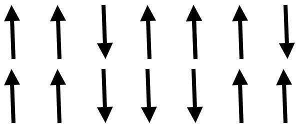
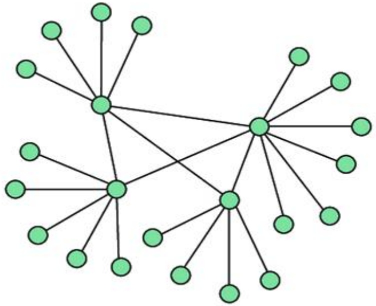
Intrinsic Dimension for discrete data

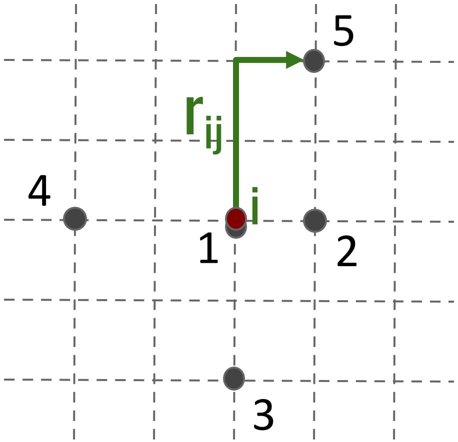
Antonietta Mira

Università della Svizzera italiana, Lugano

Discrete data spaces are ubiquitous







GAAGGTCTTCGGAT
GAAGGTTTTCGGAT
GACGGCCTTCGGGT

- Natural metric ≠ L^2 $\rightarrow$ L^1 / Hamming / Edit
- Multiplicity (repetitions in the data): $r_1 = 0$
- Degeneracy (many equidistant points): $r_3 = r_4$

Intrinsic Dimension for Discrete Data = I3D

Dimension of a (hyper) cubic lattice where the original data points can be (locally) projected without information loss

Macocco et al.
Phys. Rev. Lett. 2023

Thanks for slides

Overview

- Derivation of I3D
- Compare with benchmarks on fractals

ID for unweighted networks

- ID signature as summary statistics in ABC
- ID-based generative model

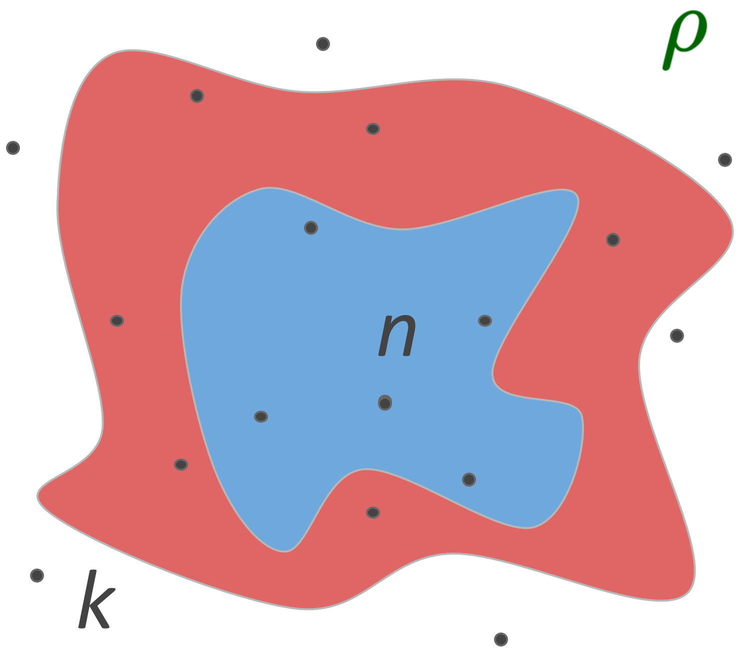
How many data points fall within a volume V if the density is constant? Poisson distributions as in BIDE

$$\mathcal{P}(k\,|V_2) = \frac{(\rho V_2)^k}{k!} e^{-\rho V_2}$$

$$\mathcal{P}(n\,|V_1) = \frac{(\rho V_1)^n}{n!} e^{-\rho V_1}$$

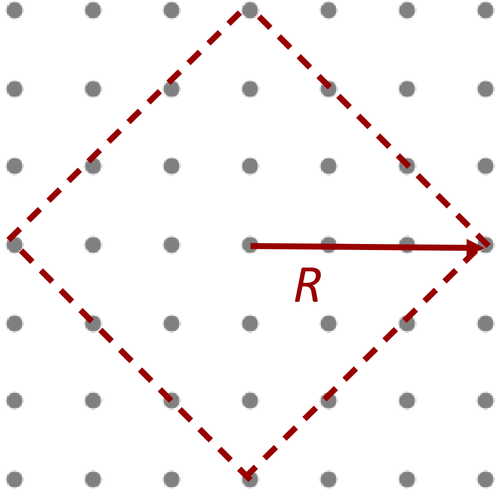
$$n\,|\,k \sim \mathrm{Binomial}(k, p)$$

$$p = \frac{\cancel{\rho} V_1}{\cancel{\rho} V_2}$$



[6] Moltchanov, D. Distance distributions in random networks. Ad Hoc Networks10, (2012)

Measuring the volume in discrete spaces: enumerate the lattice points with Ehrhart polynomials (1977)



$$V(R, d) = \sum_{i=0}^{d} \binom{d}{i} \binom{R-i+d}{d}$$ [3]

