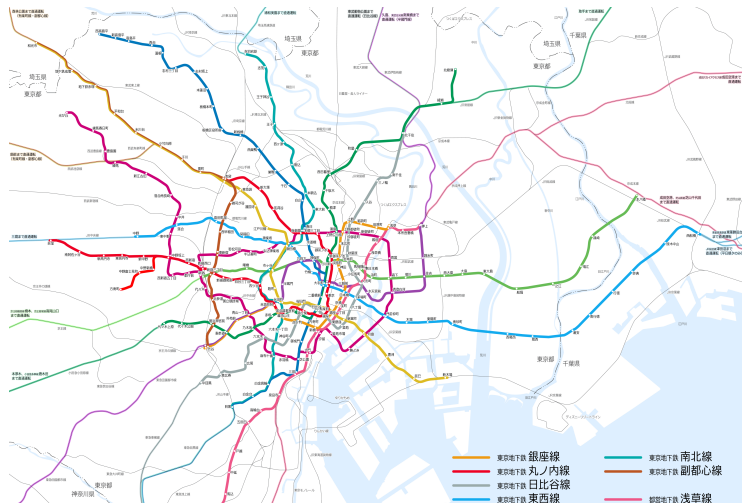
knowledge and information networks

- nodes store information
- links **associate** information
 - **citation** network (directed acyclic)
 - the web (directed)
 - peer-to-peer networks
 - word networks
 - networks of trust
 - software graphs
 - bluetooth networks
 - home page/blog networks



technological networks

- networks built for **distribution of a commodity**
 - the internet, power grids, telephone networks, airline networks, transportation networks



biological networks

- **biological systems** represented as networks
 - **protein-protein interaction** networks
 - gene regulation networks
 - gene co-expression networks
 - metabolic pathways
 - the **food web**
 - neural networks

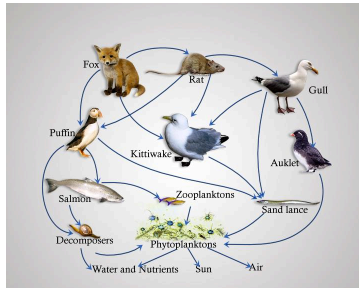
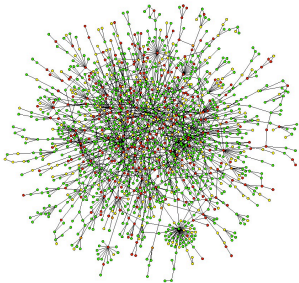


photo-sharing site

flickr from YAHOO!

[Home](#) [You](#) [Organize & Create](#) [Contacts](#) [Groups](#) [Explore](#) [Upload](#)

Signed in as [Aris Gionis](#) [Log Out](#) [Help](#)

[Search](#)

[Favorite](#) [Actions](#) [Share](#)

[← Newer](#) [Older →](#)



Rosenborg, Copenhagen

19,365

Rosenborg Castle - where we keep the Kingdoms crown jewels.

This beautiful spot is in the heart of Copenhagen, at the Kings Garden. The photograph was shot on a nice spring day, with wonderful flickr friends on a Copenhagen walk.

[Comments and faves](#)

By [michael.dreves](#)
[Michael Dreves Beier](#) [+ Add Contact](#)

This photo was taken on April 7, 2010 in Tornebuskagade, Copenhagen, Hovedstaden, DK, using a Canon EOS 5D Mark II.



161 7,584 [390](#) [190](#) [0](#)

This photo belongs to

[michael.dreves' photostream](#) (454)



This photo also appears in

- [flickr - Most Interesting](#) (set)
 - [Project 365](#) (set)
 - [HDR compilations](#) (set)
 - [Copenhagen](#) (set)
 - [***Flickr Global](#) (group)
 - [Art of Images..\(P1A3\) / Not...](#) (group)
 - [Denmark](#) (group)
 - [FlickrCentral](#) (group)
 - [FlickrToday \(only 1 pic per day\)](#) (group)
- ...and 63 more groups

People in this photo [\(add a person\)](#)

Adding people will share who is in this photo

what is the underlying graph?

