GRAPH AND SOCIAL NETWORK VISUALIZATION

Riccardo Di Bella

Dipartimento di Informatica

Università degli Studi di Milano

OUTLINE

Graph drawing

- Quality measures
- Conventions
- Crossing number
- Symmetry
- Slope number

Layout methods

- Force-directed layout
- Example 1: force-directed (gephi)
- Example 2: circular (gephi)

Networks analysis & random networks

- Empirical network features
- Power Law
- Bernoulli Random Graph
- Preferential Attachment Model

GRAPH DRAWING: QUALITY MEASURES

How can we evaluate the quality of a graph drawing, according to its aesthetic and usability? Many measures have been proposed, to name a few:

- Crossing number
- Symmetry
- Slope number
- Edge bends
- Angular resolution

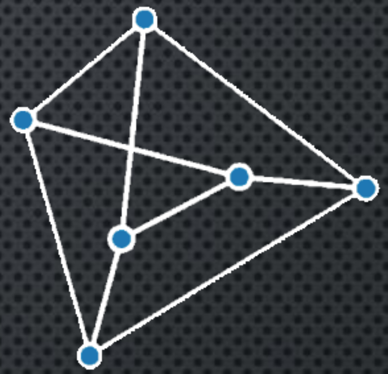
Graph drawing techniques and algorithms often try to optimize these measures. However, some of them are often conflicting, or it can be difficult algorithmically to deal with all of them at the same time.

GRAPH DRAWING: CONVENTIONS AND CONSTRAINTS

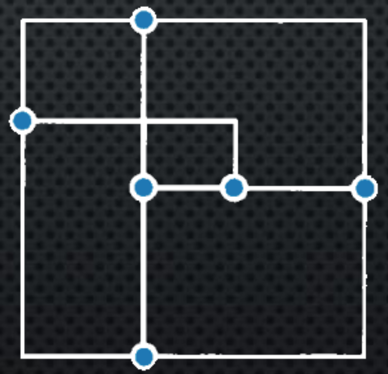
Additionally, the user can specify *conventions* and *constraints* for the drawing, which depend on the application domain. A convention is a basic rule that the drawing must satisfy. Widely used are:

- **Polyline drawing**: each edge is drawn as polygonal chain
- **Straight-line drawing**: each edge is drawn as a straight-line segment
- **Orthogonal drawing**: each edge is drawn as a polygonal chain of alternating horizontal or vertical segments
- **Grid drawing**: vertices, crossings and edge bends have integer coordinates
- **Planar drawing**: no two edges cross
- **Upward (downward) drawing**: for acyclic directed graphs, each edge is drawn as a curve monotonically increasing (decreasing).