$R = 0$: 1

$R = 1$: $1 + 2d$

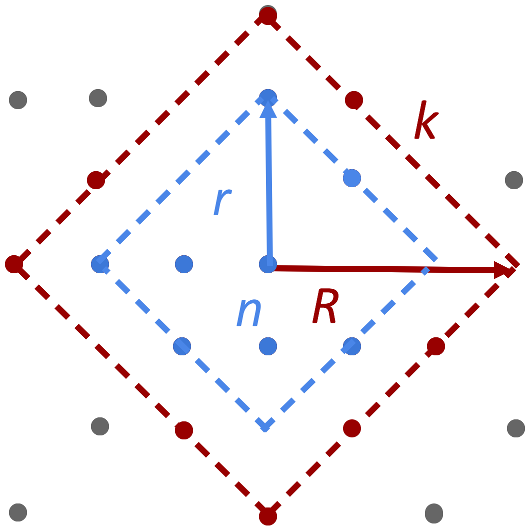
$R = 2$: $1 + 2d + 2d^2$

$R = 3$: $1 + \frac{8}{3}d + 2d^2 + \frac{4}{3}d^3$

$R = 4$: $1 + \frac{8}{3}d + \frac{10}{3}d^2 + \frac{4}{3}d^3 + \frac{2}{3}d^4$

[7] E. Ehrhart, International Series of Numerical Mathematics, Vol.35 (1977).
[8] Beck, M. & Robins, S. Computing the continuous discretely: integer-point enumeration in polyhedra. Choice Rev. 45–0923 (2007)

Poisson process on lattices



$$n \mid k \sim \mathrm{Binomial}(k, p)$$

$$p = V(r, d)/V(R, d)$$

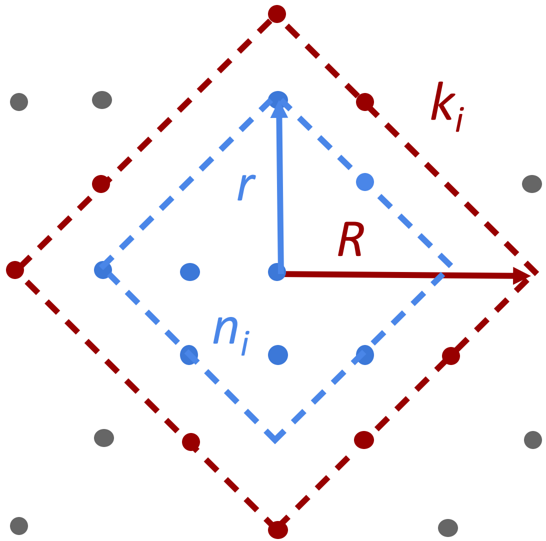
$$V(r, d) = \sum_{i=0}^{d} \binom{d}{i} \binom{r-i+d}{d}$$

$$V(R, d) = \sum_{i=0}^{d} \binom{d}{i} \binom{R-i+d}{d}$$

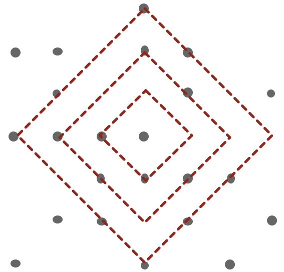
ID estimate through MLE → I3D

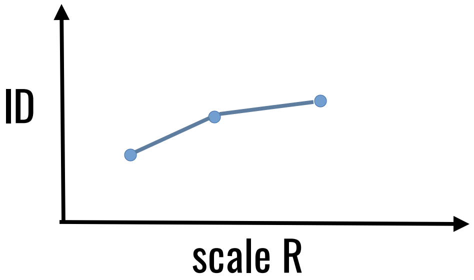
$$\mathcal{L}(d\,|\{n_i, k_i\}) = \prod_i^N \binom{k_i}{n_i} \left(\frac{V(r,d)}{V(R,d)}\right)^{n_i} \left(1 - \frac{V(r,d)}{V(R,d)}\right)^{k_i - n_i}$$

$$\frac{\partial \ln(\mathcal{L})}{\partial d} = \frac{V(r,d)}{V(R,d)} - \frac{\langle n \rangle}{\langle k \rangle} = 0 \longrightarrow \hat{d}$$



Relevant feature:
explicit scale selection by changing *R*





[9] **Macocco** et al. "Intrinsic dimension estimation for discrete metrics." Physical Review Letters 130.6 (2023)

Bayesian approach gives an analytical error estimate

$$\mathcal{L} = \mathcal{B}(n, k \,|\, p) \longrightarrow \mathcal{P} = \mathscr{B}(p \,|\, \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \qquad p = \frac{V_1}{V_2} = r^d$$

Beta posterior (of p) parameters:

$$\alpha_f = \alpha + \sum_i^N n_i$$

$$\beta_f = \beta + \sum_i^N (k_i - n_i)$$

Posterior of d

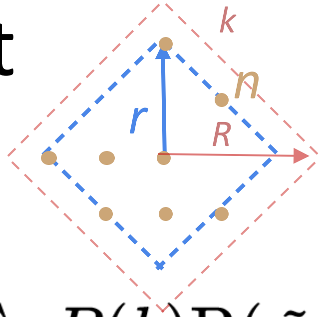
$$P(d) = \mathscr{B}(r^d \,|\, \alpha, \beta)\, r^d |\ln r|$$

$$\langle d \rangle = \frac{\psi_0(\alpha) - \psi_0(\alpha+\beta)}{\ln r}$$

$$\mathrm{Var}(d) = \frac{\psi_1(\alpha) - \psi_1(\alpha+\beta)}{(\ln r)^2}$$

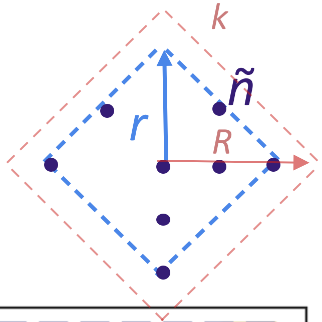
Minimize asymptotic variance: $r_{opt} \sim 0.2^{\,1/d}$

