Network Analysis and Mining
5. Random graph models II

Maximilien Danisch, Lionel Tabourier

LIP6 – CNRS and Sorbonne Université

first_name.last_name@lip6.fr

November 3 2020

Notion of uniform generation

Until now, we have seen two different kinds of models:

1. Erdős-Rényi
2. Watts-Strogatz, Barabási-Albert

Why are they fundamentally different?

1. ER: there is a target set (graphs with fixed density)
   all graphs have the same probability to be produced
2. BA, WS: no explicit target set...

$\Rightarrow$ ER model is uniform (or homogeneous)

Notion of uniform generation

Until now, we have seen two different kinds of models:

1. Erdős-Rényi
2. Watts-Strogatz, Barabási-Albert

Why are they fundamentally different?

1. ER: there is a target set (graphs with fixed density)
   all graphs have the same probability to be produced
2. BA, WS: no explicit target set...

$\Rightarrow$ ER model is uniform (or homogeneous)

Notion of uniform generation

Until now, we have seen two different kinds of models:

1. Erdős-Rényi
2. Watts-Strogatz, Barabási-Albert

Why is it important?

Because we cannot say that a BA is a *standard* SF graph
or that a WS is a *standard* graph with small-world properties

$\Rightarrow$ more relevant to have uniform graph generation

Notion of uniform generation

Until now, we have seen two different kinds of models:

1. Erdős-Rényi
2. Watts-Strogatz, Barabási-Albert

Why is it important?

Because we cannot say that a BA is a *standard* SF graph or that a WS is a *standard* graph with small-world properties

$\Rightarrow$ more relevant to have uniform graph generation

Outline

The configuration model

Degree distribution

$p_1, p_2, p_3, \ldots$

Draw nodes degree according to the distribution

$\rightarrow$ degree sequence

1 2 4 3 2 1 3

Associate to any node half-edges (stubs)

Draw random pairs of stubs and connect them

Deal with possible loops or multi-edges

The configuration model

Degree distribution

$p_1, p_2, p_3, \ldots$

Draw nodes degree according to the distribution

$\rightarrow$ degree sequence

1 2 4 3 2 1 3

Associate to any node half-edges (stubs)

Draw random pairs of stubs and connect them

Deal with possible loops or multi-edges

The configuration model

Degree distribution

$p_1, p_2, p_3, \ldots$

Draw nodes degree according to the distribution

$\rightarrow$ degree sequence

1 2 4 3 2 1 3

Associate to any node half-edges (stubs)

Draw random pairs of stubs and connect them

Deal with possible loops or multi-edges

The configuration model

Degree distribution

$p_1, p_2, p_3, \ldots$

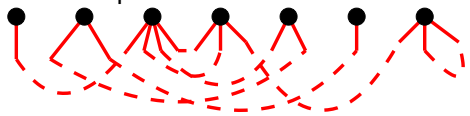
Draw nodes degree according to the distribution

$\rightarrow$ degree sequence

1 2 4 3 2 1 3

Associate to any node half-edges (stubs)

Draw random pairs of stubs and connect them



Deal with possible loops or multi-edges

The configuration model

Degree distribution

$p_1, p_2, p_3, \ldots$

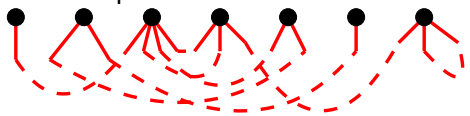
Draw nodes degree according to the distribution

$\rightarrow$ degree sequence

1 2 4 3 2 1 3

Associate to any node half-edges (stubs)

Draw random pairs of stubs and connect them



Deal with possible loops or multi-edges

Implementing the configuration model

Table : node i occurs exactly $d^\circ(i)$ times

0	1	1	2	2	2	2	3	3	3	4	4	5	6	6	6

Algorithm 1: Generating a graph with fixed degree sequence

$i = 2m$
while $i > 0$ **do**
  u = random (0 , $i - 1$)
  swap boxes u and $i - 1$
  v = random (0 , $i - 2$)
  swap boxes v and $i - 2$
  i = i-2
  edge (u, v) created*
end

$*$ to be discussed...

