Random Graphs and Complex Networks
T-79.7003

Charalampos E. Tsourakakis

Aalto University

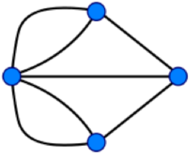
Class website

http://www.math.cmu.edu/~ctsourak/t797003-graphs-and-networks.html

introduction to graphs and networks

graphs: a simple model

- entities – set of vertices
- pairwise relations among vertices
  – set of edges
- can add directions, weights,...
- graphs can be used to model many real datasets
  - people who are friends
  - computers that are interconnected
  - web pages that point to each other
  - proteins that interact

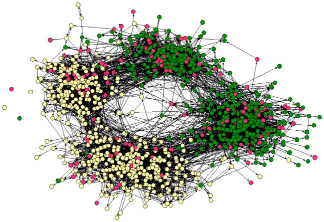


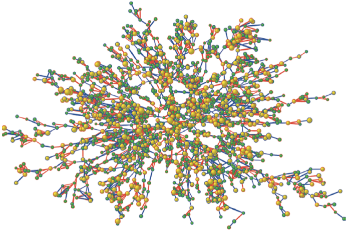
graph theory

- graph theory started in the 18th century, with Leonhard Euler
  - the problem of Königsberg bridges
  - since then, graphs have been studied extensively

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 a lot of information

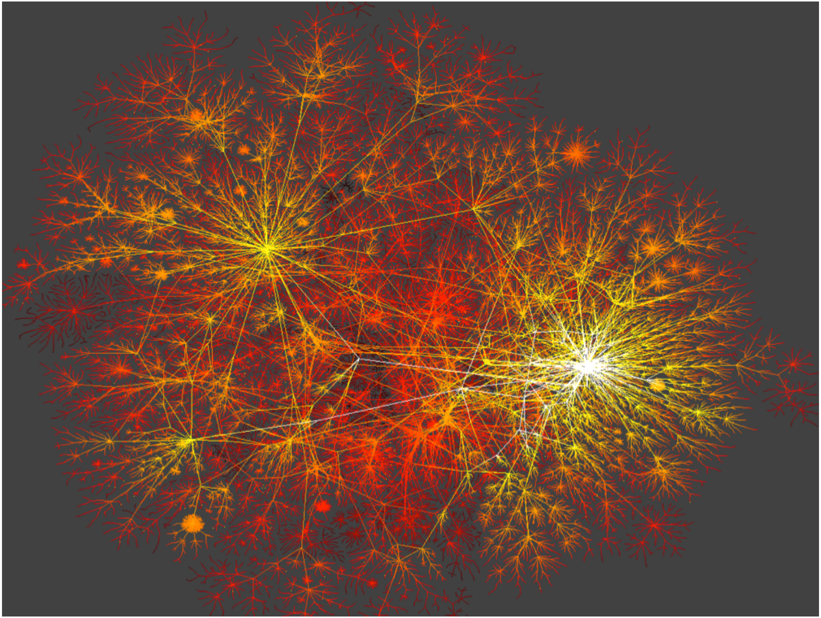




analysis of graph datasets now

- more and larger networks appear
  - products of technological advancement
    - e.g., internet, web
  - result of our ability to collect more, better-quality, and more complex data
    - e.g., gene regulatory networks
- networks of thousands, millions, or billions of nodes
  - impossible to visualize