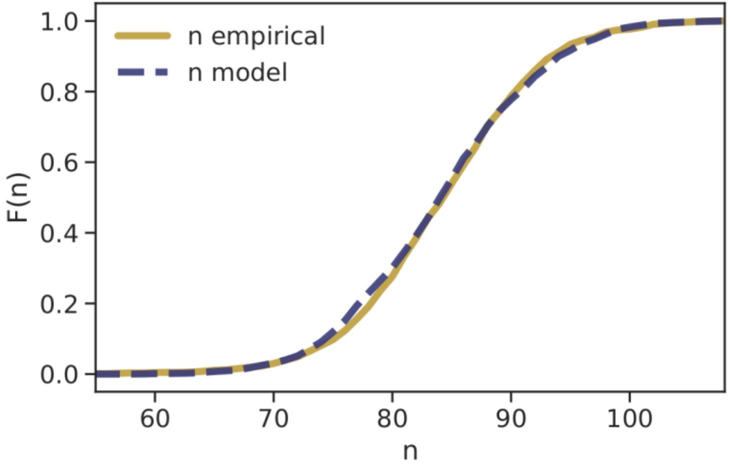
Model validation test

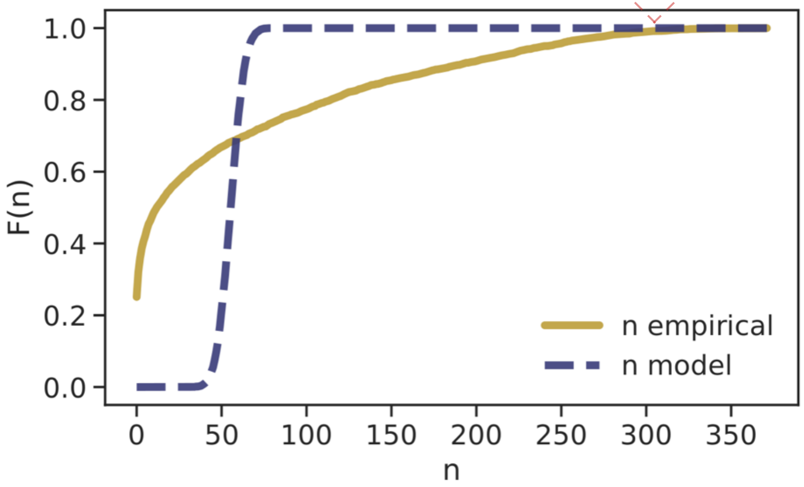


empirical distribution $P(n)$

theoretical distribution $P(\tilde{n}) = \sum_k P(k)\mathrm{B}(\,\tilde{n}; k, p(\hat{d})\,)$





