Asymptotic Properties of the ABCD Graph Benchmark with Community Structure

Bogumił Kamiński[1], Bartosz Pankratz[1,2], Paweł Prałat[2] and François Théberge*[3]

1. SGH Warsaw School of Economics, Poland,
bkamins@sgh.waw.pl, bartosz.pankratz@ryerson.ca

2. Toronto Metropolitan University, Canada, pralat@ryerson.ca

3. Tutte Institute for Mathematics and Computing, Canada, theberge@ieee.org

NetSci 2022, July 2022

ABCD in a nutshell

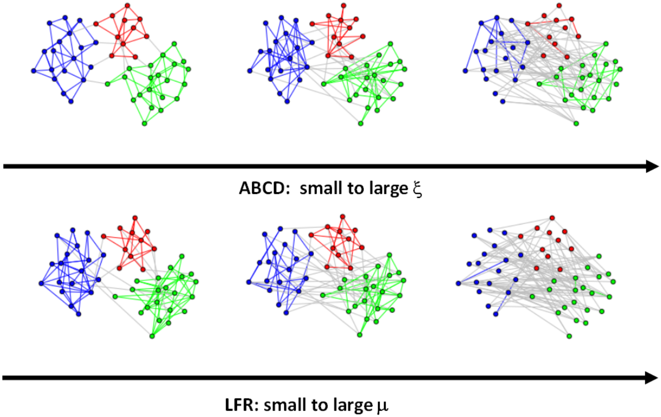
Artificial **B**enchmark for **C**ommunity **D**etection model

- power law node degree and community size distributions
  - parametrized via: (min, max, exponent)
- union of $k+1$ random graphs: k community subgraphs and one background graph (all nodes)
- parameter $\xi \in [0,1]$ controls the fraction of edges that are between communities
- graphs: **configuration** or **Chung-Lu** model
- similar properties as the **LFR** benchmark
- fast Julia code and multithreaded fork:
  `github.com/bkamins/ABCDGraphGenerator.jl`
  `github.com/tolcz/ABCDeGraphGenerator.jl`

ABCD in a nutshell

Some advantages are:

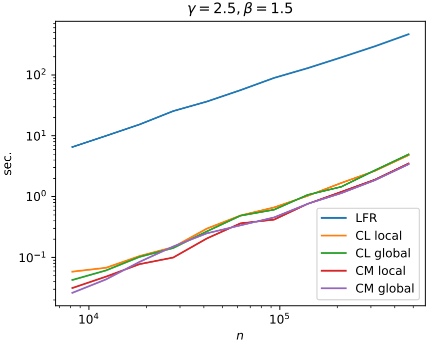
- natural interpretation of the mixing parameter ξ
  - "dimmer" from pure communities to random graph



ABCD in a nutshell

Some advantages are:

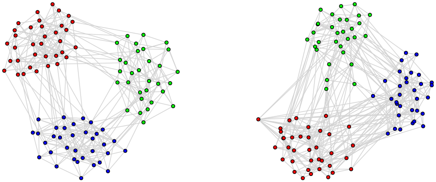
- natural interpretation of the mixing parameter ξ
  - "dimmer" from pure communities to random graph
- better scalability than LFR
  - Ref: Network Science, 9(2), 153-178 (2021)



ABCD in a nutshell

Some advantages are:

- natural interpretation of the mixing parameter ξ
  - “dimmer” from pure communities to random graph
- better scalability than LFR
  - ref: Network Science, 9(2), 153-178 (2021)
- its simplicity, which allows for **theoretical analysis**



ABCD graphs with $\xi = 0.2$ and $\xi = 0.4$

ABCD properties - degree distribution

The node degrees in **ABCD** are generated randomly following the (truncated) *power-law distribution* $\mathcal{P}(\gamma, \delta, \zeta)$ with exponent $\gamma \in (2, 3)$, minimum value δ, and maximum value $D = n^{\zeta}$ where $\zeta \in (0, 1)$.

