Directed Random Graphs with Given Degree Distributions

Mariana Olvera-Cravioto

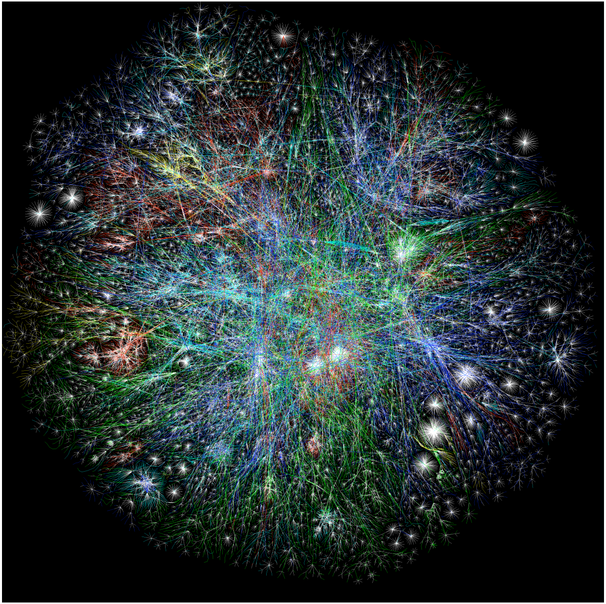
Columbia University

molvera@ieor.columbia.edu

Joint work with Ningyuan Chen

July 23th, 2012

The motivating example: WWW



Opte project. Part of the MoMA permanent collection

The World Wide Web

- WWW seen as a directed graph (webpages = nodes, links = edges).
- Empirical observations:

$$\text{fraction pages} > k \text{ in-links} \propto k^{-\alpha}, \qquad \alpha = 1.1$$

$$\text{fraction pages} > k \text{ out-links} \propto k^{-\beta}, \qquad \beta = 1.72$$

- We want a directed random graph model that matches the degree distributions.

Degree distributions

- Directed graph on n nodes $V = \{v_1, \dots, v_n\}$.
- In-degree and out-degree:
  - $m_i =$ in-degree of node $v_i =$ number of edges pointing to v_i.
  - $d_i =$ out-degree of node $v_i =$ number of edges pointing from v_i.
- $(\mathbf{m}, \mathbf{d}) = (\{m_i\}, \{d_i\})$ is called a bi-degree-sequence.
- **Target distributions:**

$$F = (f_k : k = 0, 1, 2, \dots), \quad \text{and}$$
$$G = (g_k : k = 0, 1, 2, \dots).$$

- We want the bi-degree-sequence to satisfy:

$$\frac{1}{n} \sum_{i=1}^{n} 1(m_i = k) \approx f_k \qquad \text{and} \qquad \frac{1}{n} \sum_{i=1}^{n} 1(d_i = k) \approx g_k.$$

Simple graphs

- **Definition:** We say that a directed graph is *simple* if it has no self-loops and at most one edge in each direction between any two nodes.
- **Definition:** We say that $(\mathbf{m}, \mathbf{d})$ is *graphical* if there exists a simple directed graph having $(\mathbf{m}, \mathbf{d})$ as its bi-degree-sequence.
- **Goal:** Choose a graph uniformly at random from all simple graphs having bi-degree-sequence $(\mathbf{m}, \mathbf{d})$, where $(\mathbf{m}, \mathbf{d})$ has approximately the target distributions F and G.

Simple graphs

- **Definition:** We say that a directed graph is *simple* if it has no self-loops and at most one edge in each direction between any two nodes.
- **Definition:** We say that $(\mathbf{m}, \mathbf{d})$ is *graphical* if there exists a simple directed graph having $(\mathbf{m}, \mathbf{d})$ as its bi-degree-sequence.
- **Goal:** Choose a graph uniformly at random from all simple graphs having bi-degree-sequence $(\mathbf{m}, \mathbf{d})$, where $(\mathbf{m}, \mathbf{d})$ has approximately the target distributions F and G.
- Two problems:

Simple graphs

- **Definition:** We say that a directed graph is *simple* if it has no self-loops and at most one edge in each direction between any two nodes.
- **Definition:** We say that $(\mathbf{m}, \mathbf{d})$ is *graphical* if there exists a simple directed graph having $(\mathbf{m}, \mathbf{d})$ as its bi-degree-sequence.
- **Goal:** Choose a graph uniformly at random from all simple graphs having bi-degree-sequence $(\mathbf{m}, \mathbf{d})$, where $(\mathbf{m}, \mathbf{d})$ has approximately the target distributions F and G.
- Two problems:
  - Construct an appropriate bi-degree-sequence that with high probability will be graphical.

