Modelling self-organizing networks with a hidden metric

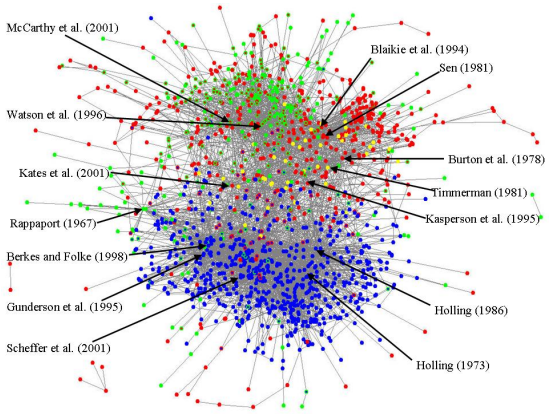
Jeannette Janssen

Dalhousie University

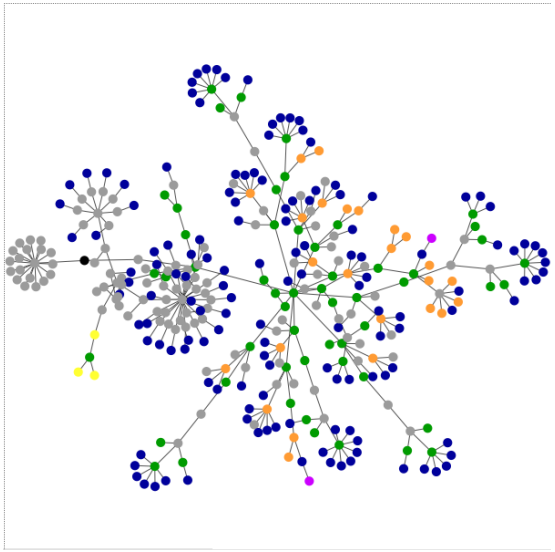
joint work with
W. Aiello, A. Bonato, C. Cooper, P. Pralat

Self-organizing networks

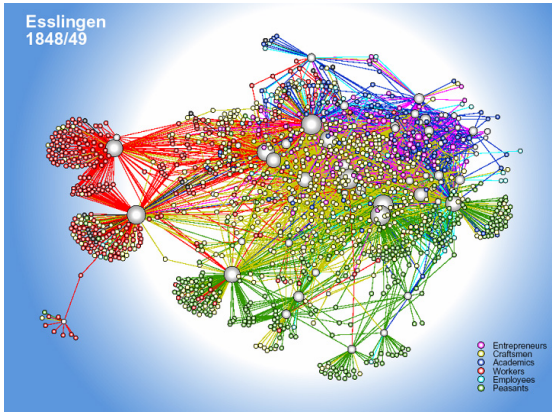
Networks created by the interaction of many autonomous agents