the internet map

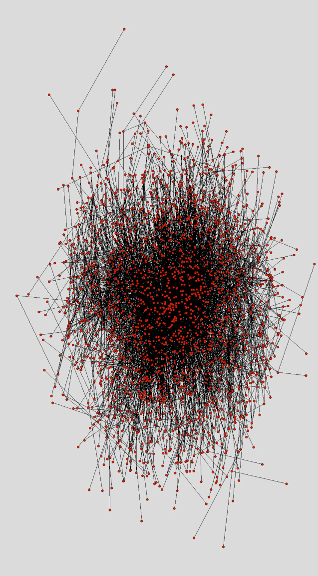


types of networks

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

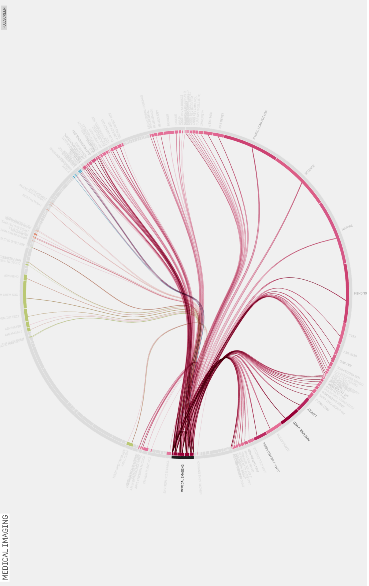
social networks

- 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



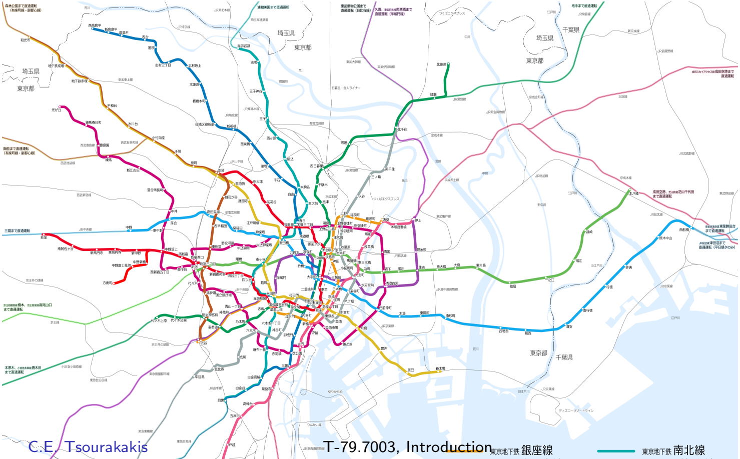
knowledge and information networks

- 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

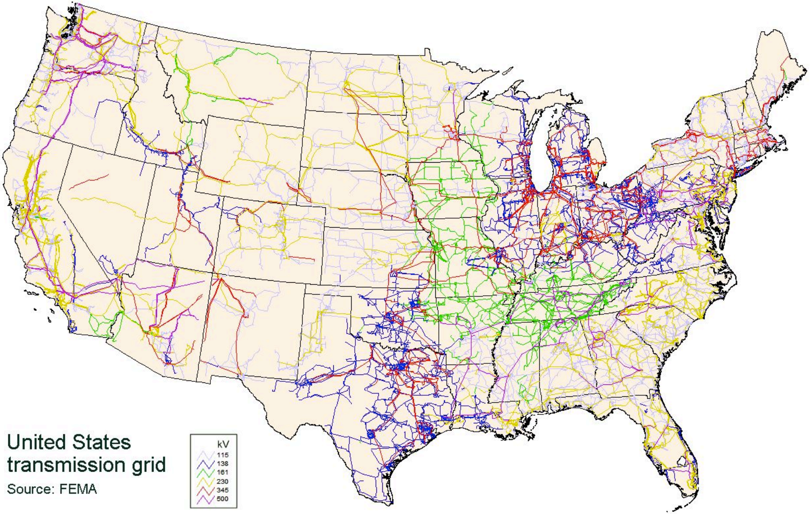


technological networks

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

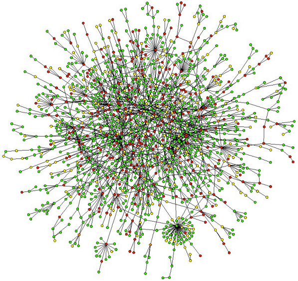


US power grid



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



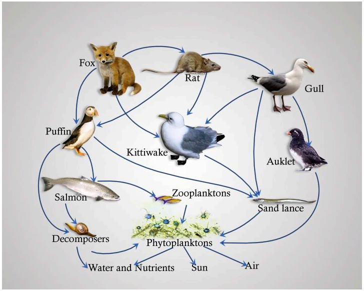
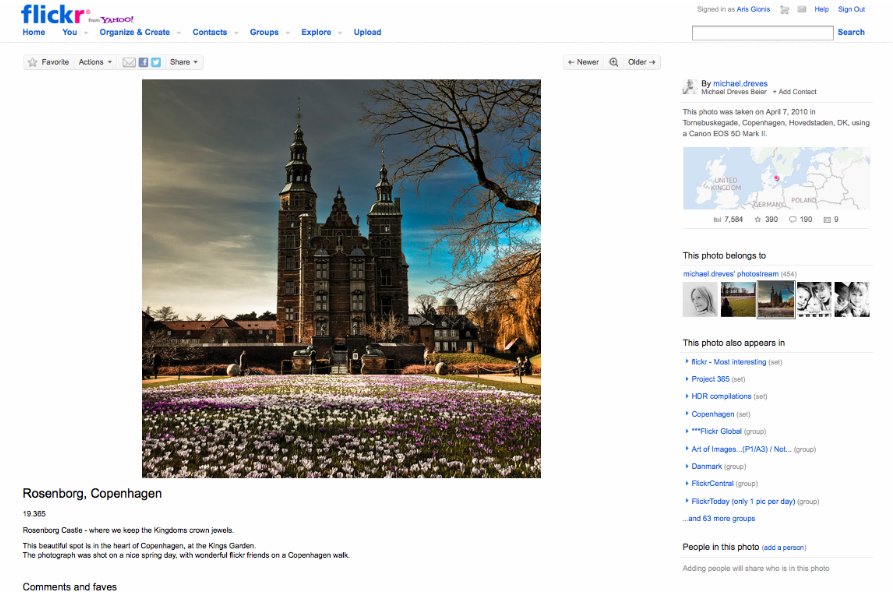


photo-sharing site



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, ...
- also many interesting induced graphs
  - tag graph: based on photos
  - tag graph: based on users
  - user graph: based on favorites
  - user graph: based on groups
- which graph to pick — not an easy choice

recurring theme

- social media, user-generated content
- user interaction is composed by many atomic actions
  - post, comment, like, mark, join, comment, fave, thumps-up, ...
  - generates all kind of interesting graphs to mine

now what?

- 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 that make use of the network structure

concepts to study

- paths and connectivity
- path lengths and diameter — small-world phenomena
- degree distributions and degree correlations
- communities and clusters
- flows and cuts
- processes
    - evolution, random walks, information cascades, epidemics, ...

What we will see in this class?

- stochastic graph models
  - Erdös-Rényi random graphs
  - Preferential attachment
  - Small world networks
- graph partitioning
  - Cheeger's inequality
  - Spectral partitioning
  - Every graph is essentially sparse

Class projects