Simple graphs

- **Definition:** We say that a directed graph is *simple* if it has no self-loops and at most one edge in each direction between any two nodes.
- **Definition:** We say that $(\mathbf{m}, \mathbf{d})$ is *graphical* if there exists a simple directed graph having $(\mathbf{m}, \mathbf{d})$ as its bi-degree-sequence.
- **Goal:** Choose a graph uniformly at random from all simple graphs having bi-degree-sequence $(\mathbf{m}, \mathbf{d})$, where $(\mathbf{m}, \mathbf{d})$ has approximately the target distributions F and G.
- Two problems:
  - Construct an appropriate bi-degree-sequence that with high probability will be graphical.
  - Choose uniformly at random a simple graph from such bi-degree-sequence.

The configuration model (Wormald '78, Bollobas '80)

- For undirected graphs, given a degree sequence $\mathbf{d} = (d_1, \dots, d_n)$:
  - assign to each node v_i a number d_i of stubs or half-edges;
  - for the first half-edge of node v_1 choose uniformly at random from all other half-edges, and if the selected half-edge belongs to, say, node v_j, draw an edge between node v_1 and v_j;
  - proceed in the same way for all remaining unpaired half-edges, i.e., choose uniformly from the set of unpaired half-edges and draw an edge between the current node and the node to which the selected half-edge belongs.
- The result is a *multigraph* (e.g., with self-loops and multiple edges) on nodes $\{v_1, \dots, v_n\}$.
- If we discard any realization that is not simple, we obtain a uniformly chosen simple graph.

The directed configuration model

- For directed graphs, given a bi-degree-sequence $(\mathbf{m}, \mathbf{d})$:
  - assign to each node v_i a number m_i of inbound stubs and a number d_i of outbound stubs;
  - pair outbound stubs to inbound stubs to form directed edges by matching to each inbound stub an outbound stub chosen uniformly at random from the set of unpaired outbound stubs.
- The result is again a multigraph, but if we discard realizations that have self-loops or multiple edges we obtain a uniformly chosen simple graph.
- **Questions:**

The directed configuration model

- For directed graphs, given a bi-degree-sequence $(\mathbf{m}, \mathbf{d})$:
  - assign to each node v_i a number m_i of inbound stubs and a number d_i of outbound stubs;
  - pair outbound stubs to inbound stubs to form directed edges by matching to each inbound stub an outbound stub chosen uniformly at random from the set of unpaired outbound stubs.
- The result is again a multigraph, but if we discard realizations that have self-loops or multiple edges we obtain a uniformly chosen simple graph.
- **Questions:**
  - What is the probability of the resulting graph being simple?

The directed configuration model

- For directed graphs, given a bi-degree-sequence $(\mathbf{m}, \mathbf{d})$:
  - assign to each node v_i a number m_i of inbound stubs and a number d_i of outbound stubs;
  - pair outbound stubs to inbound stubs to form directed edges by matching to each inbound stub an outbound stub chosen uniformly at random from the set of unpaired outbound stubs.
- The result is again a multigraph, but if we discard realizations that have self-loops or multiple edges we obtain a uniformly chosen simple graph.
- **Questions:**
  - What is the probability of the resulting graph being simple?
  - Under what conditions is it bounded away from zero as $n \to \infty$?

Probability of graph being simple

- For the undirected configuration model it is known that if $\mathbf{d}$ satisfies certain *regularity conditions*, the number of self-loops, S_n, and the number of multiple edges, M_n, satisfy

$$(S_n, M_n) \Rightarrow (S, M) \qquad n \to \infty,$$

  where S and M are independent Poisson r.v.s. (Janson '09, Van der Hofstad '08-'12).
- Then,

$$\lim_{n\to\infty} P(\text{graph is simple}) = P(S = 0, M = 0) > 0.$$

- The same should be true for the directed version.