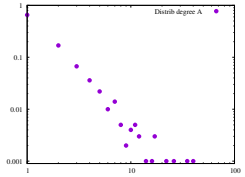
Deal with possible loops or multi-edges

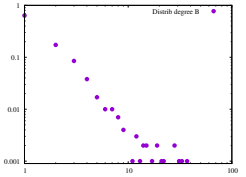
Answer 1: generation with rejection

Loop or multi-edge generated, restart the generation process

Deal with possible loops or multi-edges

- advantage: uniform generation
- drawback: can be long...





average number of trials (1000 nodes):

2180 17300

average generation time (1000 nodes):

1.2s 8.1s

Quiz: for what kind of distribution can it be long?

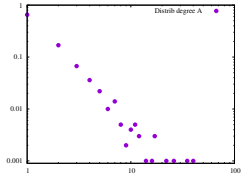
Deal with possible loops or multi-edges

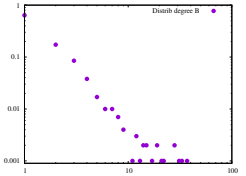
Answer 2: suppress loops or multiple edges

When a loop or a multiedge is generated, exclude it

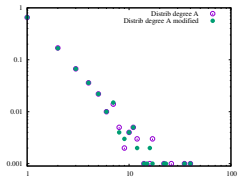
Deal with possible loops or multi-edges

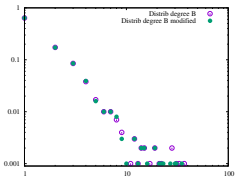
- advantage: fast
- drawback: does not have the exact degree sequence





after loops and multi-edges deletion, become:

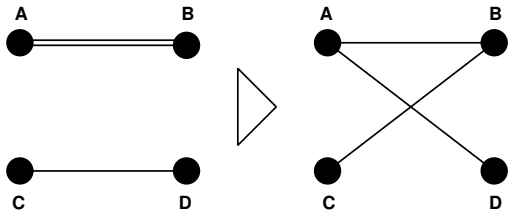




Deal with possible loops or multi-edges

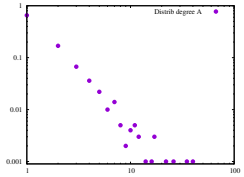
Answer 3: reconnect

When a loop or a multiedge is generated, switch to destroy it

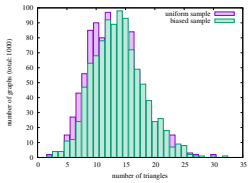


Deal with possible loops or multi-edges

- advantage: relatively fast, have the exact sequence
- drawback: not uniform = biased

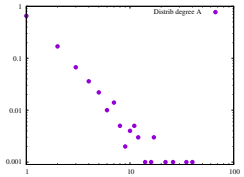


number of triangles for 1000 graphs

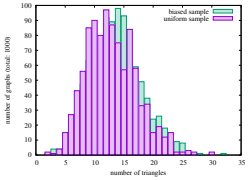


Deal with possible loops or multi-edges

- advantage: relatively fast, have the exact sequence
- drawback: not uniform = biased



number of triangles for 1000 graphs



Properties – Comparison

	real	fixed d.d.
density	low	?
connectedness	giant comp.	?
distances	low	?
degree	heterogeneous	?
clustering	high	?
communities	yes	?

Properties – Comparison

	real	fixed d.d.
density	low	low
connectedness	giant comp.	giant comp.
distances	low	low
degree	heterogeneous	heterogeneous
clustering	high	lower
communities	yes	no

→ heterogeneous degree only partly accounts for the c.c.
→ see practical work

Other implementation: switching method

Principle

- start from a graph with the given degree sequence
- iterate switching of edge ends
- after a *sufficient amount* of switches, the graph produced is a random element of the set of graphs

Other implementation: switching method

Why does it work?