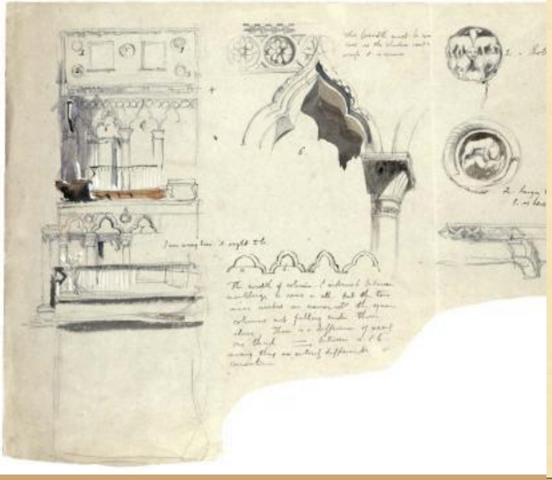


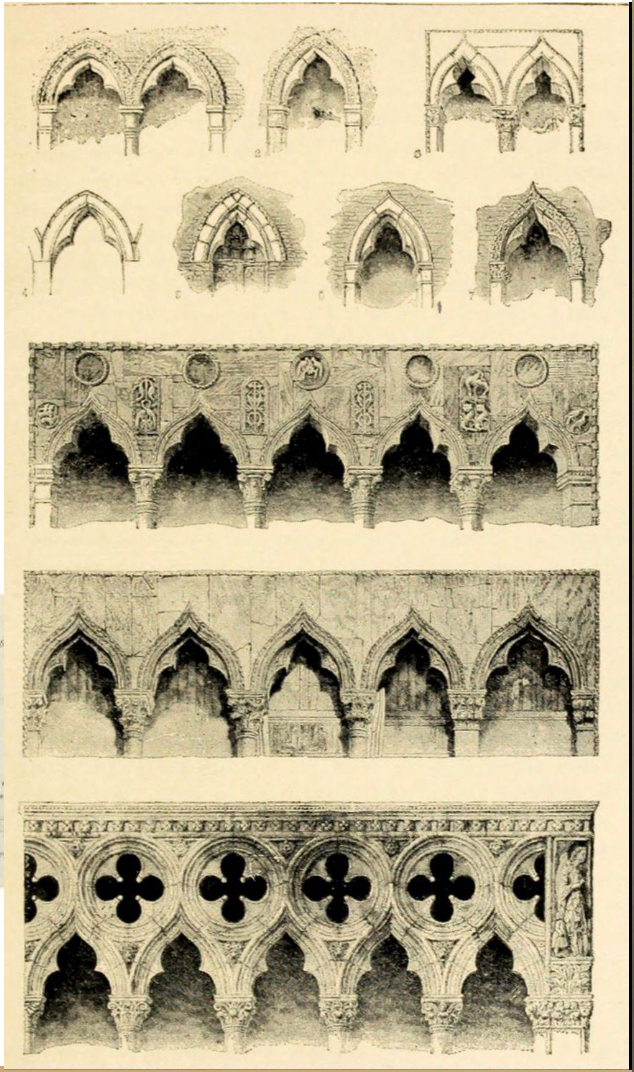
Overview

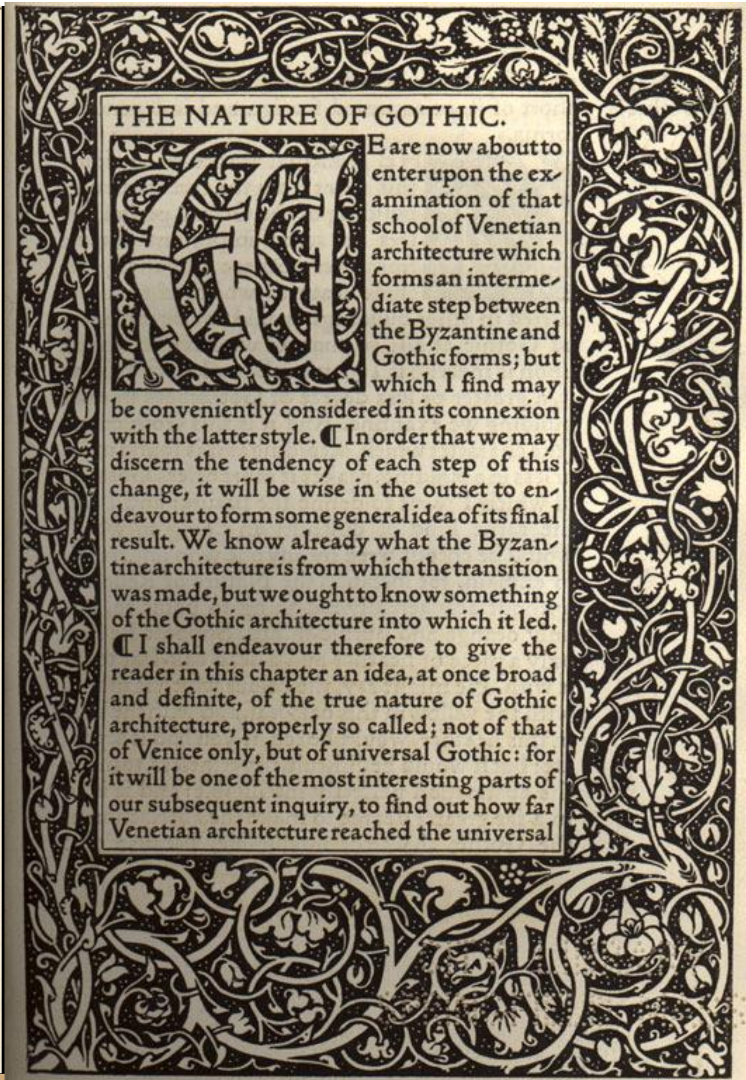
- Building the ID estimator for Discrete Datasets (I3D)
  - Derivation
  - Benchmarks on fractals
- ID for unweighted networks
  - ID signature and comparison with other fractal methods
  - ID signature as summary statistics for generative models
  - ID-based generative model

John Ruskin
The Stones of Venice
1851 - 1853

Architecture of the Venetian Byzantine, Gothic and Renaissance periods





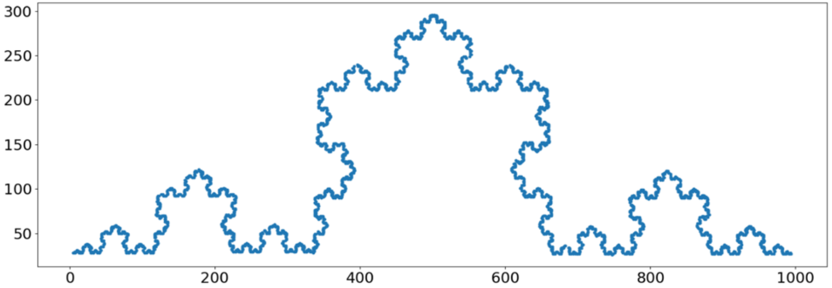


THE NATURE OF GOTHIC.

WE are now about to enter upon the examination of that school of Venetian architecture which forms an intermediate step between the Byzantine and Gothic forms; but which I find may be conveniently considered in its connexion with the latter style. ¶ In order that we may discern the tendency of each step of this change, it will be wise in the outset to endeavour to form some general idea of its final result. We know already what the Byzantine architecture is from which the transition was made, but we ought to know something of the Gothic architecture into which it led. ¶ I shall endeavour therefore to give the reader in this chapter an idea, at once broad and definite, of the true nature of Gothic architecture, properly so called; not of that of Venice only, but of universal Gothic: for it will be one of the most interesting parts of our subsequent inquiry, to find out how far Venetian architecture reached the universal

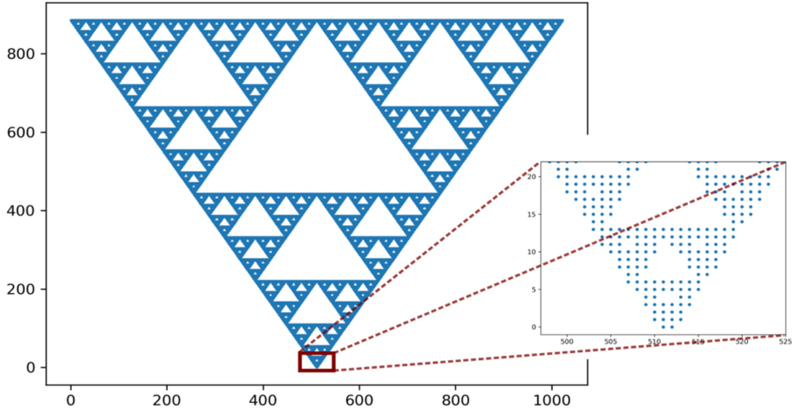
Behaviour of different estimators on geometrical fractals