Citation network



World Wide Web



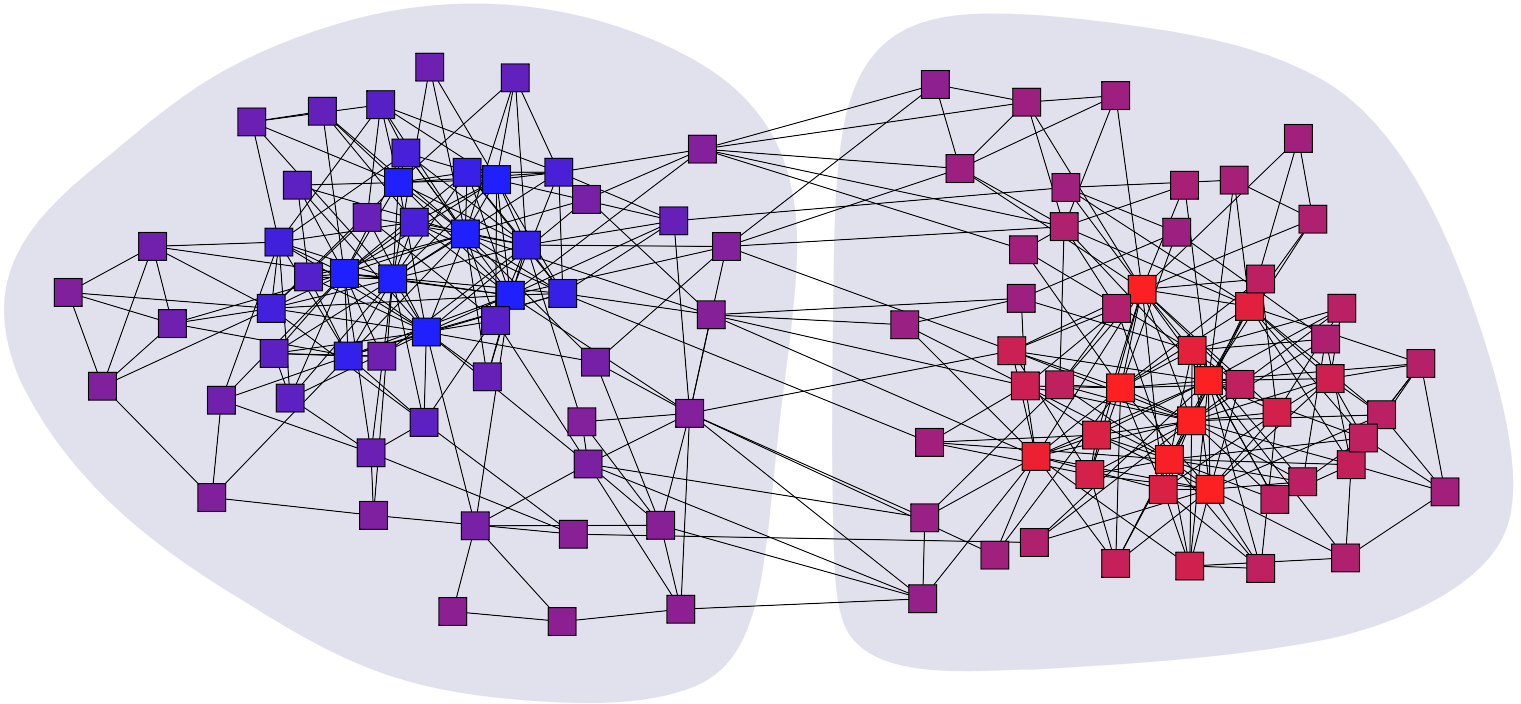
Social networks

A good graph model should...

1. ...reproduce experimentally observed graph properties:
   - Degree distribution follows a power law
   - Small average distance between nodes, ("Small world")
   - Locally dense, globally sparse.
   - Expansion properties (conductance)
   - Others..
2. ...include a credible model for agent behaviour guiding the formation of the link structure
3. ...agents should not need global knowledge of the network to determine their link environment

Common assumptions in the study of real-life networks

Communities in a social network can be recognized as densely linked subgraphs.



Newman, 2006

Common assumptions

Web pages with many common neighbours contain related topics.

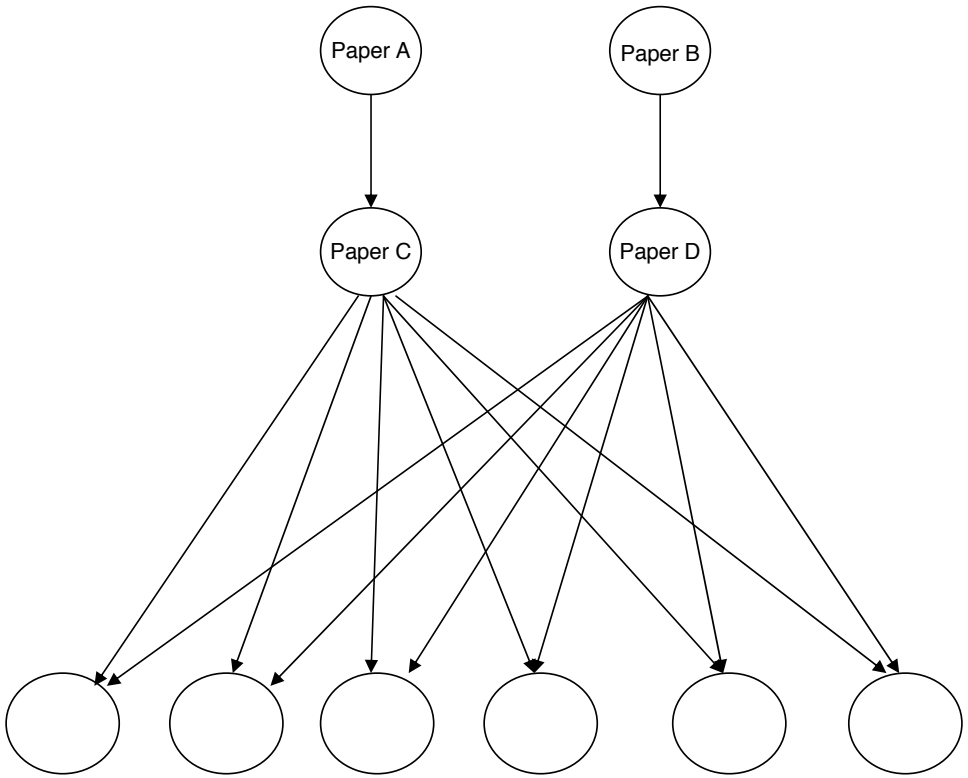


Fig. 1 Relation between papers inferred from citation graph

Underlying metric

Such assumptions, commonly used in experimental and heuristic treatments of real-life networks, imply that there is an a priori "community structure" or "relatedness measure" of the nodes, which is reflected by the link structure of the graph.

The network is a visible manifestation of an underlying hidden reality.

Spatial graph models

- Nodes correspond to points in a (high-dimensional) feature space
- The metric distance between nodes is a measure of "closeness"
- The edge generation is influenced by the position and relative distance of the nodes

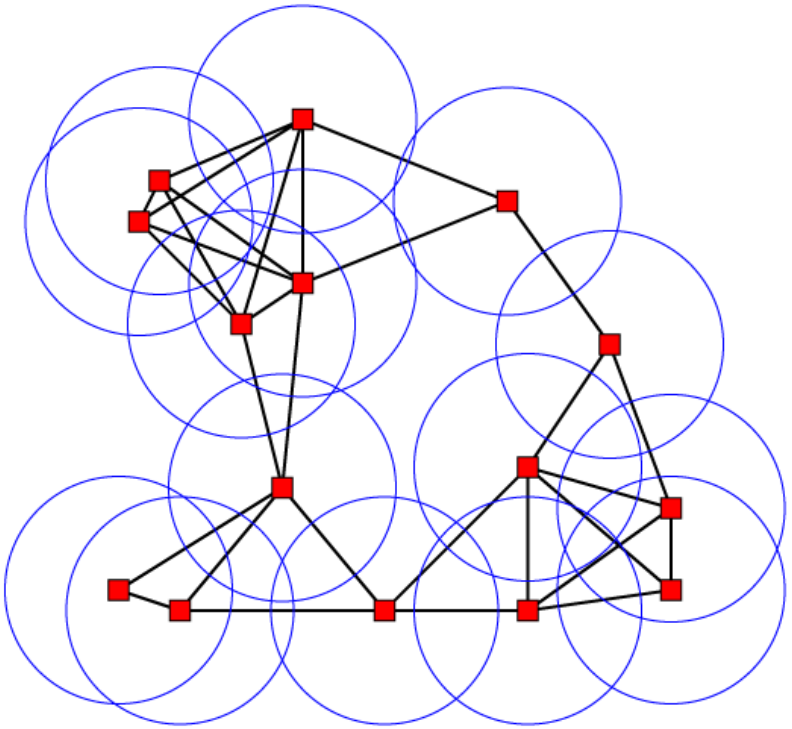
This gives a basis for reverse engineering: given a graph, and assuming a spatial model, it is possible to estimate the distribution of nodes in the feature space from information contained in the graph structure.