- **nodes:** photos, tags, users, groups, albums, sets, collections, geo, query, . . .
- **edges:** upload, belong, tag, create, join, contact, friend, family, comment, fave, search, click, . . .
- tons of interesting graphs to work with
 - **tag graph:** based on photos
 - **tag graph:** based on users
 - **user graph:** based on favorites
 - **user graph:** based on groups
- which graph to pick? — **depends on the application**

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

heavy tails

What do the proteins in our bodies, the Internet, a cool collection of atoms and sexual networks have in common? One man thinks he has the answer and it is going to transform the way we view the world.

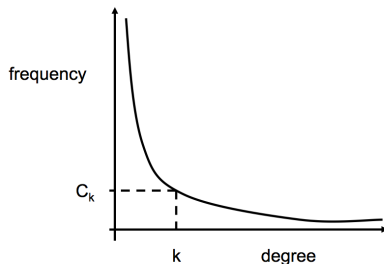
Scientist 2002



Albert-László Barabási

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

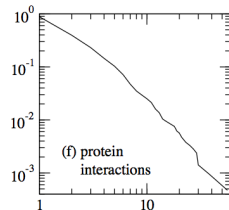
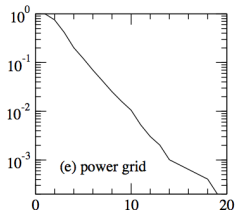
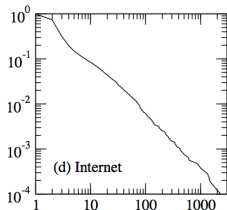
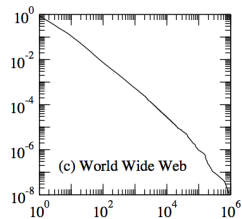
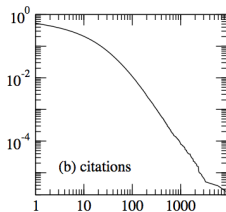
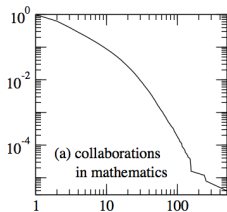
$$C_k = c k^{-\gamma}$$

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

power-law degree distribution

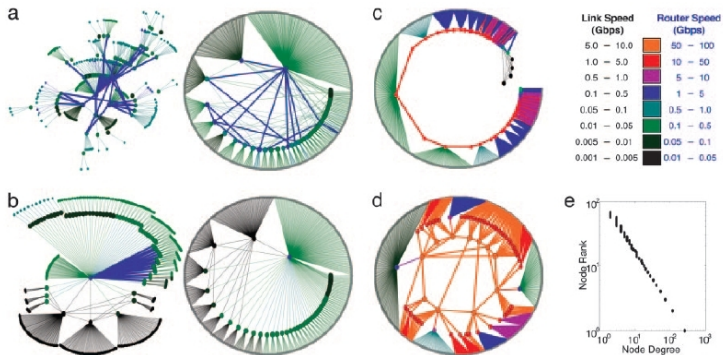
do the degrees follow a power-law distribution?

three **problems** with the initial studies

- graphs generated with **traceroute sampling**, which produces power-law distributions, even for regular graphs [Lakhina et al., 2003].
- methodological flaws in determining the exponent see [Clauset et al., 2009] for a proper methodology
- other distributions could potentially fit the data better but were not considered, e.g., **lognormal**.

disclaimer: we will be referring to these distributions as **heavy-tailed**, avoiding a specific characterization

power-law degree distribution



all networks above have the same degree sequence but
structurally are very different [Li et al., 2005]

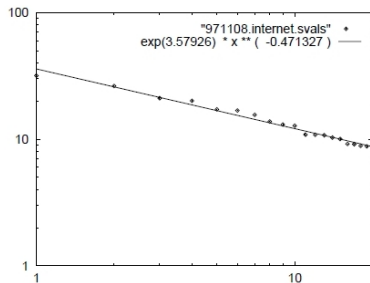
maximum degree

- for random graphs, the maximum degree is highly concentrated around the average degree z
- for power-law graphs