Koch snowflake



ID=log(4)/log(3) ~ 1.3

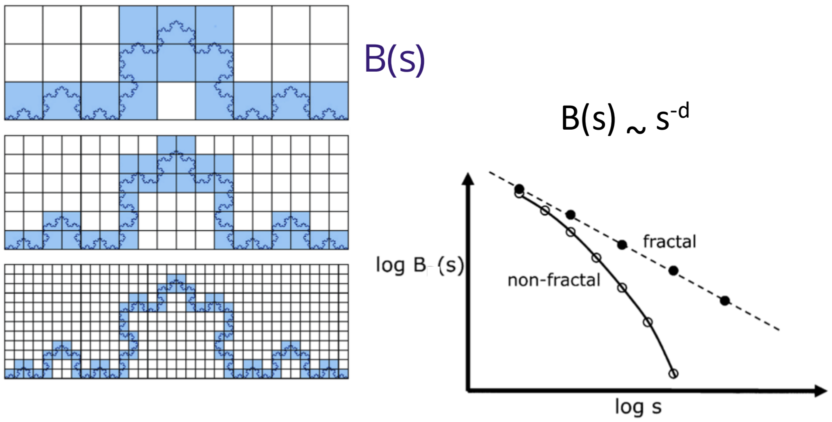
Sierpinski gasket



ID=log(3)/log(2) ~ 1.6

Methods presently used for discrete spaces

Box-Counting

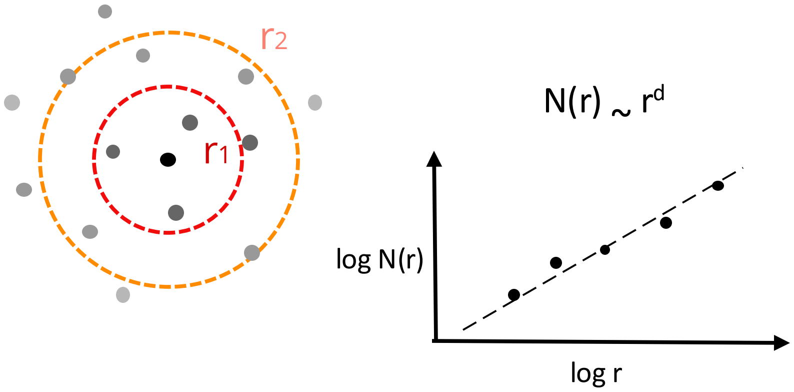


Limitations:

➔ Computationally demanding in high d

[3] K. Falconer, Fractal geometry: mathematical foundations and applications, J. Wiley & Sons (2004)
[4] A. Block, W. von Bloh, and H. J. Schellnhuber, Phys. Rev. A 42, 1869 (1990)

Fractal Dimension



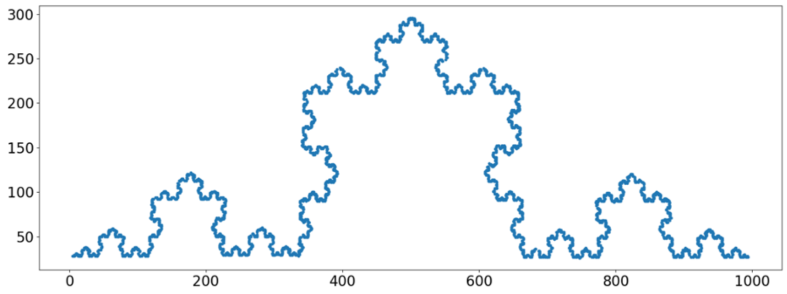
Limitations:

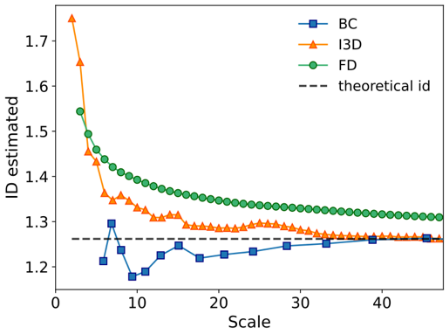
➔ No particular adaptation for discrete spaces

[5] L. Niemeyer, L., Pietronero, & H.J. Wiesmann, Fractal dimension of dielectric breakdown. Physical Review Letters, 52(12), 1033 (1984).

Behaviour of different estimators on geometrical fractals

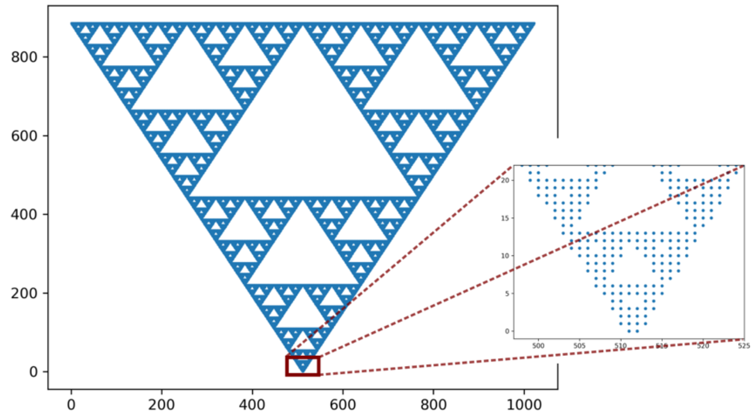
Koch snowflake

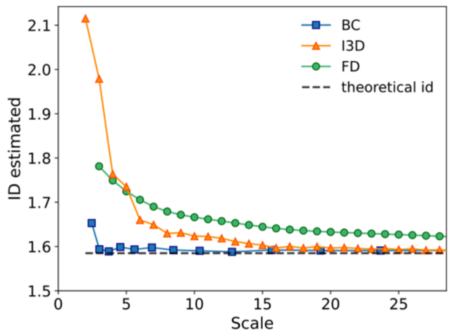




ID=log(4)/log(3) ~ 1.3

Sierpinski gasket





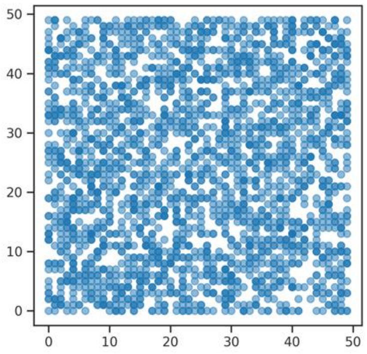
Box counting (BC)

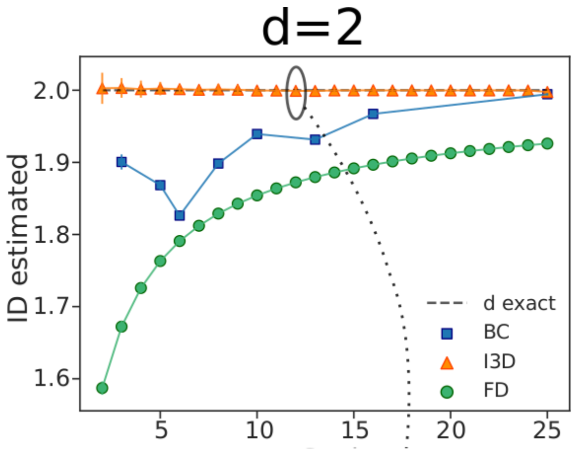
I3D

Fractal dimension (FD)