If $X \in \mathcal{P}(\gamma, \delta, \zeta)$, then for any $k \in \{\delta, \delta + 1, \ldots, D\}$,

$$
\begin{aligned}
q_k &= \Pr(X = k) = \frac{\int_k^{k+1} x^{-\gamma} dx}{\int_\delta^{D+1} x^{-\gamma} dx} \\
&= \left(1 + \mathcal{O}(n^{-\zeta(\gamma-1)}) + \mathcal{O}(k^{-1})\right) k^{-\gamma} (\gamma - 1) \delta^{\gamma-1}.
\end{aligned}
$$

ABCD properties - degree distribution

Two related lemmas:

- an upper bound for the maximum degree; in particular we can assume: $\zeta \in (0, 1/(\gamma - 1)]$
- the degree distribution is well concentrated around the expectation

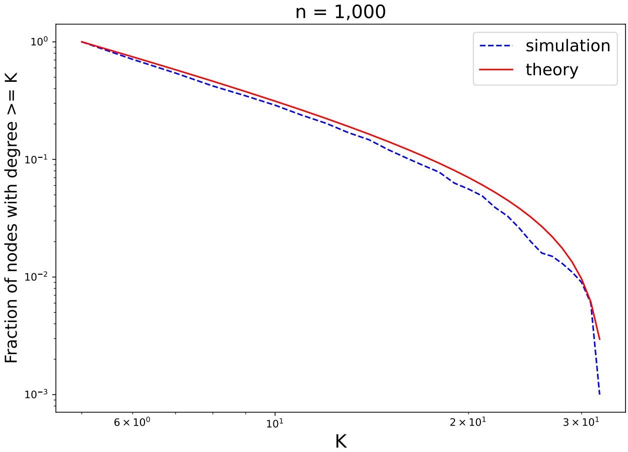
We also show the following corollary:

The volume of all nodes in an ABCD graph is *w.e.p.*[1] equal to

$$\mathrm{vol}(V) = (1 + \mathcal{O}((\log n)^{-1}))\, dn, \text{ where } d := \sum_{k=\delta}^{D} kq_k$$

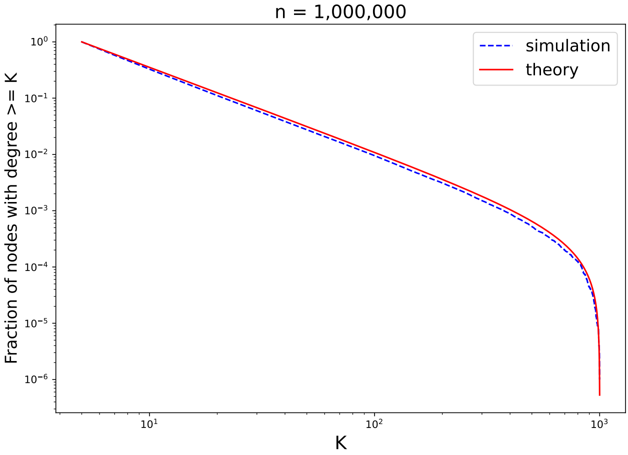
[1] with probability at least $1 - \exp(-\Omega((\log n)^2))$ where $f(n) = \Omega(g(n))$ if $g(n) = \mathcal{O}(f(n))$

ABCD properties - degree distribution



Complement of cumulative degree distribution for graphs with $\mathcal{P}(2.5, 5, 1/2)$, $n = 1,000$.

ABCD properties - degree distribution



Complement of cumulative degree distribution for graph with $\mathcal{P}(2.5, 5, 1/2)$, $n = 1,000,000$.

ABCD properties - community size distribution

Community sizes in **ABCD** are generated randomly following the (truncated) *power-law distribution* $\mathcal{P}(\beta, s, \tau)$ with exponent $\beta \in (1, 2)$, minimum value s, and maximum value $S = n^{\tau}$ with $\tau \in (\zeta, 1)$.

If $X \in \mathcal{P}(\beta, s, \tau)$, then for any $k \in \{s, s+1, \ldots, S\}$,

$$
\begin{aligned}
p_k &= \Pr(X = k) = \frac{\int_k^{k+1} x^{-\beta} dx}{\int_s^{S+1} x^{-\beta} dx} \\
&= \left(1 + \mathcal{O}(n^{-\tau(\beta-1)}) + \mathcal{O}(k^{-1})\right) k^{-\beta} (\beta - 1) s^{\beta-1}
\end{aligned}
$$