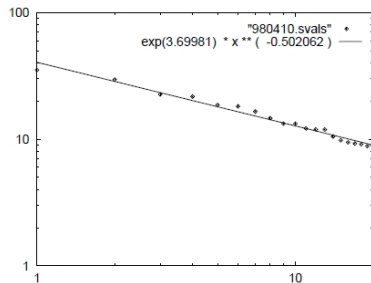
$$d_{\max} \approx n^{1/(\alpha-1)}$$

- hand-waving argument: solve $n \Pr[X \geq d] = \Theta(1)$

heavy tails, eigenvalues



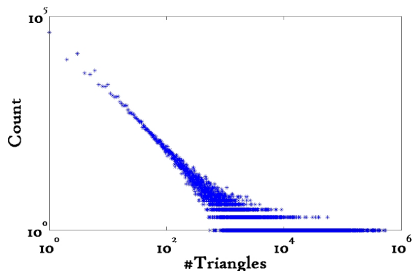
(a) Int-11-97



(b) Int-04-98

- log-log plot of eigenvalues of the Internet graph in decreasing order
- again a power law emerges [Faloutsos et al., 1999]

heavy tails, triangles



- triangle distribution in flickr
- figure shows the count of nodes with k -triangles vs. k in log-log scale
- again, heavy tails emerge [Tsourakakis, 2008]

clustering coefficients

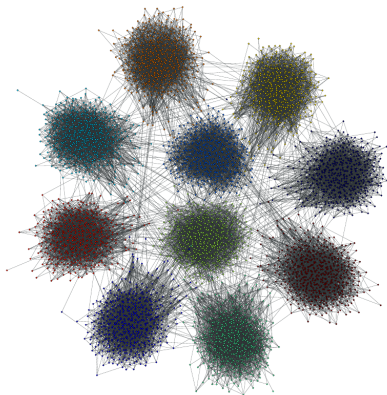
- a proposed measure to capture local clustering is **graph transitivity**

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

- 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:

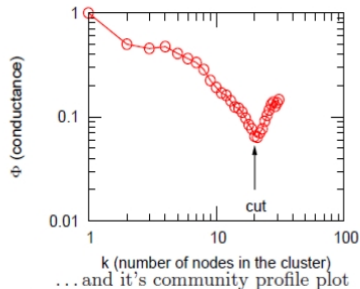
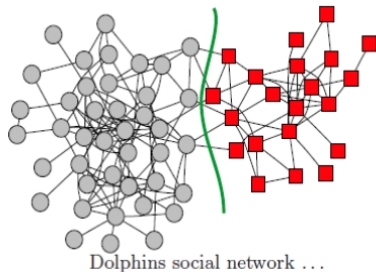
<http://projects.skewed.de/graph-tool/>

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**

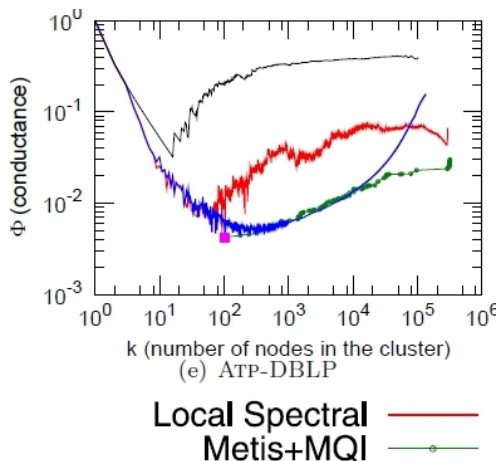
community structure



dolphins 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 (~ 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].

time-evolving networks



J. Leskovec



J. Kleinberg



C. Faloutsos

[Leskovec et al., 2005]

- densification power law:

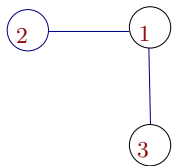
$$|E_t| \propto |V_t|^\alpha \quad 1 \leq \alpha \leq 2$$

- shrinking diameters: diameter is shrinking over time.

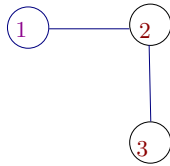
Erdős-Rényi graphs

random graphs

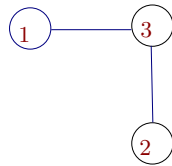
- a random graph is a **set of graphs** together with a **probability distribution** on that set
- **example**



Probability $\frac{1}{3}$



Probability $\frac{1}{3}$

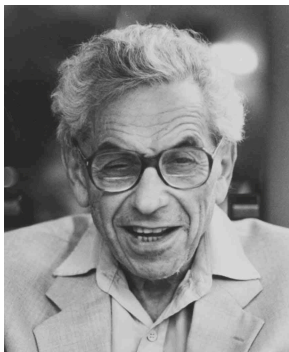


Probability $\frac{1}{3}$

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

random graphs

- Erdős-Rényi (or Gilbert-Erdős-Rényi) random graph model



Paul Erdős
1913 – 1996



Alfréd Rényi
1921 – 1970

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**

properties of random graphs

- a property P holds almost surely/with high probability (whp $\rightarrow 1 - o(1)$) if

$$\lim_{n \rightarrow \infty} \Pr[G \text{ has } P] = 1$$

- which properties hold as p increases?