- The degree of any node remains unchanged
  so we keep the degree sequence unchanged
- The process is a Markov chain
  - can be seen as a **random walk** in the set of graphs (defined by this degree sequence)
  - after a while, we visit all elements with the same probability (not proved here)
  - if we make enough switches, we obtain a random graph with this degree sequence

Other implementation: switching method

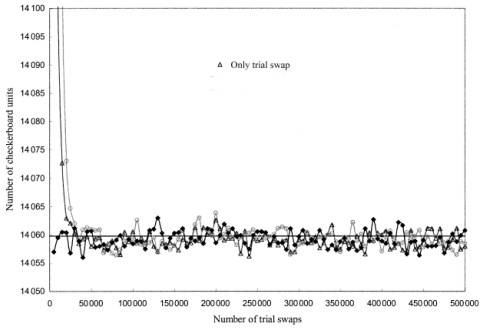
Why does it work?

- The degree of any node remains unchanged
  so we keep the degree sequence unchanged
- The process is a Markov chain
  - can be seen as a **random walk** in the set of graphs (defined by this degree sequence)
  - after a while, we visit all elements with the same probability (not proved here)
  - if we make enough switches, we obtain a random graph with this degree sequence

Other implementation: switching method

When to stop switchings?

Measuring some features (ex: clustering) during the process until these features do not evolve any more...

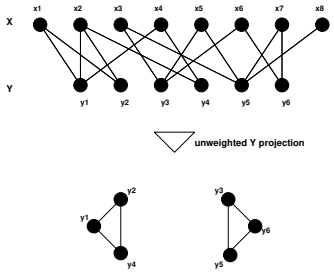


credits image: I.Miklós and J.Podani

Graph with fixed degree sequence: the bipartite case

Newman, Watts, Strogatz - *PNAS, 2002*

- Bipartite graph: two distinct types of nodes U and V
  $\rightarrow$ links between U and V
- Projection: if u_1 and u_2 connected to v in bipartite
  $\rightarrow$ u_1 and u_2 are connected in the U-projection



bipartite data richer, but not always available

Graph with fixed degree sequence: the bipartite case

Newman, Watts, Strogatz - *PNAS, 2002*

Underlying bipartite structure $\Rightarrow$ cliques in the projection

Bipartite configuration model

- fixed degree sequence for nodes X: $d_1^x, d_2^x, \ldots, d_{n_X}^x$
- fixed degree sequence for nodes Y: $d_1^y, d_2^y, \ldots, d_{n_Y}^y$
- random connections

$\rightarrow$ no possible self-loops, but multiedges still a problem

Graph with fixed degree sequence: the bipartite case

Newman, Watts, Strogatz - *PNAS, 2002*

Experimental results - comparison of the projections:

	clustering coef.		average degree	
projected network	Model	Real	Model	Real
Company directors	0.590	0.588	14.53	14.44
Movie actors	0.084	0.199	125.6	113.4
Physics collaboration	0.192	0.452	16.74	9.27

Conclusions:

- much more realistic clustering
- but still no large-scale structure
  visible on scientific collaboration networks

Graph with fixed degree sequence: the bipartite case

Newman, Watts, Strogatz - *PNAS, 2002*

Experimental results - comparison of the projections:

	clustering coef.		average degree	
projected network	Model	Real	Model	Real
Company directors	0.590	0.588	14.53	14.44
Movie actors	0.084	0.199	125.6	113.4
Physics collaboration	0.192	0.452	16.74	9.27

Conclusions:

- much more realistic clustering
- but still no large-scale structure
  visible on scientific collaboration networks

Perspective: more models

- Fix other constraints beyond degree distribution? but how?
- Exponential Random Graphs
- Stochastic Block Model
- Spatial models
- ...

→ still many open research questions

Outline

About this section

From Part A (Complex Networks Analysis tools)
→ to Part B (Graph mining)

- use our knowledge and tools to explore real networks
- adapt them to specific problems

→ problem-oriented view

Illustration: a few Social Network Analysis concepts

About this section

From Part A (Complex Networks Analysis tools)
→ to Part B (Graph mining)

- use our knowledge and tools to explore real networks
- adapt them to specific problems

→ problem-oriented view

Illustration: a few Social Network Analysis concepts

What is homophily?

From Greek, *homo:* same, similar and *philos:* friend of, to like
→ *"birds of a feather flock together"*

Observed for a long time in sociology:

- smoking habits, food habits
- residential segregation
- voting behavior
- ...

How to observe this phenomenon through SNA?

What is homophily?