Regularity conditions

Given $\{(\mathbf{m}_n, \mathbf{d}_n)\}_{n \in \mathbb{N}}$ satisfying $\sum_{i=1}^n m_{ni} = \sum_{i=1}^n d_{ni}$ for all n, let

$$P((N^{[n]}, D^{[n]}) = (i, j)) = \frac{1}{n} \sum_{k=1}^n 1(m_{nk} = i, d_{nk} = j).$$

1. *Weak convergence.* For some γ, ξ with $E[\gamma] = E[\xi] > 0$,

$$(N^{[n]}, D^{[n]}) \Rightarrow (\gamma, \xi), \qquad n \to \infty.$$

2. *Convergence of the first moments.*

$$\lim_{n \to \infty} E[N^{[n]}] = E[\gamma] \qquad \text{and} \qquad \lim_{n \to \infty} E[D^{[n]}] = E[\xi].$$

Regularity conditions... continued

3. *Convergence of the covariance.*

$$\lim_{n\to\infty} E[N^{[n]} D^{[n]}] = E[\gamma\xi].$$

4. *Convergence of the second moments.*

$$\lim_{n\to\infty} E[(N^{[n]})^2] = E[\gamma^2] \qquad \text{and} \qquad \lim_{n\to\infty} E[(D^{[n]})^2] = E[\xi^2].$$

- **Note:** $(N^{[n]}, D^{[n]})$ denote the in-degree and out-degree of a randomly chosen node.

Poisson Limit for Self-Loops and Multiple Edges

- **Proposition:** (Chen, O-C '12) If $\{(\mathbf{m}_n, \mathbf{d}_n)\}_{n\in\mathbb{N}}$ satisfies the regularity conditions with $E[\gamma] = E[\xi] = \mu > 0$, then

$$(S_n, M_n) \Rightarrow (S, M)$$

as $n \to \infty$, where S and M are independent Poisson r.v.s with means

$$\lambda_1 = \frac{E[\gamma\xi]}{\mu} \qquad \text{and} \qquad \lambda_2 = \frac{E[\gamma(\gamma-1)]E[\xi(\xi-1)]}{2\mu^2}.$$

- Proof adapted from the undirected case (Van der Hofstad '08 -'12).
- **Theorem:** Under the same assumptions,

$$\lim_{n\to\infty} P(\text{graph obtained from } (\mathbf{m}_n, \mathbf{d}_n) \text{ is simple}) = e^{-\lambda_1-\lambda_2} > 0.$$

The repeated and erased models

- **Repeated model:** If all four regularity conditions are satisfied, then repeat the random pairing until a simple graph is obtained.

The repeated and erased models

- **Repeated model:** If all four regularity conditions are satisfied, then repeat the random pairing until a simple graph is obtained.
- If condition (4) is not satisfied, the probability of obtaining a simple graph converges to zero as $n \to \infty$.

The repeated and erased models

- **Repeated model:** If all four regularity conditions are satisfied, then repeat the random pairing until a simple graph is obtained.
- If condition (4) is not satisfied, the probability of obtaining a simple graph converges to zero as $n \to \infty$.
- **Erased model:** Simply erase the self-loops and merge multiple edges in the same direction into a single edge.

The repeated and erased models

- **Repeated model:** If all four regularity conditions are satisfied, then repeat the random pairing until a simple graph is obtained.
- If condition (4) is not satisfied, the probability of obtaining a simple graph converges to zero as $n \to \infty$.
- **Erased model:** Simply erase the self-loops and merge multiple edges in the same direction into a single edge.
- **Question:** (yet to answer) What are the in-degree and out-degree distributions of the resulting simple graphs?

Degree sequences

- How to construct an appropriate bi-degree-sequence?

Degree sequences

- How to construct an appropriate bi-degree-sequence?
- For the undirected case one can take $\mathbf{D}_n = \{D_1, \ldots, D_n\}$, where the $\{D_i\}$ are i.i.d. r.v.s.

Degree sequences

- How to construct an appropriate bi-degree-sequence?
- For the undirected case one can take $\mathbf{D}_n = \{D_1, \ldots, D_n\}$, where the $\{D_i\}$ are i.i.d. r.v.s.
- **Question:** When is an i.i.d. sequence graphical?