ABCD properties - community size distribution

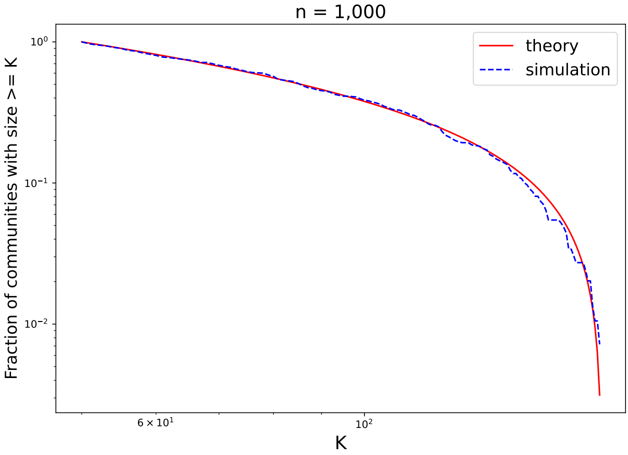
Lemma: *w.e.p.* the number of communities is equal to

$$\ell(n) = (1 + \mathcal{O}((\log n)^{-1}))\, \hat{c}\, n^{1-\tau(2-\beta)},$$

where

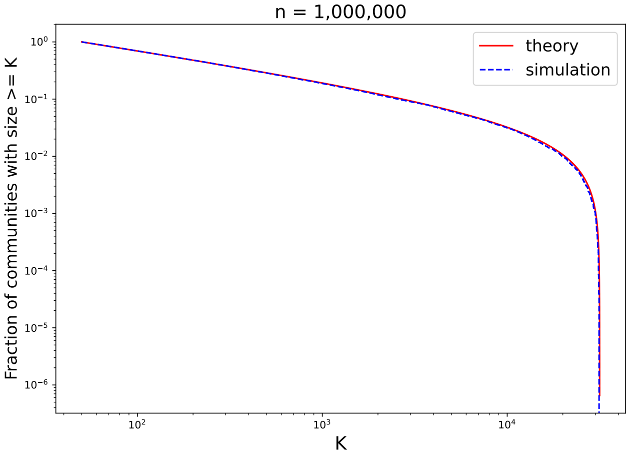
$$\hat{c} = \frac{2-\beta}{(\beta-1)s^{\beta-1}}.$$

ABCD properties - community size distribution



Complement of cumulative community size distribution for graph with $\mathcal{P}(1.5, 50, 3/4)$, $n = 1,000$.

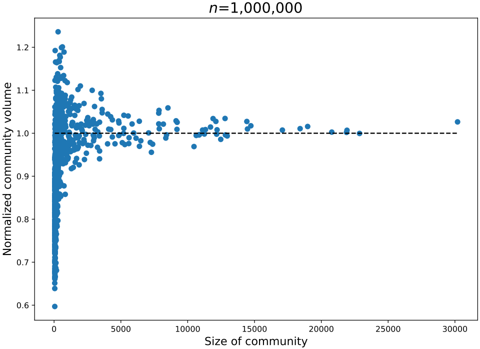
ABCD properties - community size distribution



Complement of cumulative community size distribution for graph with $\mathcal{P}(1.5, 50, 3/4)$, $n = 1,000,000$.

ABCD properties - community size distribution

We can also compare the **volume** of each community with the theoretical value. As expected, larger communities show good concentration but small ones deviate from the expectation.



ABCD properties - modularity

For a graph $G = (V, E)$ and a partition $\mathbf{A} = \{A_1, A_2, \ldots, A_\ell\}$ of V, the *modularity function* is:

$$q(\mathbf{A}) \quad = \quad \sum_{A_i \in \mathbf{A}} \frac{e(A_i)}{|E|} - \sum_{A_i \in \mathbf{A}} \left(\frac{\mathrm{vol}(A_i)}{\mathrm{vol}(V)} \right)^2$$

where $e(A) = |\{uv \in E : u, v \in A\}|$ is the *edge contribution*; $\mathrm{vol}(A) = \sum_{v \in A} \deg(v)$ is the *volume* of set A, and the second term is the expected value of the first under the Chung-Lu random null model.

ABCD properties - modularity

We now investigate the modularity function for the **ABCD** model $\mathcal{A}$.

We use notation $q^*(\mathcal{A})$ for the maximum modularity.