From Greek, *homo:* same, similar and *philos:* friend of, to like
→ *"birds of a feather flock together"*

Observed for a long time in sociology:

- smoking habits, food habits
- residential segregation
- voting behavior
- ...

How to observe this phenomenon through SNA?

Structural homophily, based on degree

Degree-based assortativity

Do high-degree nodes connect to high-degree nodes?

We denote:

- q_k probability distribution of the remaining degree:
  if p_k is the degree distribution then $q_k = \frac{(k+1)p_{k+1}}{\sum_j j.p_j}$
- e_{jk} joint probability of remaining degree distribution (rdd),
  i.e. probability to pick an edge which ends have remaining degree j and k

We measure the assortativity coefficient r:

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n}\sum_k q_k(k - \overline{k})^2$$

Structural homophily, based on degree

Degree-based assortativity

Do high-degree nodes connect to high-degree nodes?

We measure the assortativity coefficient r:

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n}\sum_k q_k (k - \overline{k})^2$$

A few remarks:

- σ_q^2 is the variance of the rdd q_k
- r is a normalized quantity ($\in [-1:1]$)
- r: ratio between covariance across ties of the rdd and variance of the rdd

Newman - *Phys. Rev. E, 2003*

Structural homophily, based on degree

Typical degree-assortativity on social networks:

- astrophysics coauthorship: $r = 0.235$ (Georgia Tech data)
- actor collaboration: $r = 0.227$ (Notre-Dame Univ data)
- friendship network: $r = 0.039$ (Livejournal data)

→ social networks are usually degree-assortative

Note that all complex networks are not degree-assortative:

- Internet AS level: $r = -0.215$ (UCLA data)
- human protein network: $r = -0.126$ (Vidal data)
- US power grid network: $r = 0.003$ (Tore Opsahl data)

Structural homophily, based on degree

Typical degree-assortativity on social networks:

- astrophysics coauthorship: $r = 0.235$ (Georgia Tech data)
- actor collaboration: $r = 0.227$ (Notre-Dame Univ data)
- friendship network: $r = 0.039$ (Livejournal data)

→ social networks are usually degree-assortative

Note that all complex networks are not degree-assortative:

- Internet AS level: $r = -0.215$ (UCLA data)
- human protein network: $r = -0.126$ (Vidal data)
- US power grid network: $r = 0.003$ (Tore Opsahl data)

Structural homophily: other kinds of assortativity

Degree-based assortativity

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n}\sum_k q_k (k - \overline{k})^2$$

r: ratio covariance across ties / variance of remaining degree k

General assortativity

$$\rho = \frac{\sum_{\lambda,\kappa} \lambda.\kappa.(e_{\lambda\kappa} - \chi_\lambda \chi_\kappa)}{\sigma_\chi^2} \text{ with } \sigma_\chi^2 = \frac{1}{n}\sum_\kappa \chi_\kappa (\kappa - \overline{\kappa})^2$$

r: ratio covariance across ties / variance of trait κ
χ_κ: fraction of edges that start and end at vertices with value κ
examples for κ: age, salary, number of children...

Structural homophily: other kinds of assortativity

Degree-based assortativity

$$r = \frac{\sum_{j,k} j.k.(e_{jk} - q_j q_k)}{\sigma_q^2} \text{ with } \sigma_q^2 = \frac{1}{n}\sum_k q_k(k - \overline{k})^2$$

r: ratio covariance across ties / variance of remaining degree k

General assortativity

$$\rho = \frac{\sum_{\lambda,\kappa} \lambda.\kappa.(e_{\lambda\kappa} - \chi_\lambda \chi_\kappa)}{\sigma_\chi^2} \text{ with } \sigma_\chi^2 = \frac{1}{n}\sum_\kappa \chi_\kappa(\kappa - \overline{\kappa})^2$$

r: ratio covariance across ties / variance of trait κ
χ_κ: fraction of edges that start and end at vertices with value κ
examples for κ*:* age, salary, number of children...