A list with a wide range of topics will be suggested to you in the 3rd lecture. Some keywords and few words for now.

- Random graphs and models of real-world networks
  - Erdös-Rényi
  - scale-free
  - highly optimized tolerance (HOT)
  - random geometric graphs etc
  - strategic network formation
  - ...
- Random processes on static graphs (cascades, rumor spreading, Moran process etc.)
- Evolving graphs
- Algorithmic issues related to graph partitioning
  - Solving Laplacians in nearly linear time
  - Max flows
  - Graph sparsification
  - Learning and graphs

Administrivia

Grading

- 3 homeworks, 15%
- 2 exams, 30%+30%
- Project, 20%

Homework policy

- You may discuss the problems with other students but you must write your solutions on your own and list your collaborators.
- You may not search the Web for solutions.
- You may consult outside materials, but you must cite your sources in your solutions.

Administrivia

Textbook

- There is no required textbook.
- A suggested set of books available online and notes is available in the class Web page. I will provide you with lecture notes and slides. If you are interested in learning more, you are urged to read the references therein.

Administrivia

Projects will be consist of a project report and a presentation in class.

- Students who work on theoretical computer science/discrete mathematics are welcome to read a paper and present it. In case that after you choose your favorite paper you find it hard to understand, I will be happy to discuss it with you beforehand.
- Students who do research in data mining are welcome to conduct an experimental project. For the latter type of projects, collaboration in groups of two is welcomed. If you want to collaborate with more than one person, please contact me first.

empirical properties of graphs and networks

Properties of real-world networks

diverse collections of graphs arising from different phenomena

are there typical patterns?