Polyline



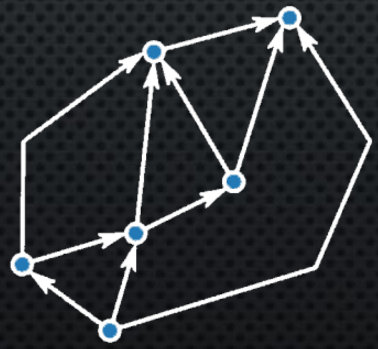
Straight-line



Orthogonal

Grid

Planar polyline



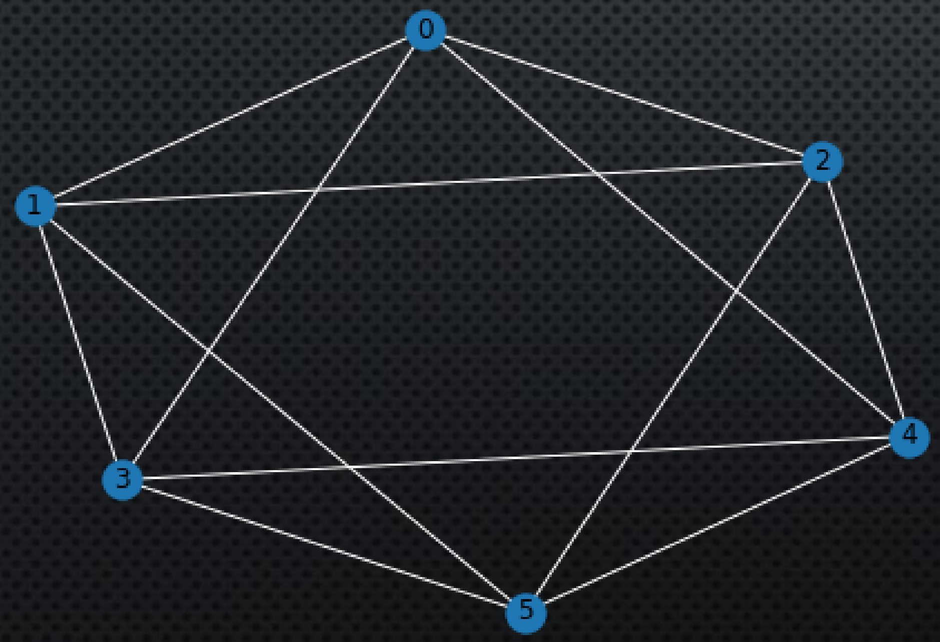
Upward planar polyline

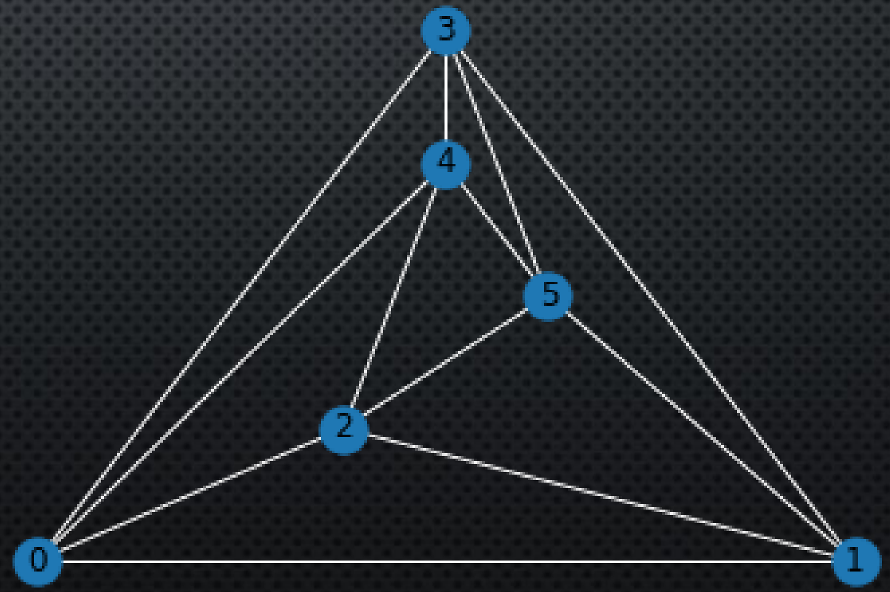
CROSSING NUMBER

Definition

The crossing number cr(G) of a graph G is the lowest number of edge crossings of a plane drawing of the graph G. A graph is ***planar*** *if its crossing number is zero, and thus is possible to draw it so that none of the edges cross.*

Determining the crossing number is important in graph drawing, since user studies have shown that drawings with few crossings are easier to understand.





Non-planar and planar drawing of the same (planar) graph

EULER'S FORMULA FOR PLANAR GRAPHS

When a planar graph is drawn without edges crossing it divides the plane into regions, which we call **faces**. Euler has shown that there's a relationship between the number of vertices (v), the number of edges (e) and the number of faces (f):

Euler's Formula: $v - e + f = 2$

Using Euler's Formula and other simple properties of planar graphs we can easily show that any planar graph with v vertices and e edges satisfies:

$$e \leq 3v - 6$$

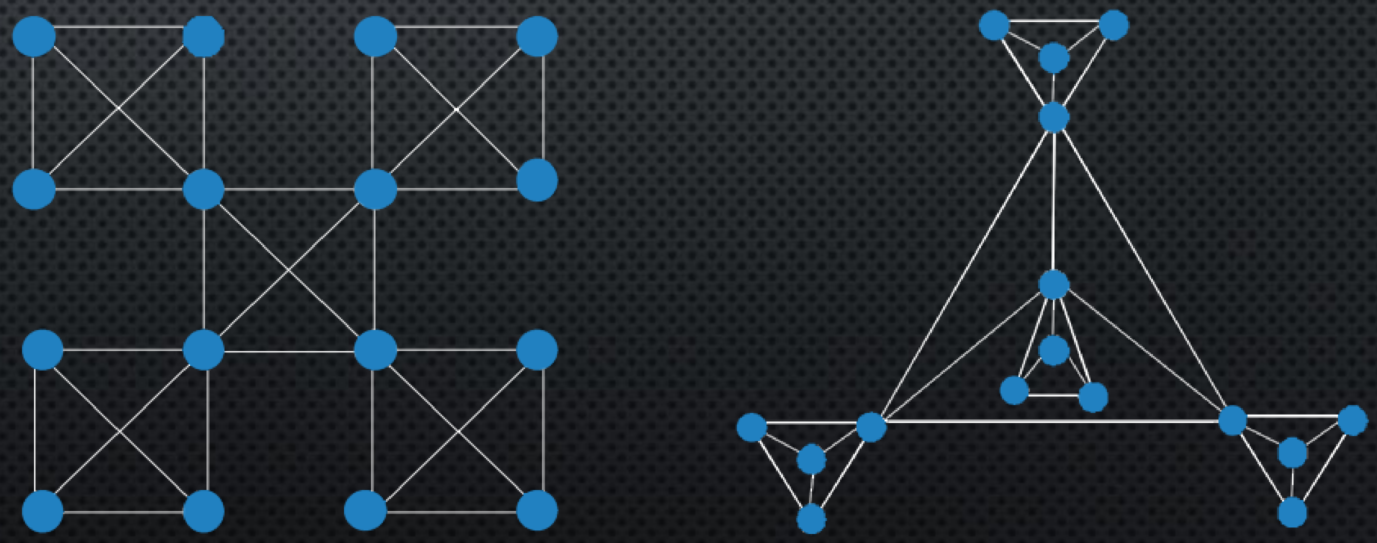
As can be inferred from this, *dense* graphs (in which the number of edges is $\theta(v^2)$) are usually not planar.

SYMMETRY

Symmetry is another important criteria, both for aesthetic and to reveal structural properties of the graph. Ideally, we would like our drawing to exhibit as much symmetry as possible.

Definition

Formally, a symmetry corresponds to an ***isomorphism on the graph****: a permutation* σ *on the vertex set* V*, such that the pair of vertices* (u, v) *forms an edge if and only if the pair* $\big(\sigma(u), \sigma(v)\big)$ *also forms an edge.*

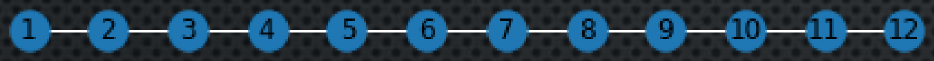


Two drawings of the same graph.
The left one exhibits a lot of symmetries,
the right one has only the axial symmetry
but has no edge crossing.