Transitivity and clustering

Reminder: social networks have a high average clustering

Triadic closure phenomenon

Old concept in sociology (Simmel, 1908).
Hypothesis on the growth dynamics of a network:

Consequences:

- high clustering
- large number of cliques (complete subgraphs)
- densification over time

Transitivity and clustering

Reminder: social networks have a high average clustering

Triadic closure phenomenon

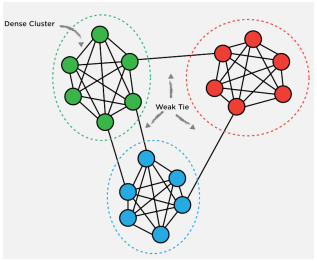
Old concept in sociology (Simmel, 1908).
Hypothesis on the growth dynamics of a network:

Consequences:

- high clustering
- large number of cliques (complete subgraphs)
- densification over time

Transitivity and clustering

$\Rightarrow$ schematic picture of social networks

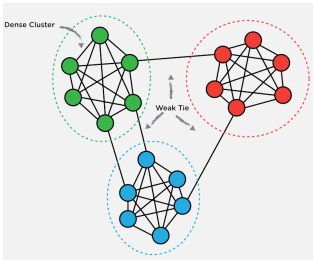


credits image: V.Gauthier

Warning: only a schematic representation
misses overlaps in clusters, groups hierarchy, core/periphery...

Transitivity and clustering

$\Rightarrow$ schematic picture of social networks



credits image: V.Gauthier

Warning: only a schematic representation
misses overlaps in clusters, groups hierarchy, core/periphery...

About weak ties...

Hypothesis of the strength of weak tie Granovetter - *1973*

A "weak tie" is a link in a social network which represents a relation which is not frequently maintained

It is argued that weak ties play an essential role as they **ensure connections between groups**

What is considered a strong tie in social sciences?

- frequent contacts
- strong affinity (if measurable)
- structural criterion: many mutual neighbors

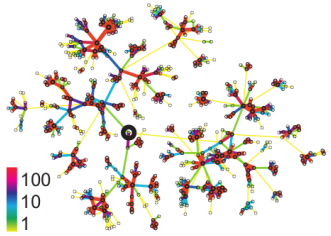
About weak ties...

Onnela et al. - *2007*

Experimental validation on a phonecall network
$\rightarrow$ are weaker links between clusters?

- strength of a relationship = cumulative duration of calls
- link between groups measured with link betweenness

weight (color) = cumulative duration of calls



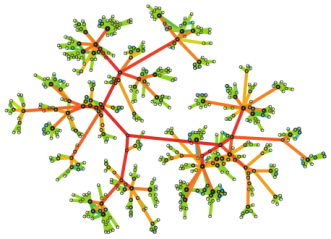
About weak ties...

Onnela et al. - *2007*

Experimental validation on a phonecall network
$\rightarrow$ are weaker links between clusters?

- strength of a relationship = cumulative duration of calls
- link between groups measured with link betweenness

weight (color) = link betweenness



About weak ties. . .

Onnela et al. - *2007*

Experimental validation on a phonecall network
→ are weaker links between clusters?

Quiz: what could you plot to check the correlation ?

Effects on spreading in a social network

Examples: innovation spreading, rumor spreading, adversiting. . .

What can we expect from the previous observations?
- fast spreading within a community
- use of weak links to spread from a group to another

In practice hard to measure experimentally:
- “contagion” hard to track and isolate
- spreading rarely reaches a large part of a network

→ very active field of research

Uniform graph generation with fixed degree distribution
Applying our tools to Social Network Analysis
The homophily phenomenon
Local density and community structure
Impact on social contagion

Effects on spreading in a social network

Examples: innovation spreading, rumor spreading, adversiting. . .

What can we expect from the previous observations?
- fast spreading within a community
- use of weak links to spread from a group to another

In practice hard to measure experimentally:
- “contagion” hard to track and isolate
- spreading rarely reaches a large part of a network

→ very active field of research

Uniform graph generation with fixed degree distribution
Applying our tools to Social Network Analysis
The homophily phenomenon
Local density and community structure
Impact on social contagion

Effects on spreading in a social network

Examples: innovation spreading, rumor spreading, adversiting. . .

What can we expect from the previous observations?
- fast spreading within a community
- use of weak links to spread from a group to another

In practice hard to measure experimentally:
- “contagion” hard to track and isolate
- spreading rarely reaches a large part of a network

→ very active field of research