Random geometric graphs

[see book by Penrose '03]

n node points are randomly distributed in Euclidean space according to a given distribution.

Node points are joined by an edge if and only if their distance is less than a threshold value t.



Link neighbourhood based on local environment, but graph properties do not match.

Digression: Rank-based attachment

Preferential attachment:
Linking probability of a node v is proportional to $deg(v)$.

> Is it possible to modify the exponent of the power law by varying the linking probability to $(deg(v))^{\alpha}$?

- **No.** Only if $\alpha = 1$ does the process lead to a power law graph.

  Krapivsky, Redner, Leyvraz, 2000

- Instead, rank all the vertices. The linking probability of a node v is proportional to $R(v)^{-\alpha}$, where $R(v)$ is the rank of v.

  Fortunato, Flammini, Menczer, 2006

Rank-based attachment

Model:

- Parameters: initial degree $d \in \mathbb{N}$, linking probability coefficient $\alpha \in (0, 1)$.
- At each time t, node v has rank $R(v, t) \in [1, \ldots, t]$. Note that $R(\cdot, t)$ is a bijection.
- G_1 consists of a node v_1 with rank 1, and d loops.
- Add each time step t, a new node v_{t+1} is added, together with d edges with endpoint v_{t+1}. The other endpoints are chosen so that the probability that v_{t+1} links to a node u is proportional to $R(u, t)^{-\alpha}$.
- A rank is assigned to v_{t+1}, and the ranks of the existing nodes are updated, according to a pre-defined ranking scheme.

Ranking schemes

1 Each new node v_t is given a prestige label ℓ_t chosen randomly from $[0, 1]$ The ranking is based on the order of the labels.

Fortunato, Flammini, Menczer '06

$$R(v_i, t) \sim \ell_i t.$$

2 Ranking by age. Special case of *Protean graphs*

Pralat, Wormald '07

$$R(v_i, t) = i.$$

3 Random ranking. The new node v_t is given a rank r_t chosen u.a.r. from $[1, \ldots, t]$.

$$R(v_i, t) \sim (r_i/i)t$$

All these ranking schemes give a power law degree distribution with exponent $1 + 1/\alpha$.

Ranking schemes

- Novelty ranking: youngest node gets rank 1.

$$R(v_i, t) = t - i + 1.$$

Degree distribution is exponential.

- Ranking by degree: nodes are ranked in order of decreasing degree, secondary criterium is age.

  Not possible to predict $R(v, t)$.

Power law with exponent $1 + 1/\alpha$.

Geometric Preferential Attachment, Model

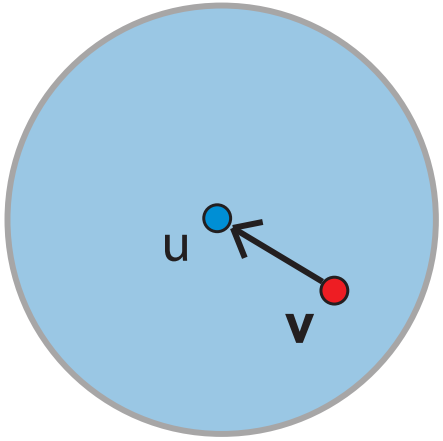
[Flaxman, Frieze, Vera '04]

- n points are randomly distributed on a sphere, in sequence.
- Each node i chooses m neighbours among the nodes that are within a present distance r of i,
- The neighbours are chosen with link probability based on global degree.

Power law degree distribution
Small separators
Connected whp if $r \geq n^{-1/2+\beta} \log n$

Spatial Preferred Attachment (SPA) Model

- Generates directed graphs.
- Nodes are points in Euclidean space.
- Each node has a "sphere of influence" centered at the node.
- The size of the sphere of influence is determined by the in-degree of the node.
- A new node v can only link to an existing node u if v falls within the sphere of influence of u.
- If v falls into the sphere of influence u, it will link to u with probability p.



Spatial Preferred Attachment (SPA) Model

- Space: S, the surface of a sphere in $\mathbb{R}^3$ of area 1
- Sphere of influence of vertex v at time t: $R(v,t)$, the cap around v that has area

$$\frac{c_1 d^-(v,t) + c_2}{t + c_3},$$

  where $d^-(v,t)$ is the in-degree of node v at time t.
- G_0 is the empty graph.
- At each time step $t > 0$, a new node v_t is chosen u.a.r. from S, and added to G_{t-1} to create G_t.
- For each vertex u of G_{t-1} so that $v \in R(u, t-1)$, independently, a directed edge (v_t, u) is created with probability p.

In-degree distribution

$N_{k,t}$ = the number of vertices of in-degree k at time t.