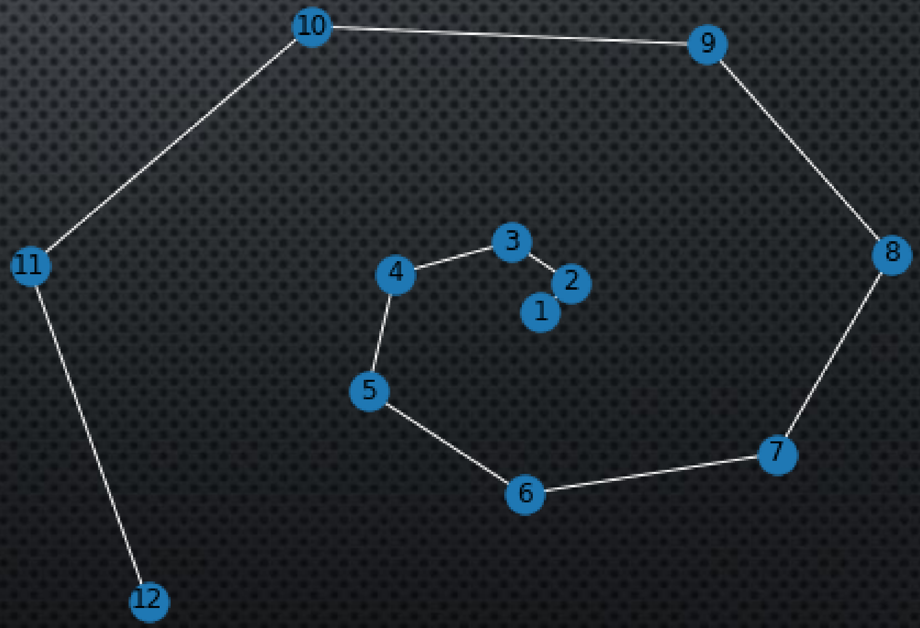
SLOPE NUMBER

Definition

The slope number of a graph is the ***minimum possible number of distinct slopes*** *of edges in a drawing of the graph (in which vertices are represented as points and edges as line segments).*



1 slope



11 different slopes

LAYOUT METHODS

Most commonly used layout techniques:

- **Force-directed layout**: modifies an initial placement of vertices by simulating a system of combined attractive and repulsive forces

- **Spectral layout**: uses eigenvectors of a matrix associated with the graph (usually the *adjacency* or the *Laplacian* matrix) to determine the vertices coordinates

- **Circular layout**: the vertices are partitioned into *clusters* and then placed onto one or multiple circles, choosing the order carefully to minimize edge crossings

FORCE-DIRECTED LAYOUT

The main idea behind these algorithms is to assign forces among the set of nodes and edges of a graph drawing. **Spring-like forces are used to attract pairs of adjacent vertices**, while simultaneously **repulsive forces are used to separate them**. Once the forces have been assigned, the behaviour of the entire graph can be simulated as if it were physical system. The algorithm is repeated until the system comes to a state of equilibrium.

Being physical simulations, force-directed layout methods do not rely on domain-specific knowledge or concepts related to graph theory. Despite this, graph drawn with these algorithms tend to be aesthetically pleasing, exhibit symmetries and produce few edge crossings.

A SIMPLE FORCE-DIRECTED LAYOUT ALGORITHM (EADES, 1984)

- Repulsive force: $$f_{rep}(p_u, p_v) = \frac{c_{rep}}{\|p_u - p_v\|^2} \overrightarrow{p_u p_v}$$

- Attractive force: $$f_{spring}(p_u, p_v) = c_{spring} \log \frac{\|p_u - p_v\|}{l} \overrightarrow{p_u p_v}$$

$$\textit{where } l \textit{ is the ideal edge length}$$

- Displacement vector for node v:

$$F_v = \sum_{u:(u,v) \notin E} f_{rep}(p_u, p_v) + \sum_{u:(u,v) \in E} f_{spring}(p_u, p_v)$$

A SIMPLE FORCE-DIRECTED LAYOUT ALGORITHM (EADES, 1984)

Input: A connected undirected graph $G = (V, E)$ with initial placement

$p = (p_v)_{v \in V}$, number of iterations $k \in \mathbb{N}$ threshold $\varepsilon > 0$, constant $\delta > 0$

Output: A new layout p with «low internal stress»

$t \leftarrow 0$;

while $t < k$ and $max_{v \in V} \|F_v(t)\| > \varepsilon$ do:

  for $v \in V$ do:

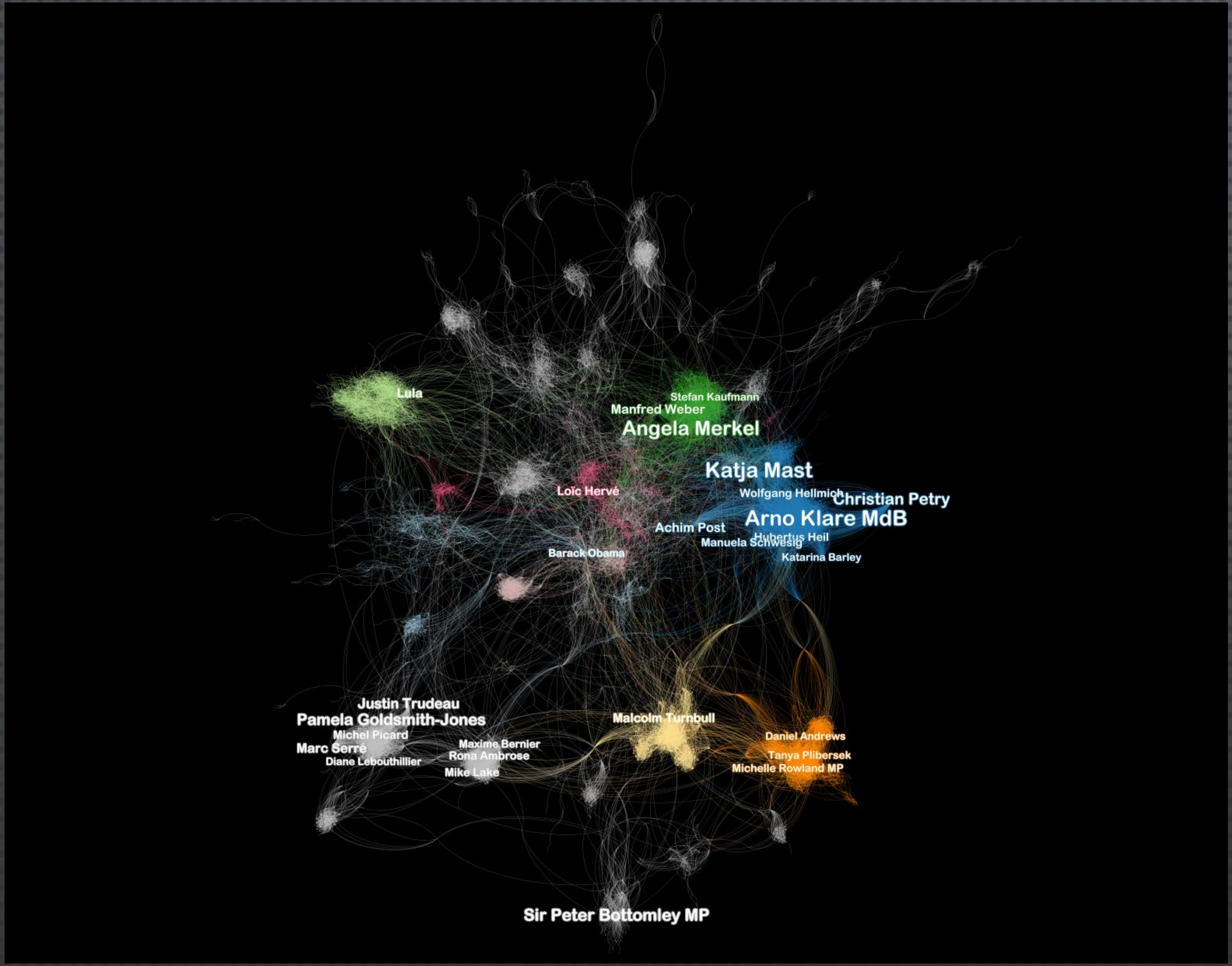
    $F_v(t) \leftarrow \sum_{u:(u,v) \notin E} f_{rep}(p_u, p_v) + \sum_{u:(u,v) \in E} f_{spring}(p_u, p_v);$