ID=log(3)/log(2) ~ 1.6

Uniform distribution on square lattice



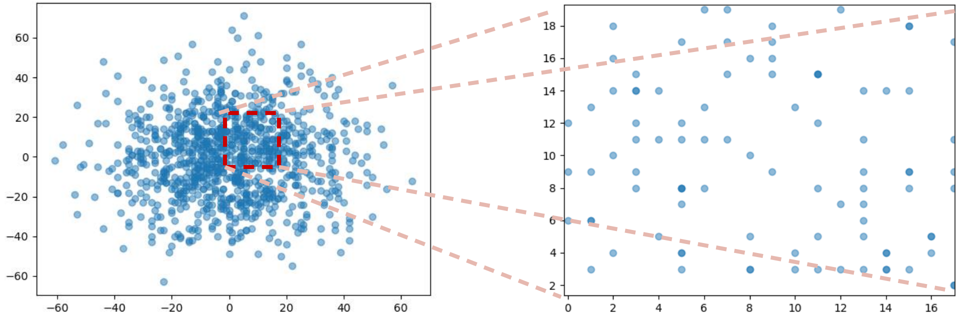


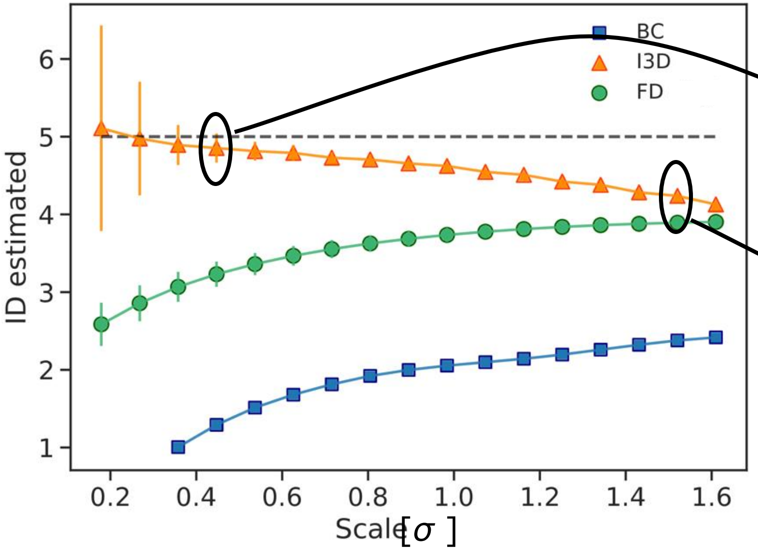
Box counting (BC)

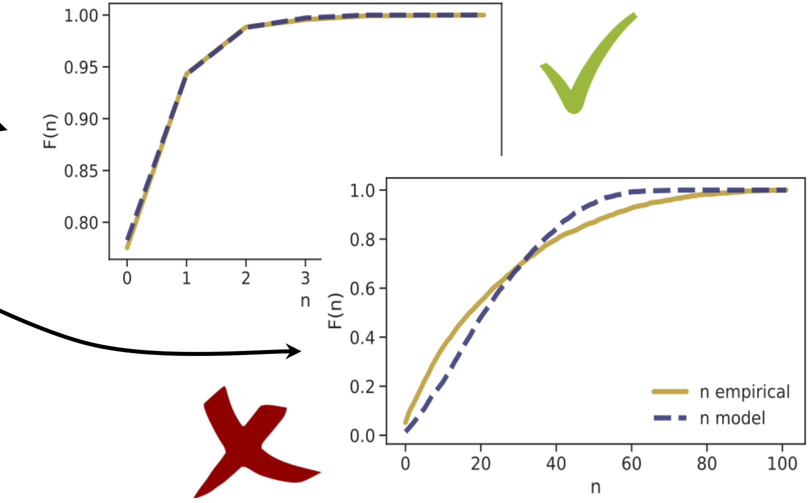
I3D

Fractal dimension (FD)

5d Gaussian lattice distribution







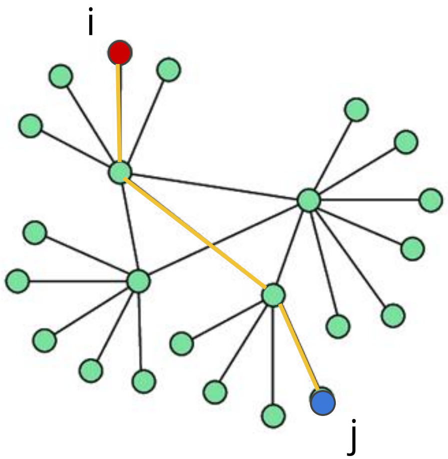
model begins to be inaccurate at this scale: non-constant density within the selected volume

Overview

- Building the ID estimator for Discrete Datasets (I3D)
  - Derivation
  - Benchmarks on fractals
- ID for unweighted networks
  - ID signature and comparison with other fractal methods
  - ID signature as summary statistics for generative models
  - ID-based generative model

Distances on unweighted networks are discrete!