Let u be a node of in-degree k.

$$\mathbb{P}(v_t \text{ lands in sphere of influence of node of } u) : \quad \frac{c_1 k + c_2}{t + c_3}$$

$$\mathbb{P}(v_t \text{ links to } u \text{ if } v_t \text{ fall in its sphere of influence}) : \quad p$$

$$\mathbb{E}(N_{0,t+1} - N_{0,t} \mid G_t) = 1 - p\frac{c_2}{t + c_3} N_{0,t-1}.$$

For $k \geq 1$:

$$\mathbb{E}(N_{k,t+1} - N_{k,t} \mid G_t) = p\frac{c_1(k-1) + c_2}{t + c_3} N_{k-1,t-1} - p\frac{c_1 k + c_2}{t + c_3} N_{k,t-1}$$

The in-degree distribution follows a power law with exponent $1+\frac{1}{pc_1}$:

$$\lim_{t\to\infty} \mathbb{E}(N_{0,t})/t = n_0, \text{ where } n_0 = \frac{1}{1+pc_2},$$

and for all $k \geq 1$,

$$\lim_{t\to\infty} \mathbb{E}(N_{k,t})/t = n_k, \text{ where } n_k \sim k^{-(1+\frac{1}{pc_1})}.$$

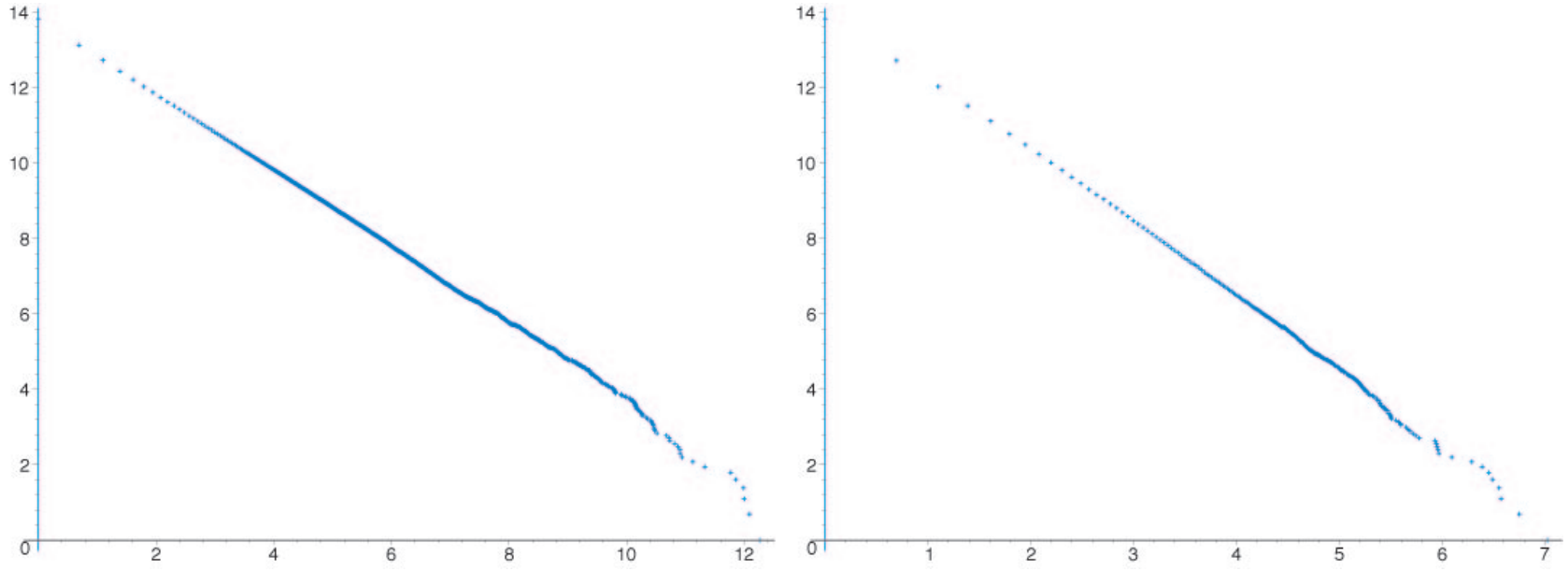
The variables $N_{k,t}$ are concentrated around the mean:

With extreme probability in n, for $0 \leq k \leq (\frac{n}{\log n^4})^{p/6p+2}$, for all $t \leq n$,

$$N_{k,t} = n_k t(1+o(1)).$$

In-degree distribution: simulations

Cumulative in-degree distribution, obtained from a graph with 1,000,000 nodes.