properties of random graphs

- a property P holds almost surely/with high probability (whp $\rightarrow 1 - o(1)$) if

$$\lim_{n \rightarrow \infty} \Pr[G \text{ has } P] = 1$$

- which properties hold as p increases?
- threshold phenomena : many properties appear suddenly
- there exist a probability p_c such that
 - for $p < p_c$ the property does not hold a.s.
 - for $p > p_c$ the property holds a.s.

the giant component

- let $z = np$ be the average degree
- if $z < 1$ the largest component has size $O(\log n)$ a.s.
- if $z > 1$ the largest component has size $\Theta(n)$ a.s.;
the second largest component has size $O(\log n)$ a.s.
- if $z = \omega(\log n)$ the graph is connected a.s.

phase transition

- if $z = 1$ there is a phase transition
 - the largest component has size $O(n^{2/3})$
 - the sizes of the components follow a power-law

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

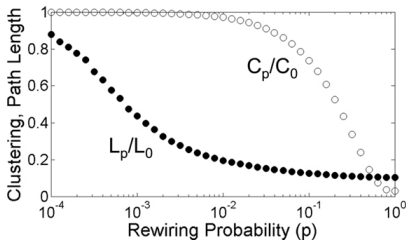
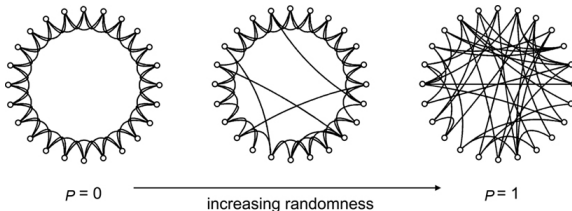


Steven Strogatz

construct a network with

- small diameter
- positive density of triangles

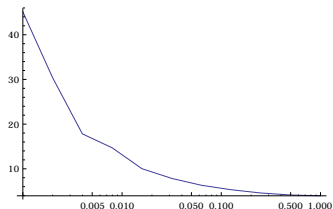
small-world models



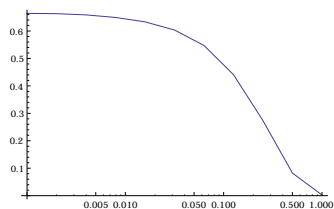
rewiring probability, p

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

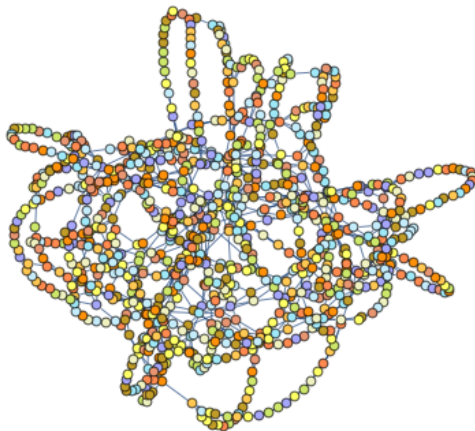


clustering coefficient

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](#)]

small-world models



Watts-Strogatz on 1 000 vertices with rewiring
probability $p = 0.05$

navigation in a small world



Jon Kleinberg

how to find short paths using only local information?

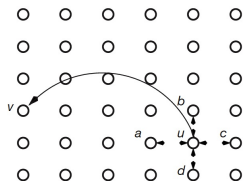
- we will use a simple directed model [Kleinberg, 2000]
- a local algorithm
 - can remember the source, the destination and its current location
 - can query the graph to find the long-distance edge at the current location

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 by 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

forest-fire model



J. Leskovec



J. Kleinberg



C. Faloutsos

[Leskovec et al., 2007] propose the forest fire model that is able to re-produce at a qualitative scale most of the established properties of real-world networks

forest-fire model

the forest-fire model is able to explain

- heavy tailed in-degrees and out-degrees
- densification power law
- shrinking diameter
- ...
- deep cuts at small size scales and the absence of deep cuts at large size scales

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*.