Degree sequences

- How to construct an appropriate bi-degree-sequence?
- For the undirected case one can take $\mathbf{D}_n = \{D_1, \ldots, D_n\}$, where the $\{D_i\}$ are i.i.d. r.v.s.
- **Question:** When is an i.i.d. sequence graphical?
- **Answer:** (Arratia and Ligget ' 05) Provided $E[D_1] < \infty$,

$$\lim_{n\to\infty} P(\mathbf{D}_n \text{ is graphical}) = \begin{cases} 1/2, & \text{if } P(D_1 = \text{odd}) > 0, \\ 1, & \text{if } P(D_1 = \text{odd}) = 0. \end{cases}$$

Degree sequences

- How to construct an appropriate bi-degree-sequence?
- For the undirected case one can take $\mathbf{D}_n = \{D_1, \ldots, D_n\}$, where the $\{D_i\}$ are i.i.d. r.v.s.
- **Question:** When is an i.i.d. sequence graphical?
- **Answer:** (Arratia and Ligget ' 05) Provided $E[D_1] < \infty$,

$$\lim_{n\to\infty} P(\mathbf{D}_n \text{ is graphical}) = \begin{cases} 1/2, & \text{if } P(D_1 = \text{odd}) > 0, \\ 1, & \text{if } P(D_1 = \text{odd}) = 0. \end{cases}$$

- *Easy fix:* either resample $\mathbf{D}_n$ until its sum is even, or simply add 1 to the last node if the sum is odd.

Bi-degree-sequence

- We want a bi-degree-sequence $(\mathbf{N}, \mathbf{D})_n$ such that the $\{N_i\}$ and the $\{D_i\}$ are close to being independent sequences of i.i.d. r.v.s from distributions F and G, resp.
- The sequences must satisfy

$$\sum_{i=1}^{n} N_i = \sum_{i=1}^{n} D_i \qquad \text{for all } n.$$

- **Problem:** In general, if $\{\gamma_i\}$ and $\{\xi_i\}$ are independent i.i.d. sequences with $E[\gamma_1] = E[\xi_1]$,

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} \gamma_i = \sum_{i=1}^{n} \xi_i\right) = 0.$$

- *The "easy fix":* add a **few** in-degrees or out-degrees to match the sums.

Assumptions on the target distributions

- In-degree target distribution F.
- Out-degree target distribution G.
- Assume F and G have support on $\{0, 1, 2, \dots\}$ and have common mean $\mu > 0$.
- Suppose further that for some $\alpha, \beta > 1$,

$$\overline{F}(x) = \sum_{k>x} f_k \leq x^{-\alpha} L_F(x) \qquad \text{and} \qquad \overline{G}(x) = \sum_{k>x} g_k \leq x^{-\beta} L_G(x),$$

  for all $x \geq 0$, where $L_F(\cdot)$ and $L_G(\cdot)$ are slowly varying.

The Algorithm

1. Fix $0 < \delta_0 < 1 - \theta$, $\theta = \max\{\alpha^{-1}, \beta^{-1}, 1/2\}$.
2. Sample $\{\gamma_1, \dots, \gamma_n\}$ i.i.d. from F; let $\Gamma_n = \sum_{i=1}^n \gamma_i$.
3. Sample $\{\xi_1, \dots, \xi_n\}$ i.i.d. from G; let $\Xi_n = \sum_{i=1}^n \xi_i$.
4. Let $\Delta_n = \Gamma_n - \Xi_n$. If $|\Delta_n| \leq n^{\theta+\delta_0}$ go to step 5; otherwise go to step 2.
5. Choose randomly $|\Delta_n|$ nodes $\mathcal{S} = \{i_1, i_2, \dots, i_{|\Delta_n|}\}$ without remplacement and let

$$N_i = \gamma_i + \tau_i, \qquad D_i = \xi_i + \chi_i, \qquad i = 1, 2, \dots, n,$$

where

$$\chi_i = \begin{cases} 1 & \text{if } \Delta_n \geq 0 \text{ and } i \in \mathcal{S}, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and}$$

$$\tau_i = \begin{cases} 1 & \text{if } \Delta_n < 0 \text{ and } i \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

Some basic properties

- The parameter θ is chosen so that

$$\lim_{n\to\infty} P(|\Delta_n| \leq n^{\theta+\delta_0}) = 1.$$

- **Proposition:** $(\mathbf{N}_n, \mathbf{D}_n)$ satisfies for any fixed $r, s \in \mathbb{N}$,

$$(N_{i_1}, \ldots, N_{i_r}, D_{j_1}, \ldots, D_{j_s}) \Rightarrow (\gamma_1, \ldots, \gamma_r, \xi_1, \ldots, \xi_s)$$

  as $n \to \infty$, where $\{\gamma_i\}$ and $\{\xi_i\}$ are independent sequences of i.i.d. random variables having distributions F and G, respectively.
- $(\mathbf{N}_n, \mathbf{D}_n)$ is an approximate equivalent of the i.i.d. degree sequence for the undirected case.