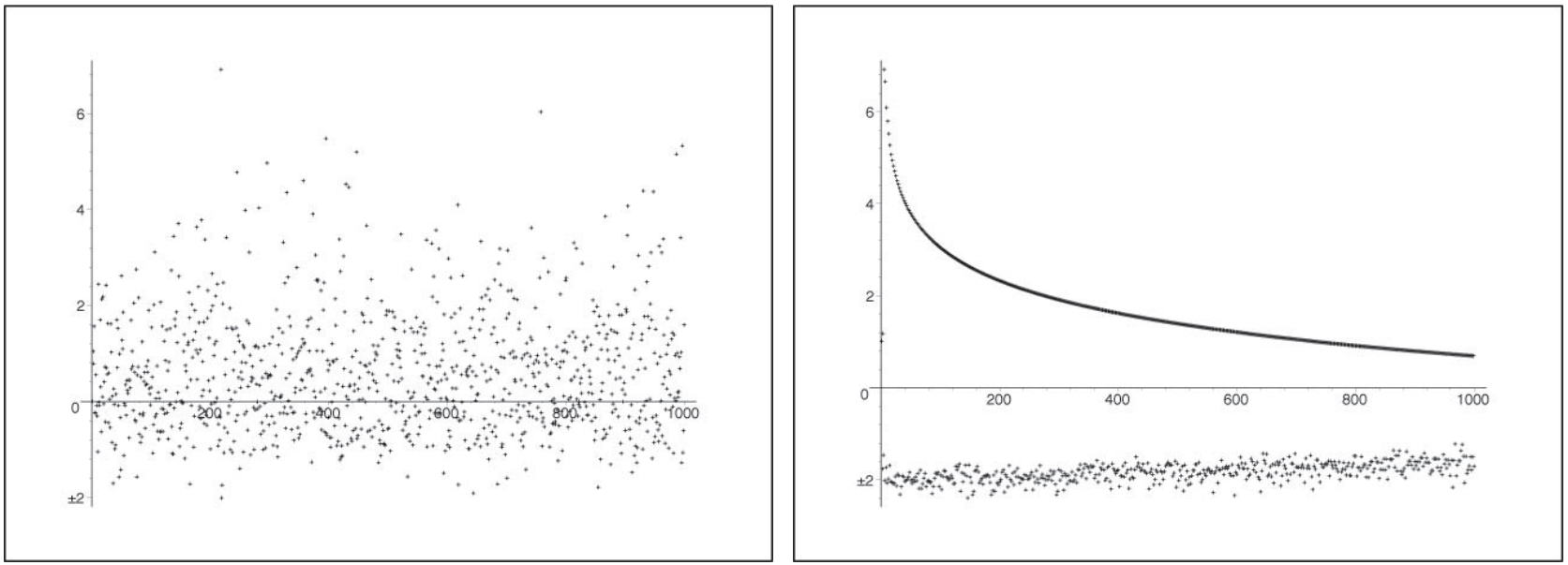
Parameters: $c_1 = c_2 = c_3 = 1$. Left: $p = 1$. Right: $p = 0.5$. Generated on a torus.

In-degree of a node

Let v_i be the node born at time i.

$$\mathbb{E}(d^-(v_i, t)) \sim \left(\frac{t}{i}\right)^{pc_1}$$

But $d^-(v_i)$ is not concentrated:

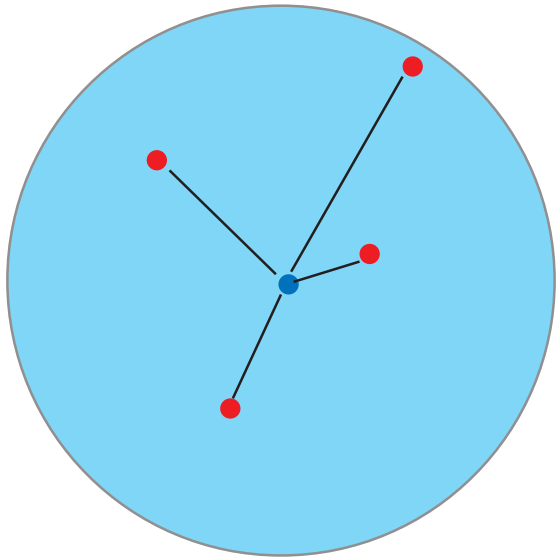


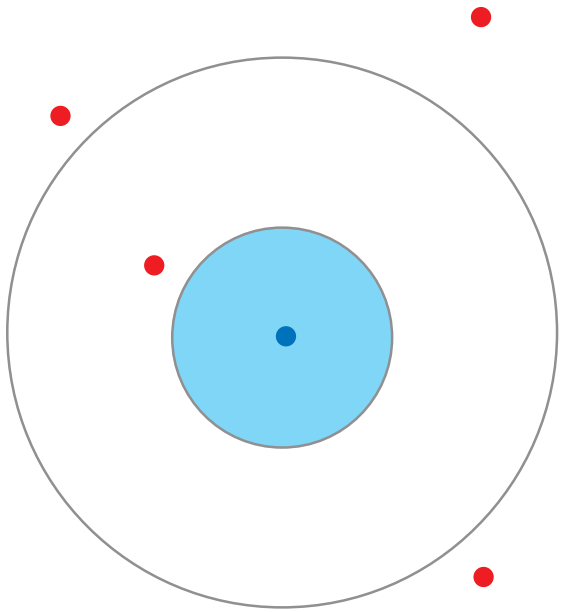
Left: time born vs. in-degree of node.

Right: time vs. maximum deviation of in-degree (log-scale) from the expected value.