G(N,E)



N = number of nodes

E = number of edges

A_{ij} = adiacency matrix

Directed

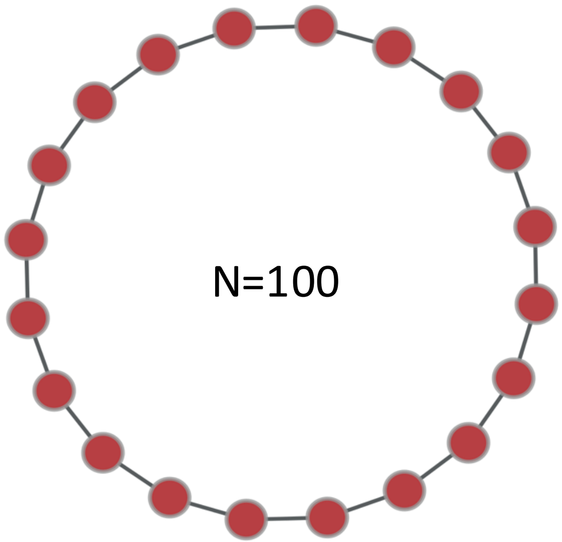
Undirected $A_{ij} = A_{ji}$

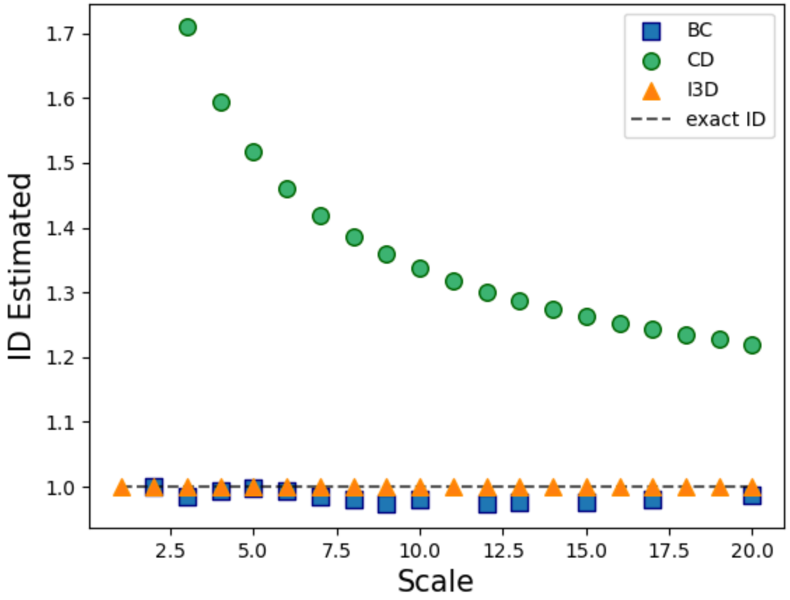
Weighted

Unweighted $A_{ij} = \{0,1\}$

$d(i,j)$ = shortest path $\in \mathbb{N}$

I3D is stable and finds the proper ID





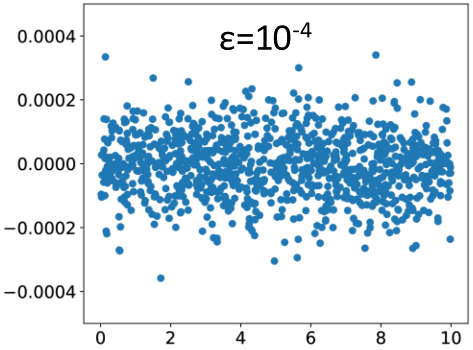
I3D computational complexity lies in the calculation of distances O(N(N+E))