- static networks
  1. heavy tails
  2. clustering coefficients
  3. communities
  4. small diameters
- time-evolving networks
  1. densification
  2. shrinking diameters
- web graph
  1. bow-tie structure
  2. bipartite cliques

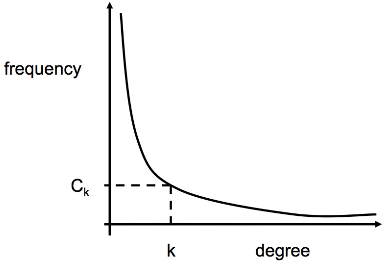
Heavy tails

> *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

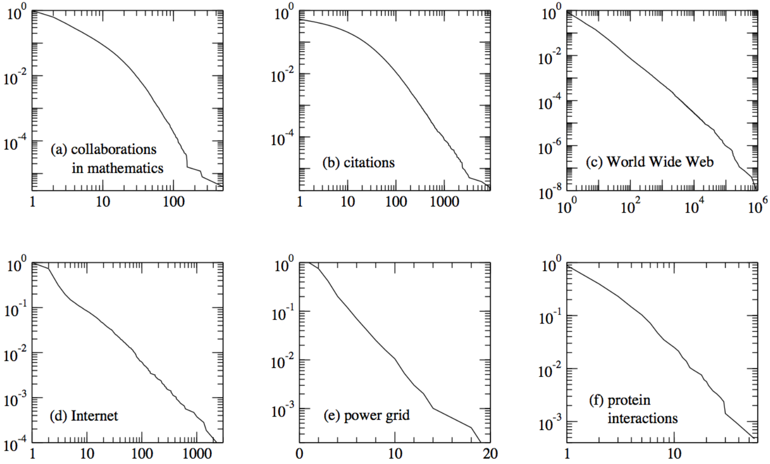
$$C_k = ck^{-\gamma}$$

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

- frequently, we hear about "scale-free networks"
  correct term is networks with scale-free degree distribution

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

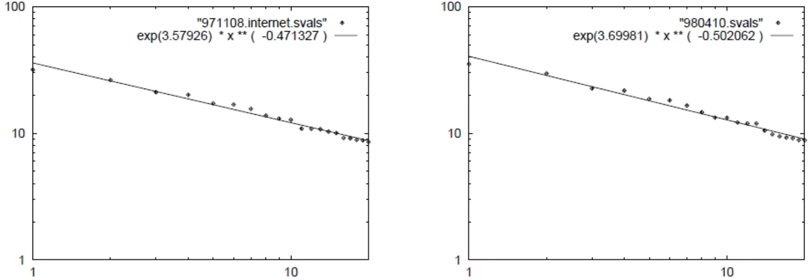
Maximum degree

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

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

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

Heavy tails, eigenvalues

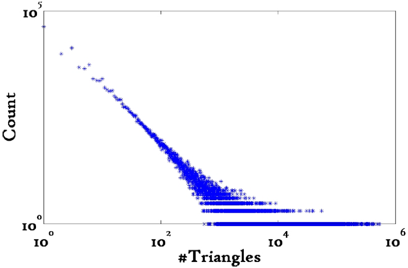


(a) Int-11-97 (b) Int-04-98

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 the 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

Clustering coefficients

- alternative definition
- local clustering coefficient

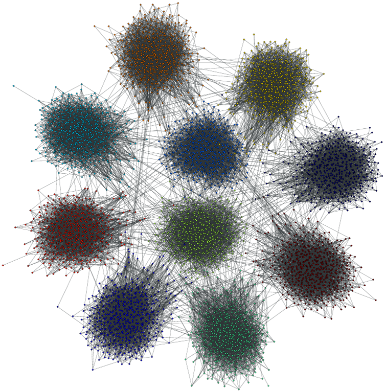
$$C_i = \frac{\text{Number of triangles connected to vertex } i}{\text{Number of triples centered at vertex } i}$$

- global clustering coefficient

$$C(G) = \frac{1}{n} \sum_i C_i$$

Community structure

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



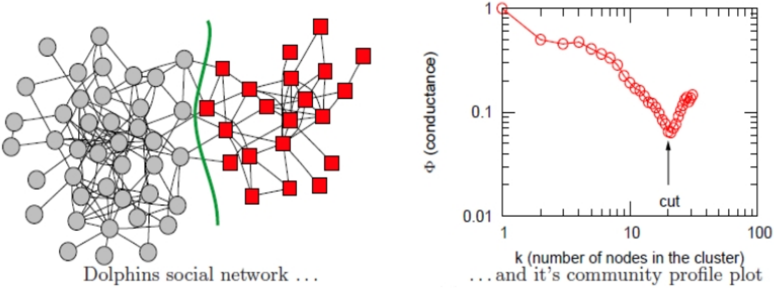
artificial communities:
http://projects.skewed.de/graph-tool/