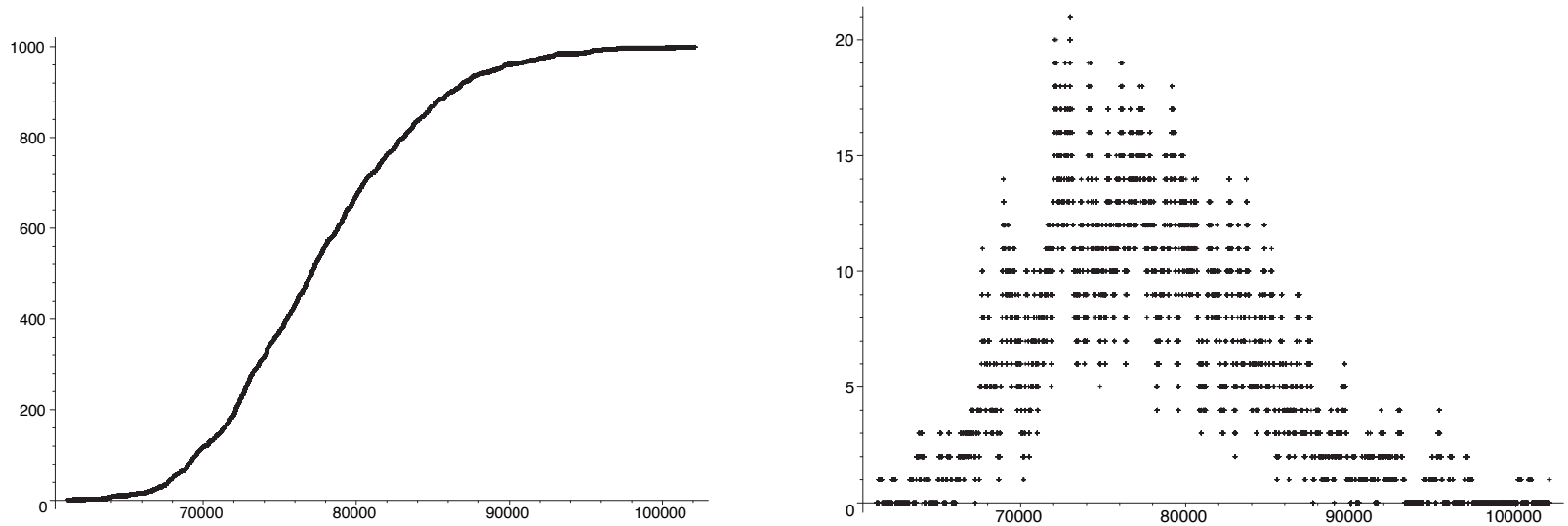
Initial steps matter!

t=5





Number of edges



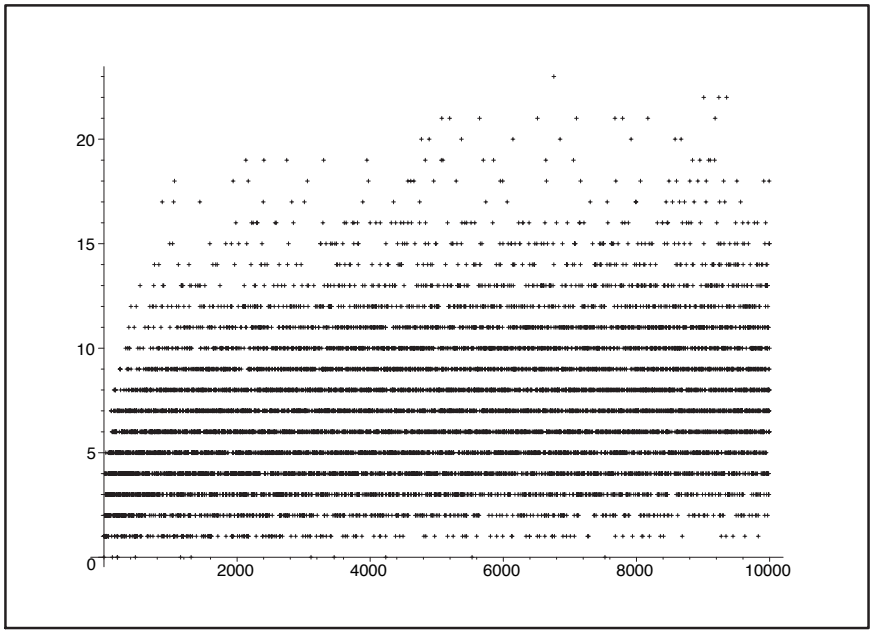
left: $(i, |\{G^i : |E(G^i)| \leq i\}|)$, **right:** $(i, |\{G^i : i - 100 \leq |E(G^i)| \leq i + 100\}|)$

From simulations of $1,000$ graphs G^i, $i \in [1 \ldots 1000]$, on $10,000$ nodes. Parameters: $c_1 = c_2 = c_3 = 1$, $p = 1$.

The sparsest graph had $61,154$ edges, the densest one had $102,152$ edges.

Out-degree

$$\mathbb{E}(d^+(v_i)) = \sum_{v \in V(G_{i-1})} p\left(\frac{c_1 d^-(v) + c_2}{i - 1 + c_3}\right) = \frac{c_2}{1 - pc_1}(1 + o(1)).$$