    $p_v \leftarrow p_v + \delta \cdot F_v(t);$

  end

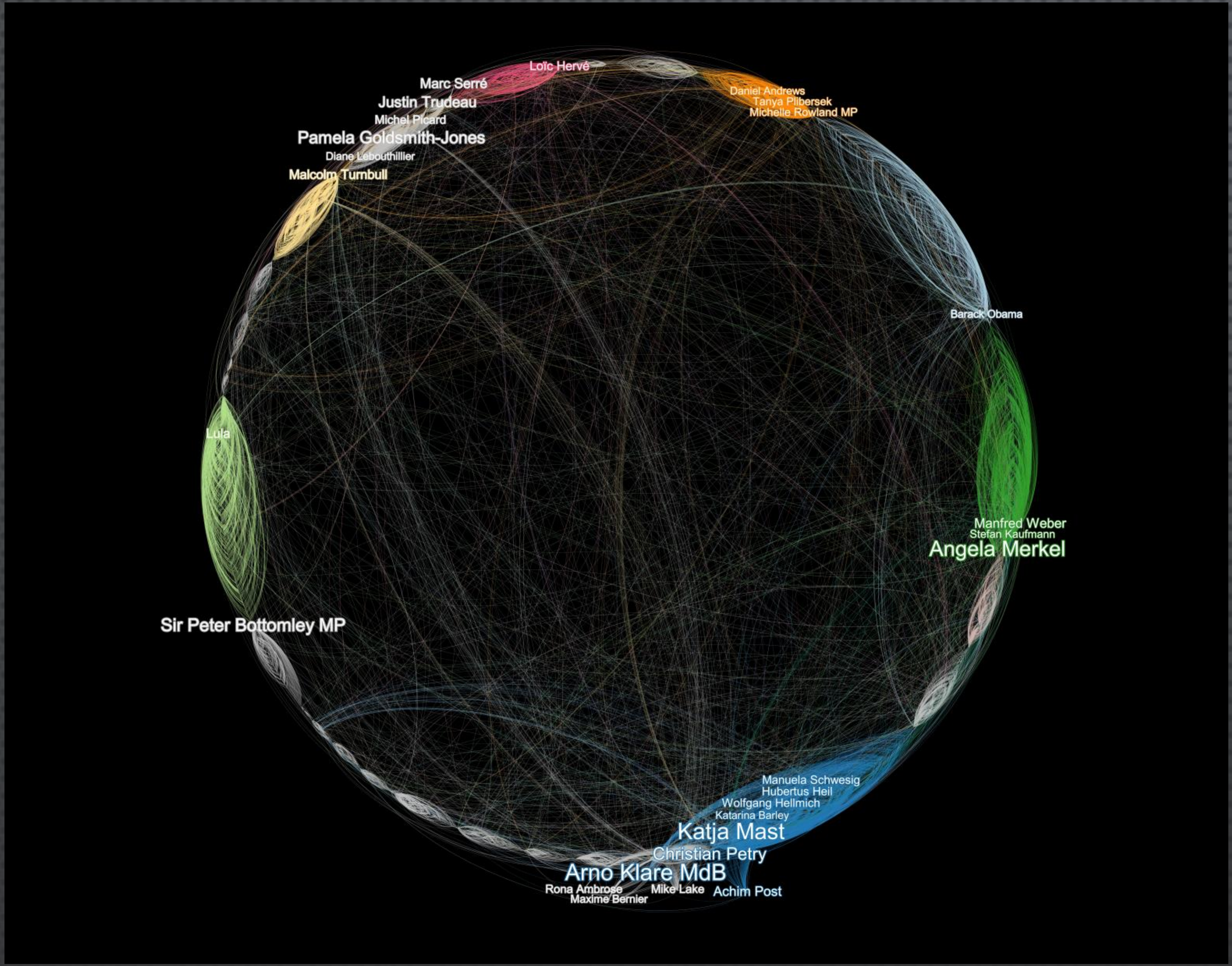
end

EXAMPLE 1: FORCE-DIRECTED (GEPHI – FORCE ATLAS 2)



*Example of a layout generated with Force-Atlas 2, an algorithm provided by the open source graph visualization tool **Gephi**. The graph represents "mutual friendship" between politicians on Facebook (dataset courtesy of Network Repository)*

EXAMPLE 2: CIRCULAR (GEPHI)



The same graph represented using a circular layout algorithm provided by Gephi. Nodes have been grouped "explicitly" according to their suppoed communities

NETWORK ANALYSIS WITH RANDOM NETWORK MODELS

One way to study networks is by creating random models which try to capture the features and properties of real networks.

Additionally, this helps us solve problems like:

- How are networks formed? Why do they have certain properties?
- How similar are networks?
- Inference of missing links or, conversely, checking if existing links are not false positives
- Predicting networks

EMPIRICAL NETWORK FEATURES

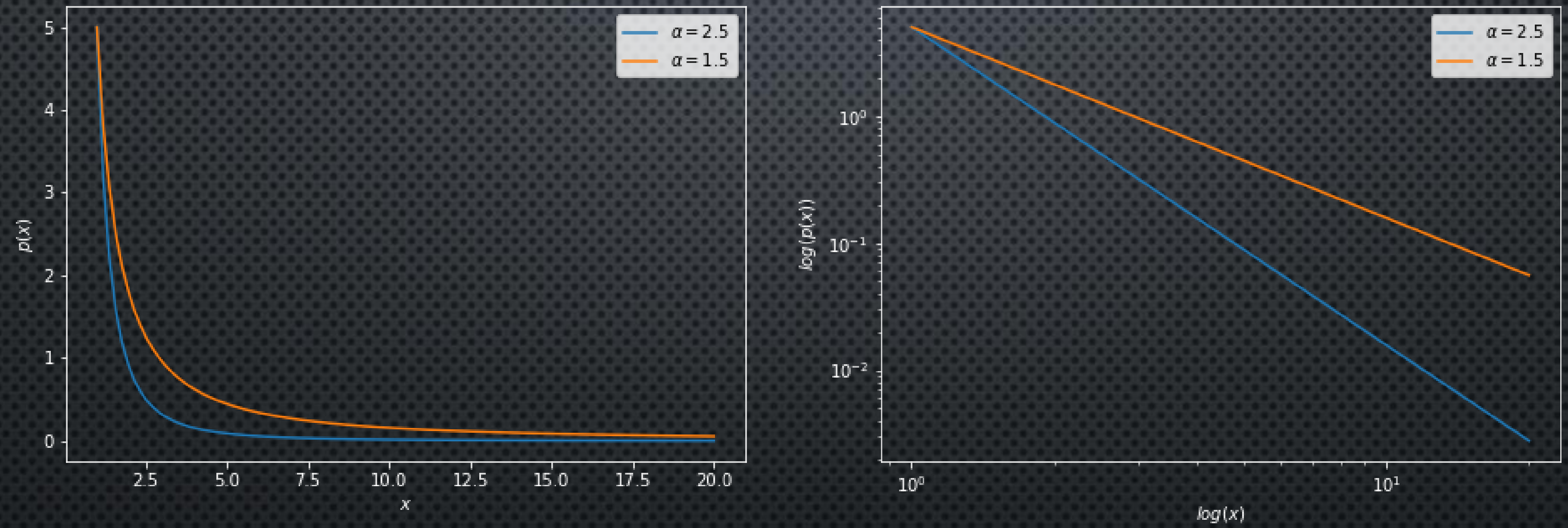
The main features that are found in real social networks are:

- Power-law (heavy tailed) degree distribution
- Small-world phenomenon: small average distance and graph diameter (e.g. “six degrees of separation”)
- High clustering coefficient (a lot of “triangles”)
- Giant connected component

POWER LAW AND SCALE-FREE NETWORKS

A lot of empirical data related to social networks follows approximately a *power law*. In general, by power law we mean a function of the form:

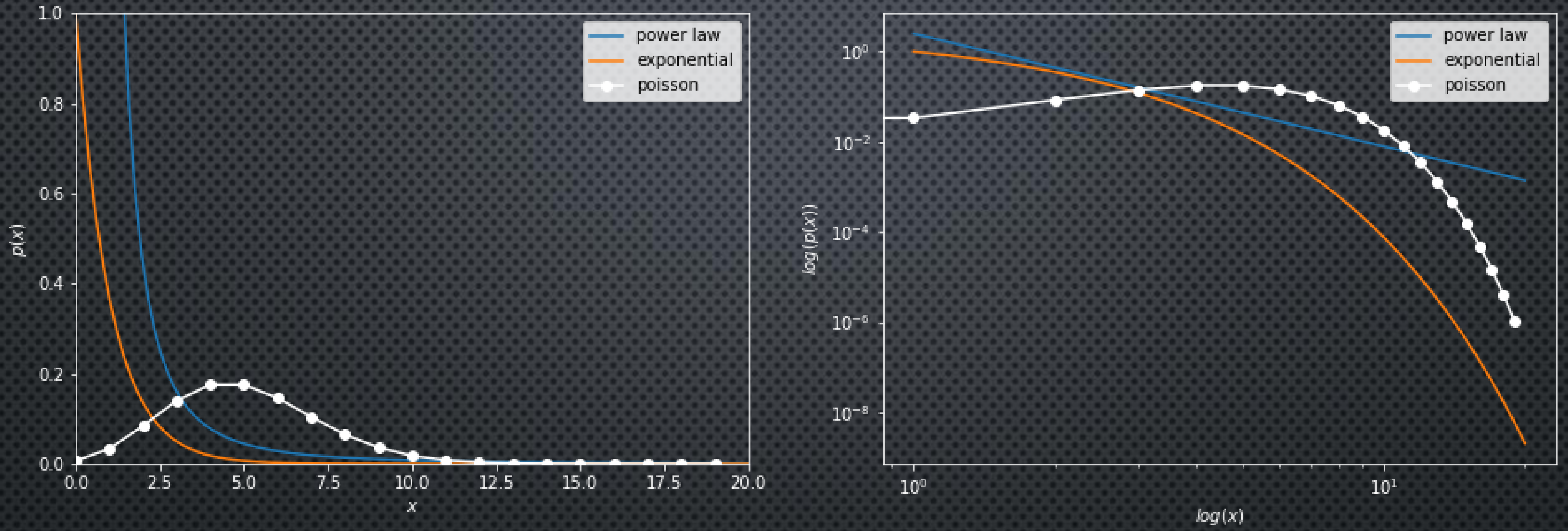
$$p(x) = Cx^{-\alpha} \qquad x \geq x_{min}$$



This function has some interesting properties:

- It's *scale-free*
- *Heavy-tailed*
- Linear in log-log scale

POWER LAW COMPARED



- When compared to exponential or Poisson distribution, power law decays much slower. This behavior is called *fat-tail*
- For small x, note how power law lays above the Poisson. When dealing with networks, this means that a *scale-free network* will have more nodes with low degree, and also more nodes with high degree.

BERNOULLI RANDOM GRAPH (ERDŐS–RÉNYI MODEL)

The most standard model is the one proposed by Erdős and Rényi in 1959.

The parameters are n (the number of nodes) and p. Any edge between two nodes is present with probability p, independently of other edges.

The number of potential edges is

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

Thus the number of expected edges is

$$E(m) = p\binom{n}{2}$$

The probability that a given node has degree k is

$$P(K_i) = P(K) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

The expected degree is $E(K) = (n-1)p$

BERNOULLI RANDOM GRAPH (ERDŐS–RÉNYI MODEL)

The average number of triangles is

$$\binom{n}{3} p^3 = \frac{n(n-1)(n-2)}{6} p^3$$

Similarly, the average number of 2-stars is:

$$\binom{n}{3} p^2$$

And we can expect an average clustering coefficient of

$$\frac{\binom{n}{3} p^3}{\binom{n}{3} p^2} = p$$

The probability of your friends being friend themselves is the same probability of two random strangers being friend!

We see the Bernoulli model doesn't properly capture an important feature of scale-free networks, the high clustering coefficient. Another problem is the degree distribution function:

$$P(K = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \text{with fixed } \lambda = np \text{ and } n \to \infty$$

PREFERENTIAL ATTACHMENT MODEL (BARABASI-ALBERT 1999)