Community structure

[Leskovec et al., 2009]

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

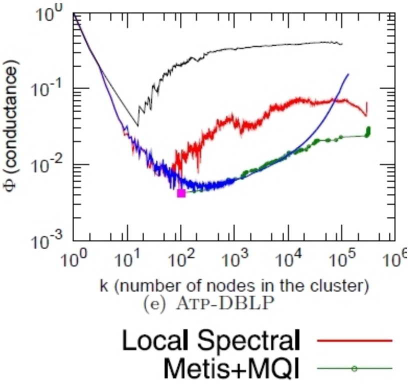
Community structure



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

Community structure

- do large-scale real-world networks have this nice artifical structure? NO!



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

Community structure

important findings of [Leskovec et al., 2009]

1. up to a certain size k ($k \sim 100$ vertices) there exist 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-World project: http://smallworld.columbia.edu/index.html

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].

Other properties

- assortativity
- distribution of size of connected components
- distribution of motifs
- ...

Time-evolving networks

J. Leskovec

J. Kleinberg

C. Faloutsos

[Leskovec et al., 2005]

- densification power law:

$$|E_t| \propto |V_t|^{\alpha} \qquad 1 \leq \alpha \leq 2$$

- shrinking diameters: diameter is shrinking over time.

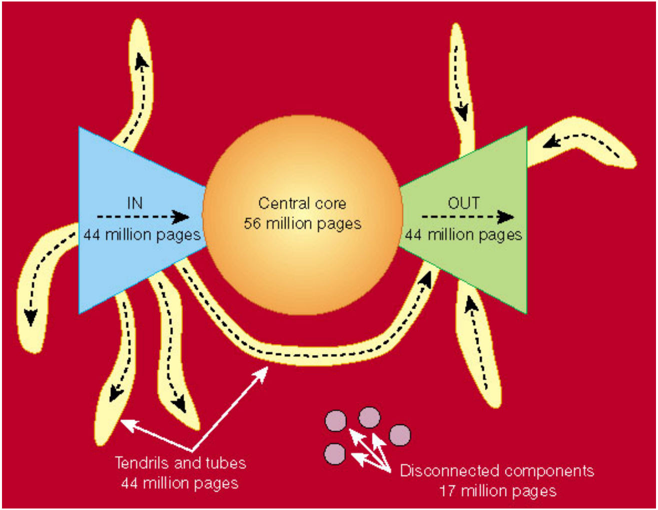
Web graph

- the Web graph is a particularly important real-world network

  > *Few events in the history of computing have wrought as profound an influence on society as the advent and growth of the World Wide Web*
  >
  > [Kleinberg et al., 1999]

- vertices correspond to static web pages
- directed edge (i, j) models a link from page i to page j
- will discuss two structural properties of the web graph:
  1. the bow-tie structure [Broder et al., 2000]
  2. abundance of bipartite cliques [Kleinberg et al., 1999, Kumar et al., 2000]

Web is a bow-tie

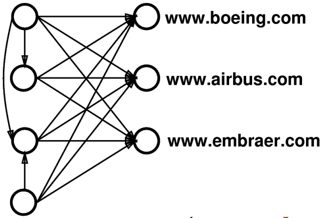


(source [Broder et al., 2000])

Bipartite subgraphs

- websites that are part of the same community frequently do not reference one another

  (competitive reasons, disagreements, ignorance) [Kumar et al., 1999].
- similar websites are *co-cited*
- therefore, web communities are characterized by dense directed bipartite subgraphs



(source [Kleinberg et al., 1999])

references I

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000).

Graph structure in the web: Experiments and models.

In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands. ACM Press.

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*.

references II

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999).

The web as a graph: Measurements, models, and methods.

In *Computing and combinatorics*, pages 1–17. Springer.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. (2000).

Stochastic models for the web graph.

In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 57–65, Redondo Beach, CA, USA. IEEE CS Press.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999).

Trawling the Web for emerging cyber-communities.

Computer Networks, 31(11–16):1481–1493.

references III

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.

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.

references IV

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.

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

The structure and function of complex networks.

references V

Tsourakakis, C. E. (2008).

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

In *ICDM*.