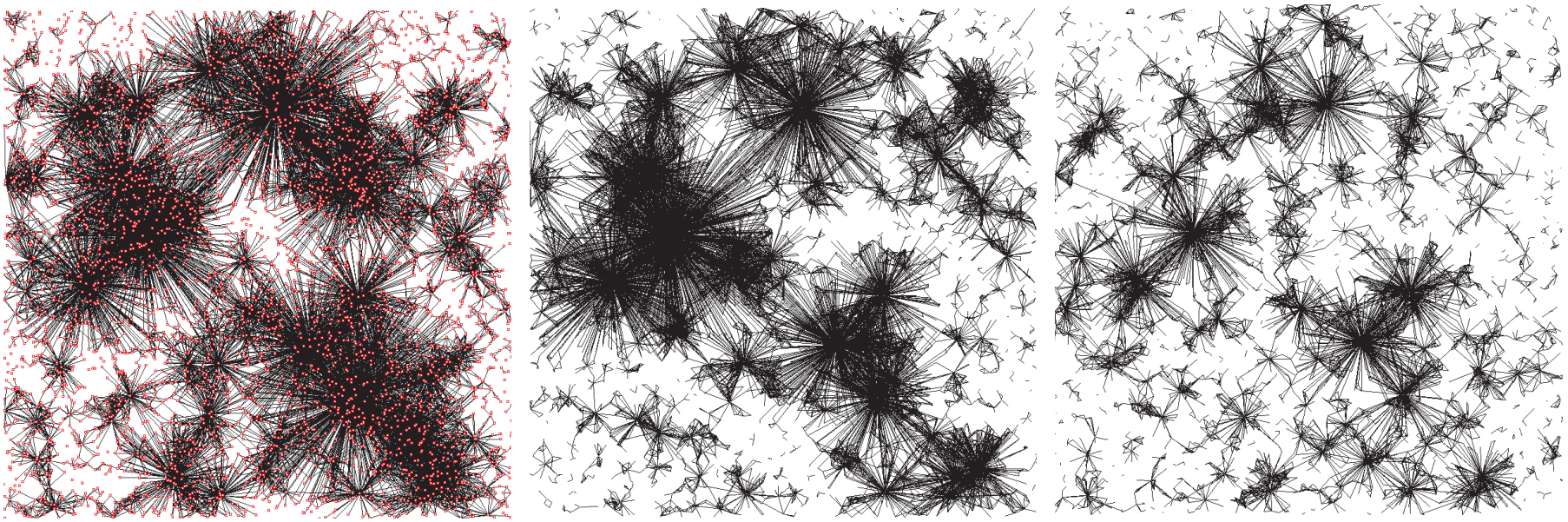


Out-degree sequence: $(i, \deg^+(v_i))$, $c_1 = c_2 = c_3 = 1$, $p = 1$.

Small world

- Graphs generated by SPA-model are acyclic
- The underlying graph is not connected; there are many isolated nodes
- Conjecture: the largest induced directed path has length $O(logn)$.
- Conjecture: Many long links: Distance between endpoints of an edge can be large.

Simulation results



$p = 1$, $p = 0.9$, and $p = 0.8$

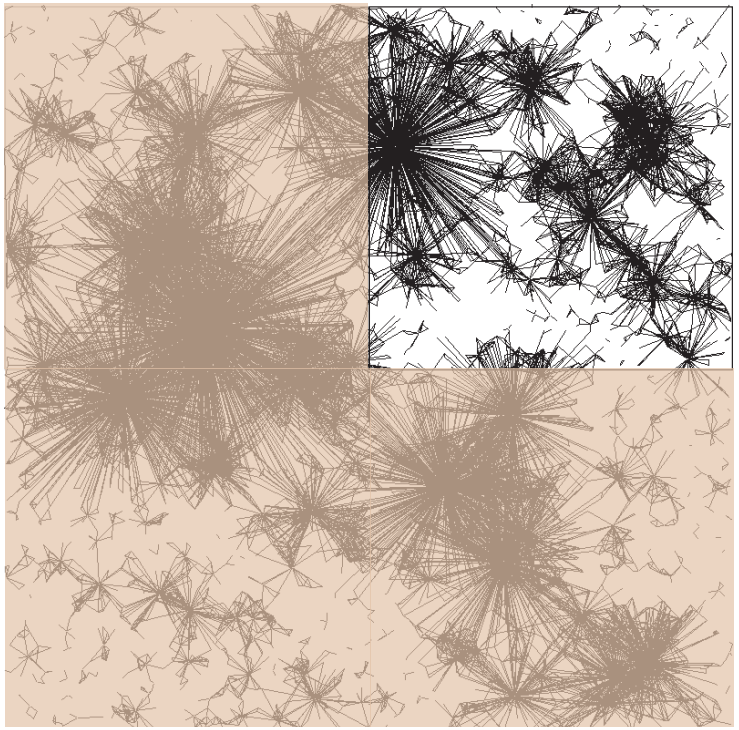
Generated on the unit square.

Scaling and self-similarity

Let $S_0 \subseteq S$ be a convex subset of S with area c. Then for all k for which we have concentration around the mean,

$$N_k(t) \cap S_0 \sim cN_k(t).$$

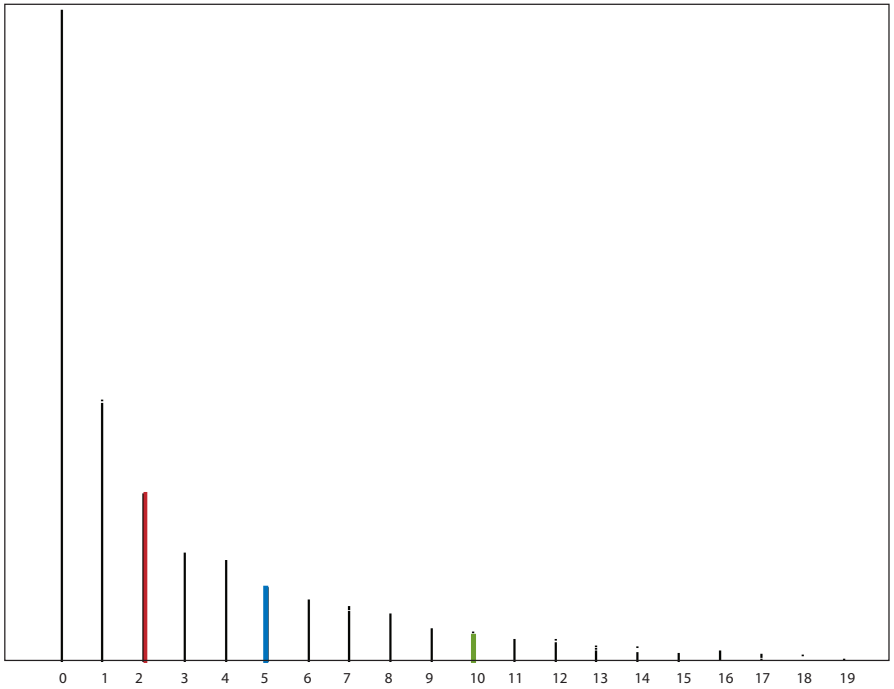
So the in-degree distribution of all nodes in S_0 follows a power law with the same exponent as the original graph.