Is the bi-degree-sequence graphical?

- **Theorem:** (Chen, O-C '12) The bi-degree-sequence $(\mathbf{N}_n, \mathbf{D}_n)$ satisfies

$$\lim_{n\to\infty} P\left((\mathbf{N}_n, \mathbf{D}_n) \text{ is graphical}\right) = 1.$$

Is the bi-degree-sequence graphical?

- **Theorem:** (Chen, O-C '12) The bi-degree-sequence $(\mathbf{N}_n, \mathbf{D}_n)$ satisfies

$$\lim_{n\to\infty} P\left((\mathbf{N}_n, \mathbf{D}_n) \text{ is graphical}\right) = 1.$$

- The proof uses a graphicality criterion from Berge '76.

Random pairing with $(\mathbf{N}_n, \mathbf{D}_n)$

- Does $(\mathbf{N}_n, \mathbf{D}_n)$ satisfy the regularity conditions?

Random pairing with $(\mathbf{N}_n, \mathbf{D}_n)$

- Does $(\mathbf{N}_n, \mathbf{D}_n)$ satisfy the regularity conditions?
- It can be shown that

$$\frac{1}{n}\sum_{k=1}^{n} 1(N_k = i, D_k = j) \xrightarrow{P} f_i g_j, \quad \text{for all } i, j \in \mathbb{N} \cup \{0\},$$

$$\frac{1}{n}\sum_{i=1}^{n} N_i \xrightarrow{P} E[\gamma_1], \quad \frac{1}{n}\sum_{i=1}^{n} D_i \xrightarrow{P} E[\xi_1], \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} N_i D_i \xrightarrow{P} E[\gamma_1 \xi_1],$$

and provided $E[\gamma_1^2 + \xi_1^2] < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} N_i^2 \xrightarrow{P} E[\gamma_1^2], \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} D_i^2 \xrightarrow{P} E[\xi_1^2].$$

- Therefore, if $E[\gamma_1^2 + \xi_1^2] < \infty$, the directed configuration model will produce a simple graph with probability bounded away from zero.

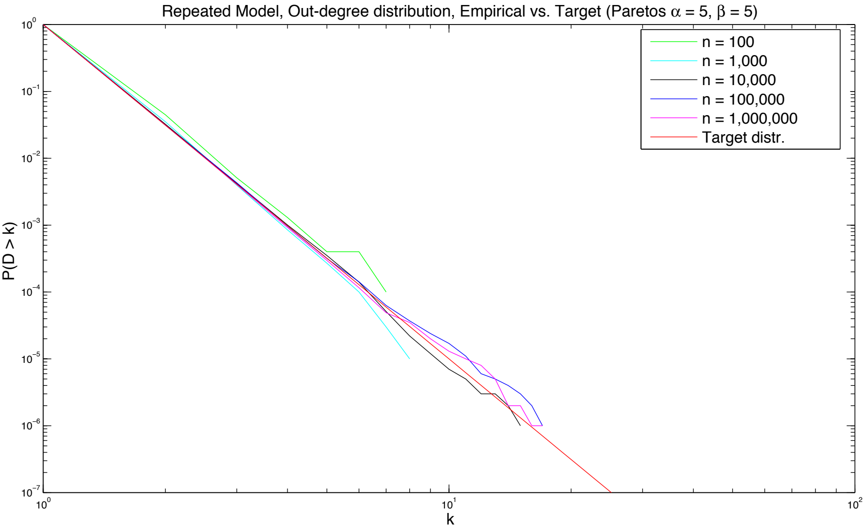
Repeated directed configuration model

- Let $N_k^{(r)}$ and $D_k^{(r)}$ be the in-degree and out-degree of node k in the resulting simple graph.
- Define for $i, j = 0, 1, 2, \ldots,$

$$h^{(n)}(i,j) = \frac{1}{n}\sum_{k=1}^{n} P(N_k^{(r)} = i, D_k^{(r)} = j),$$

$$\widehat{f_i}^{(n)} = \frac{1}{n}\sum_{k=1}^{n} 1(N_k^{(r)} = i) \quad \text{and} \quad \widehat{g_j}^{(n)} = \frac{1}{n}\sum_{k=1}^{n} 1(D_k^{(r)} = j).$$