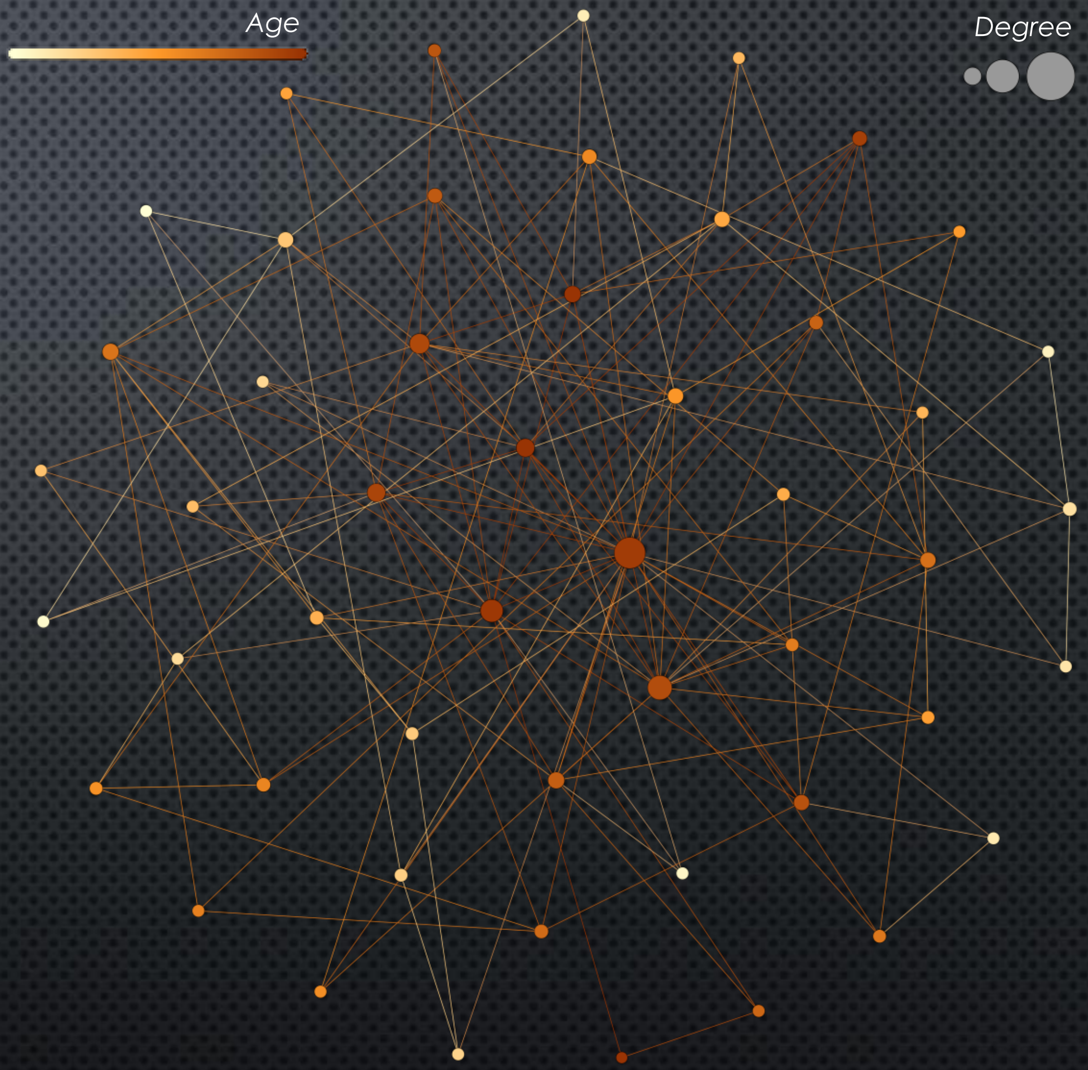
This model has two important differences from the E-R model:

- Dynamical growth
- Attachment probability is proportional to degree: "*rich get richer*"

$$p_i^l(k_i) = \frac{k_i}{\sum_i k_i}$$

We start with n_0 nodes, at every discrete time t we add a new node. Every node is born with a degree m, with m links attached to already existing nodes according to the probability stated above.

We use $k_i(t)$ to indicate the average degree of a node born at time i when t instants have passed.



A simple Barabasi-Albert random graph
N=50, m=3

PREFERENTIAL ATTACHMENT MODEL (BARABASI-ALBERT 1999)

We find that

$$k_i(t) = m\sqrt{\frac{t}{i}} \qquad t \geq i$$

We can plot this as a function of t (time passing), or we can fix t and plot it as a function of i (birthdate of the node). The latter is more useful for our purposes: we want to find what nodes have degree lower than a certain k.

After fixing t, we can solve for i and get

$$i > \frac{m^2}{k^2}t$$

The fraction of nodes with (average) degree lower than k is

$$F_k = P(K_i(t) < k) = \frac{n_0 + t - i}{n_0 + t} = \frac{n_0 + t - \frac{m^2}{k^2}t}{n_0 + t} \sim 1 - \frac{m^2}{k^2} \text{ for } t \gg n_0$$

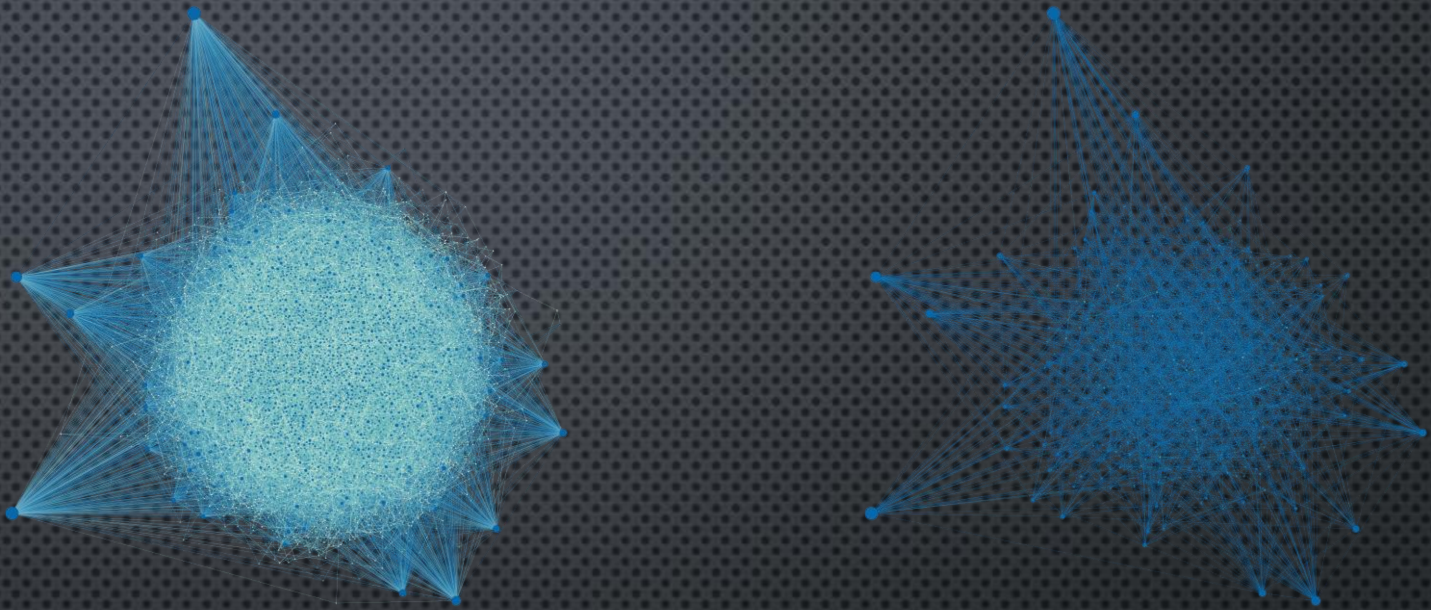
PREFERENTIAL ATTACHMENT MODEL (BARABASI-ALBERT 1999)