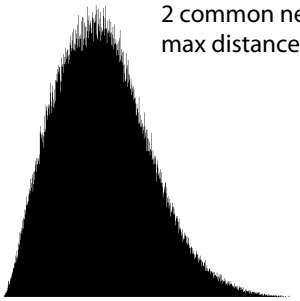


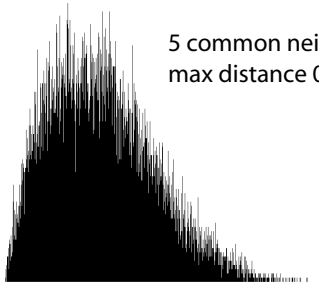
Does the graph induced by all nodes in S_0 have many of the same graph properties as the entire graph?

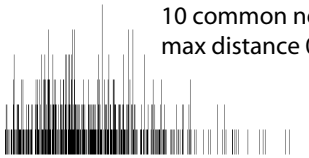
Estimating the geometry

Is the number of common out-neighbours of a pair of nodes an indicator of the distance between the nodes?









(# common neighbours, distance) for each pair

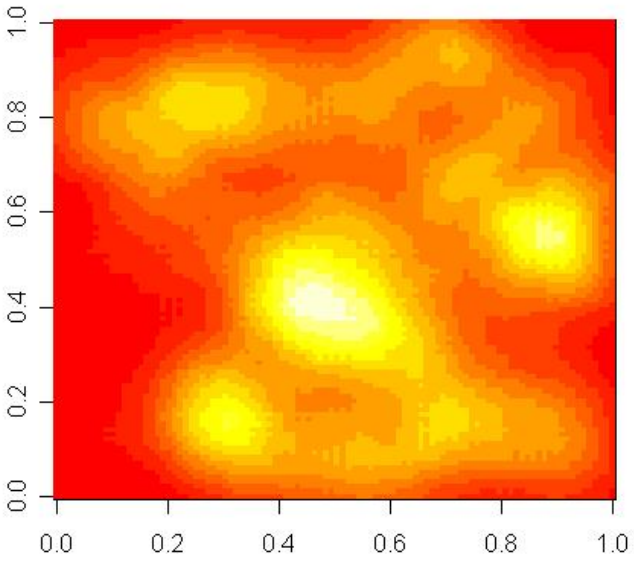
Future Work

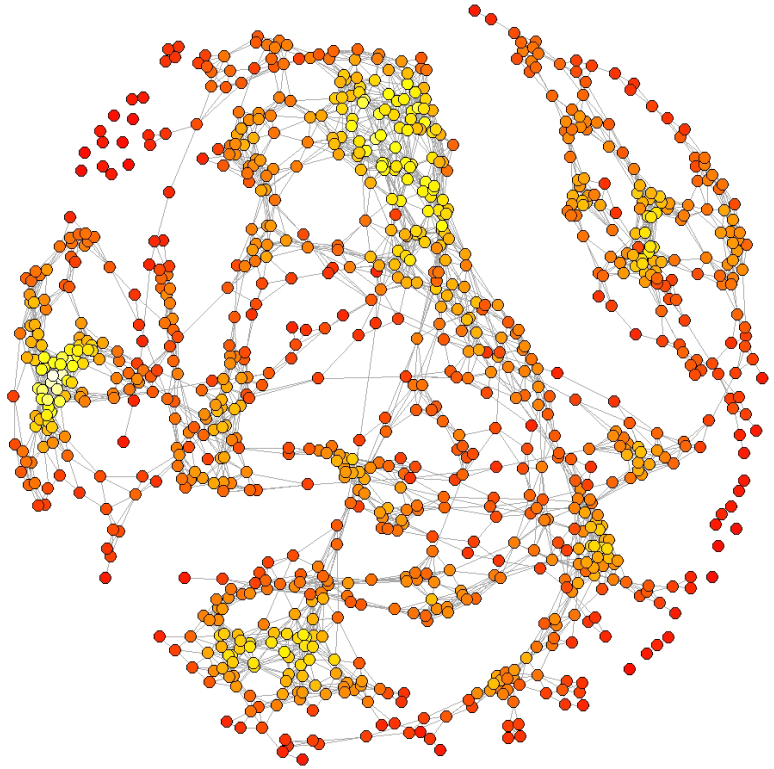
- Generalize the model:
  - Node and edge deletion
  - Adding edges to existing nodes
  - Updating the out-links of a node
- Undirected graphs
- Rank-based preferential attachment: each node v has a rank $r(v,t)$ at time t, and the area of the sphere of influence of v is proportional to

$$r(v,t)^{-\alpha}$$

Non-uniform distribution of points

Generate points using the principle of **cumulative advantage**: areas that contain many points have a higher probability of receiving new points.





$\Rightarrow$

Future Work: Adapt to real-world networks

- Adapt the model to specific types of real-world networks
- Find the right parameters from power law exponent etc.
- Validate the model by comparing graph properties
- Use the model to estimate the underlying geometry of the nodes