- **Proposition:** For the repeated directed configuration model with bi-degree-sequence $(\mathbf{N}_n, \mathbf{D}_n)$ we have for all $i, j = 0, 1, 2, \ldots,$
  1. $$h^{(n)}(i,j) \to f_i g_j \qquad \text{as } n \to \infty, \quad \text{and}$$
  2. $$\widehat{f_i}^{(n)} \xrightarrow{P} f_i \qquad \text{and} \qquad \widehat{g_j}^{(n)} \xrightarrow{P} g_j, \qquad n \to \infty.$$


Repeated Model, Out-degree distribution, Empirical vs. Target (Paretos α = 5, β = 5)
n = 100
n = 1,000
n = 10,000
n = 100,000
n = 1,000,000
Target distr.
P(D > k)
k
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^0
10^1
10^2

If the regularity conditions fail...

- If $E[\gamma_1^2 + \xi_1^2] = \infty$ but parts (1)-(3) of the regularity condition hold:

  "erase all the self-lops and merge multiple edges
  in the same direction into a single edge."

- **Note:** The resulting simple graph no longer has $(\mathbf{N}_n, \mathbf{D}_n)$ as its bi-degree-sequence.
- **Question:** Do the in-degrees and out-degrees still follow the target distributions F and G?

Erased directed configuration model

- Let $N_k^{(e)}$ and $D_k^{(e)}$ be the in-degree and out-degree of node k in the resulting simple graph.
- Define for $i, j = 0, 1, 2, \ldots,$

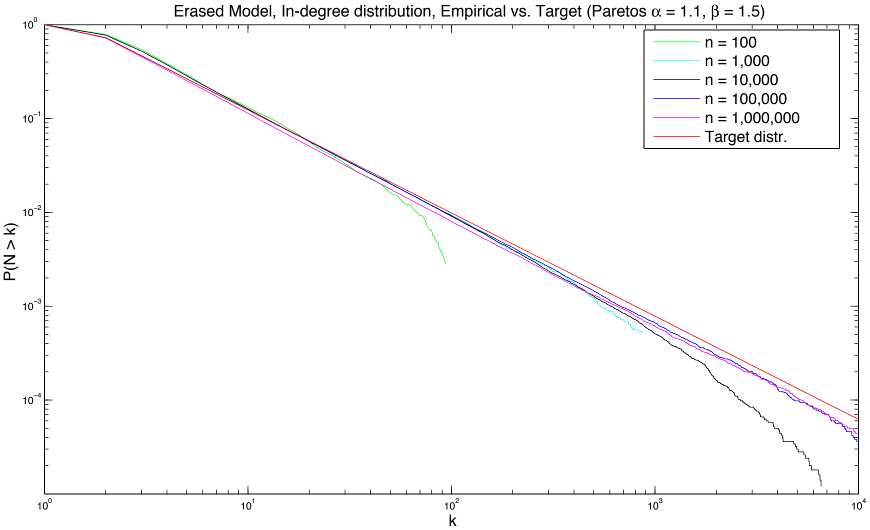
$$h^{(n)}(i,j) = \frac{1}{n}\sum_{k=1}^{n} P(N_k^{(e)} = i, D_k^{(e)} = j),$$

$$\widehat{f_i}^{(n)} = \frac{1}{n}\sum_{k=1}^{n} 1(N_k^{(e)} = i) \quad \text{and} \quad \widehat{g_j}^{(n)} = \frac{1}{n}\sum_{k=1}^{n} 1(D_k^{(e)} = j).$$

- **Proposition:** For the erased directed configuration model with bi-degree-sequence $(\mathbf{N}_n, \mathbf{D}_n)$ we have for all $i, j = 0, 1, 2, \ldots,$

1. $$h^{(n)}(i,j) \to f_i g_j \qquad \text{as } n \to \infty, \quad \text{and}$$

2. $$\widehat{f_i}^{(n)} \xrightarrow{P} f_i \qquad \text{and} \qquad \widehat{g_j}^{(n)} \xrightarrow{P} g_j, \qquad n \to \infty.$$



Erased Model, In-degree distribution, Empirical vs. Target (Paretos α = 1.1, β = 1.5)
n = 100
n = 1,000
n = 10,000
n = 100,000
n = 1,000,000
Target distr.
P(N > k)
k
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^0
10^1
10^2
10^3
10^4

Thank you for your attention.