Differentiating with respect to k we find the probability function

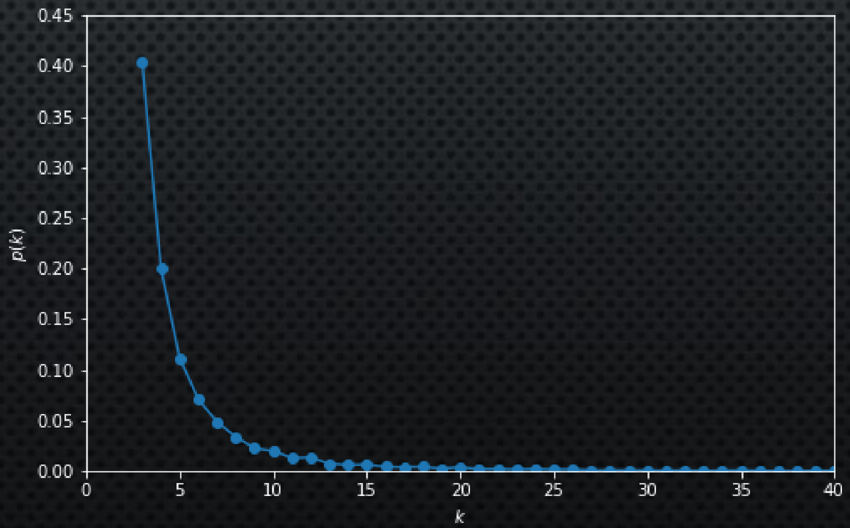
$$p(k) = \frac{d}{dx}F(k) = \frac{2m^2}{k^3}$$

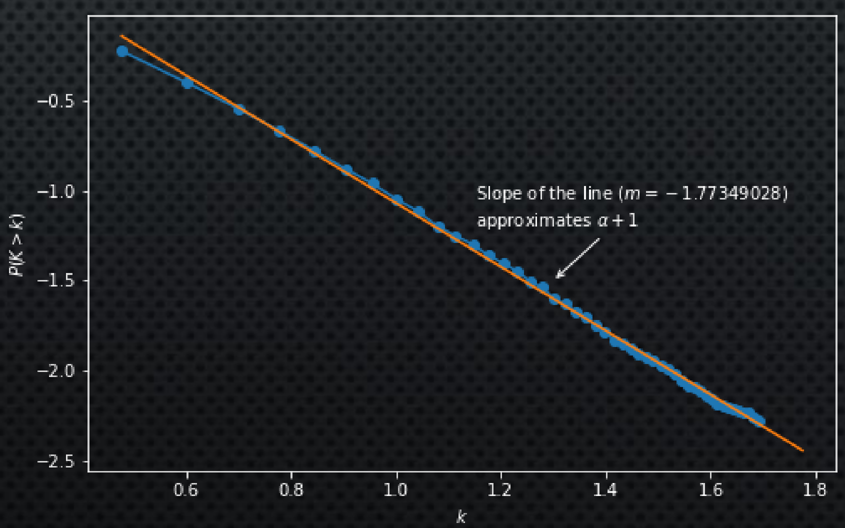
We see the degree distribution of this model follows a power law!

It can also be shown that, in general, the *small-world* property is well captured by this model. However, the clustering coefficient is not.



Another Barabasi-Albert graph, with N=10000 and m=3. On the right we kept only the nodes with degree > 14





REFERENCES AND ADDITIONAL RESOURCES

- “Graph drawing: Algorithms for the Visualization of Graphs” G. Di Battista, P. Eades, R. Tamassia, I. G. Tollis (Pearson)
- Purchase, E. (1997) "Which aesthetic has the greatest effect on human understanding?" In G. Di Battista, editor, Graph Drawing (Proc. GD ’97), volume 1353 of Lecture Notes Comput. Sci., pages 248–261. Springer-Verlag. Giant connected component
- “Handbook of Graph Drawing and Visualization”
- Barabasi, A.; Albert. R (1999). "Emergence of scaling in random networks“ . Science 286, 509 (1999)
- “Introduction to SNA” (videoclass by Lehonid Zukov) and “Social and Economic Networks” (videoclass by Matthew O. Jackson)
- Gephi
- NetworkX
- Network Repository