Theorem

Let $\mathbf{C} = \{C_1, C_2, \ldots, C_\ell\}$ be the ground-truth partition of the set of nodes of $\mathcal{A}$. Then, w.e.p.

$$q^*(\mathcal{A}) \geq q(\mathbf{C}) = \left(1 + \mathcal{O}\left((\log n)^{-(\gamma-2)}\right)\right)(1-\xi).$$

ABCD properties - modularity

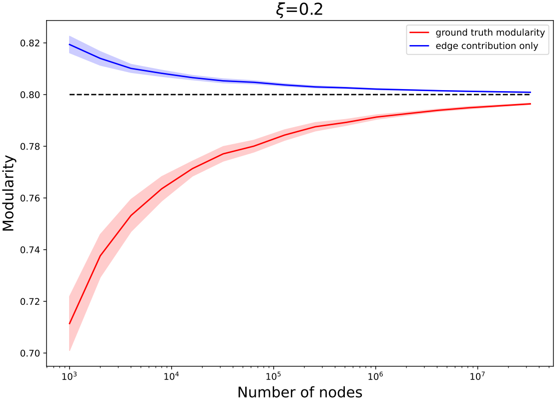


Figure: Modularity $q(\mathbf{C})$ of the ground-truth partition (red) and the corresponding edge contribution (blue) for 30 independently generated graphs. The dashed line is the asymptotic prediction; $\xi = 0.2$ and other parameters as before.

ABCD properties - modularity

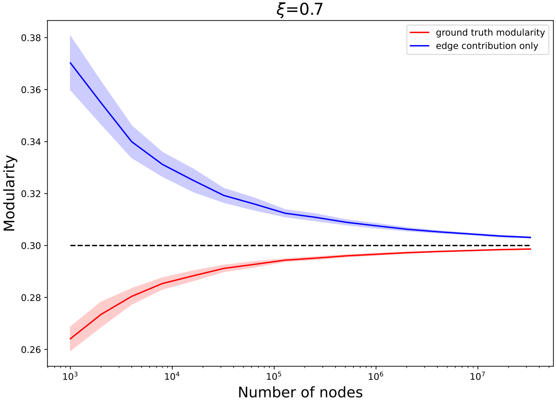


Figure: Modularity $q(\mathbf{C})$ of the ground-truth partition (red) and the corresponding edge contribution (blue) for 30 independently generated graphs. The dashed line is the asymptotic prediction; $\xi = 0.7$ and other parameters as before.

ABCD properties - modularity

We also have results comparing the **maximum** modularity $q^*(\mathcal{A})$ and the modularity of the ground truth partition $q(\mathbf{C})$.

For **noisy** graphs (large ξ), we show then we can find a partition with larger modularity than the ground-truth partition!

For **small values** of ξ, we show that *w.h.p.*[2]
$q^*(\mathcal{A}) \sim q(\mathbf{C}) \sim 1 - \xi$, provided δ is large enough.
We also show that this is not true when $\delta = 1$.

---

[2] with probability tending to 1 as $n \to \infty$

ABCD properties - modularity

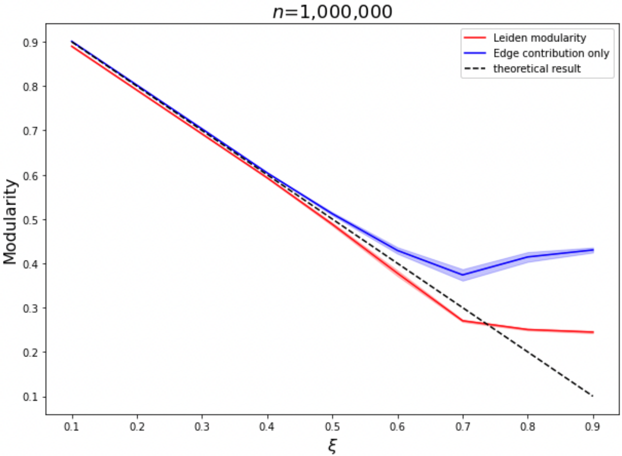


Figure: The modularity $q(\mathbf{C})$ obtained with Leiden (red) and the corresponding edge contribution (blue) for 30 independently generated graphs. The dashed line corresponds to the asymptotic prediction for the ground-truth. Same parameters as before.

References

ABCD benchmark:

- *Artificial benchmark for community detection (ABCD) - Fast random graph model with community structure*, Network Science, 1-26 (2021)
- *Properties and Performance of the ABCDe Random Graph Model with Community Structure*, arXiv:2203.14899 (2022)
- `github.com/bkamins/ABCDGraphGenerator.jl`
- `github.com/tolcz/ABCDeGraphGenerator.jl`

Pre-print for this work:

- *Modularity of the ABCD random graph model with community structure*, arXiv:2203.01480 (2022)

In the works:

- ABCD with outliers
- Hypergraph-ABCD