C

ID signature for 1d graph

Local structure at small scales: $ID_1(G) = \delta/2$

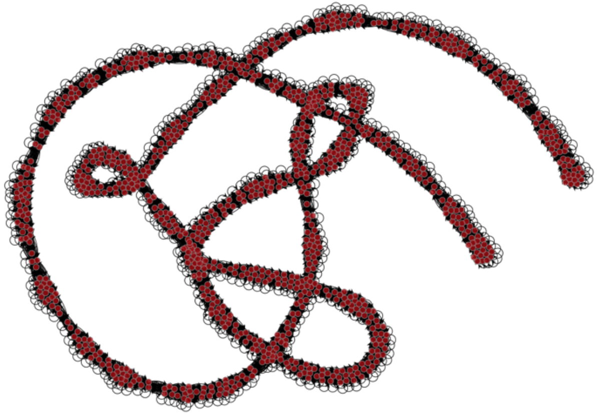
Noisy line in 100D

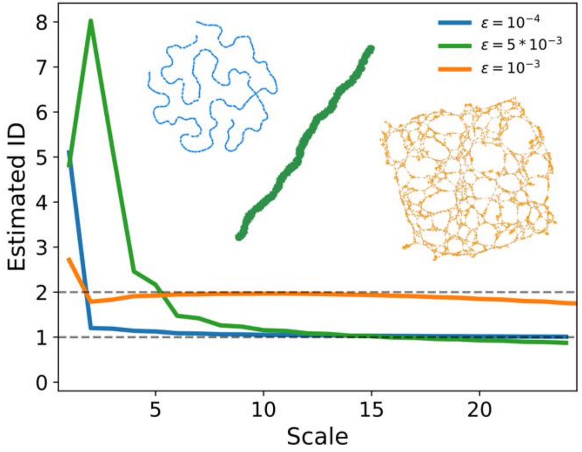


I3D

G(n,e) with ~constant degree δ

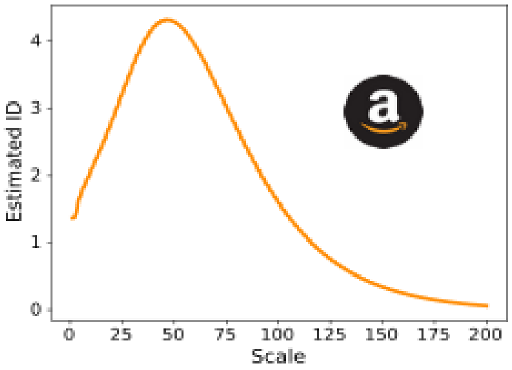
link between first δ=10 neighbors

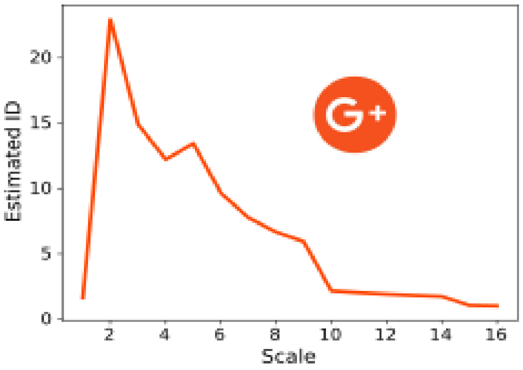


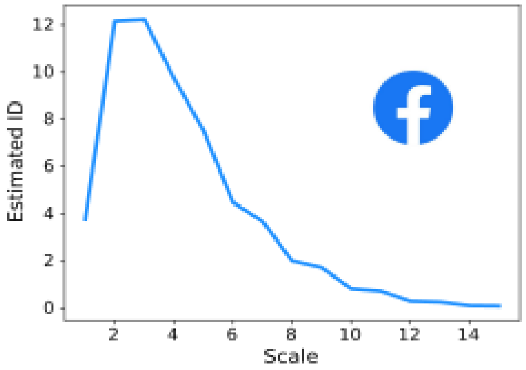


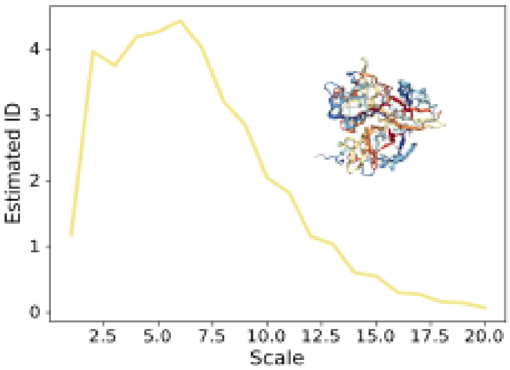
Information on the Non-local structure by analyzing the ID at the meso-scale

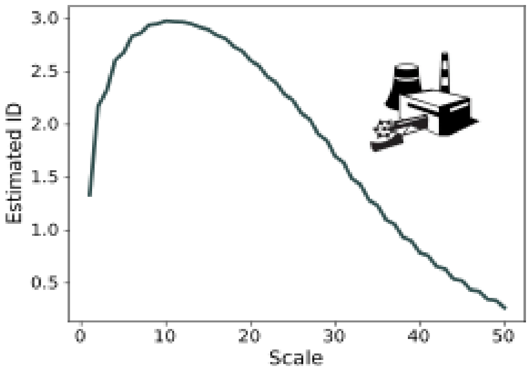
ID signature for real world networks











scale -> diameter
ID -> zero

Overview

- Building the ID estimator for Discrete Datasets (I3D)
  - Derivation
  - Benchmarks on fractals
- ID for unweighted networks
  - ID signature and comparison with other fractal methods
  - ID signature as summary statistics for generative models
  - ID-based generative model

Typical summaries are local or global

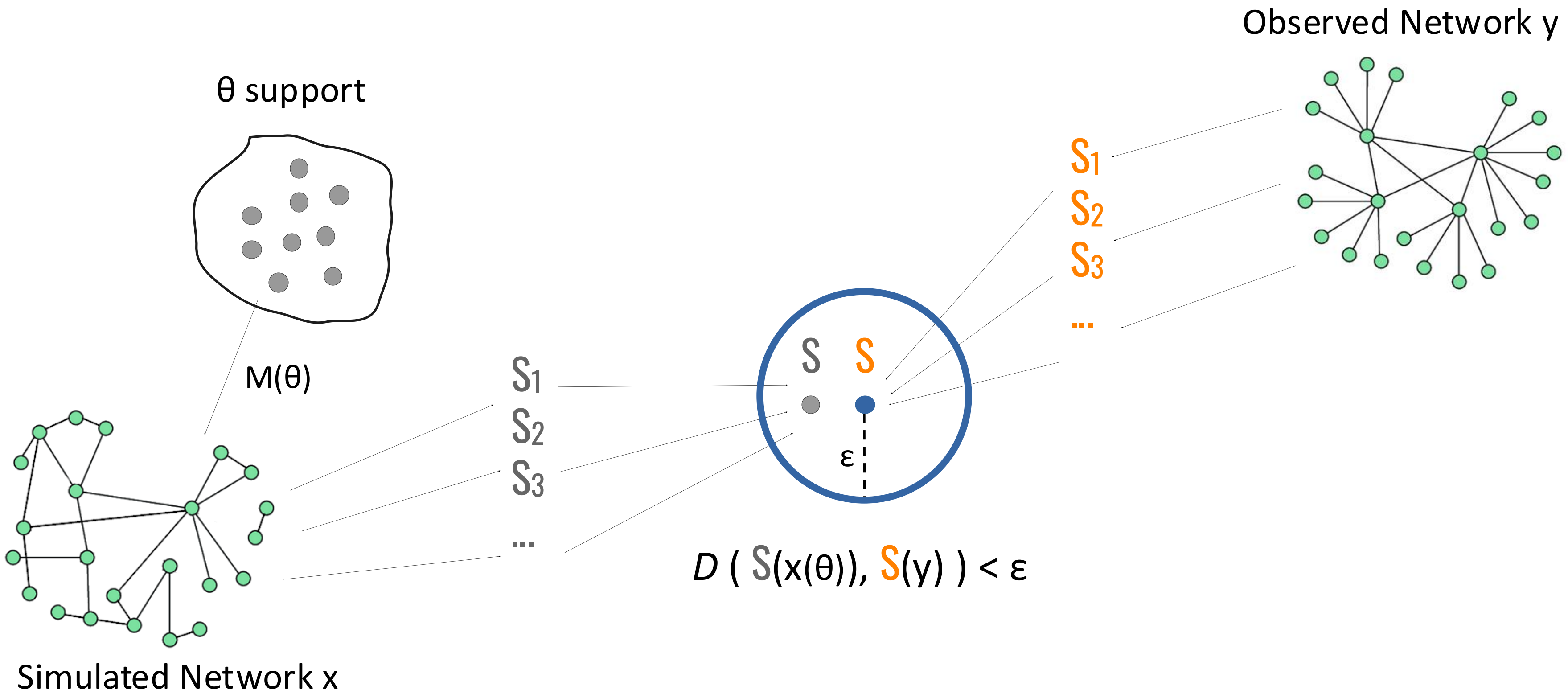
Local observables:

- degree distribution
- clustering coefficient
  #(Δ)/#(Δ+Λ)

Global observables:

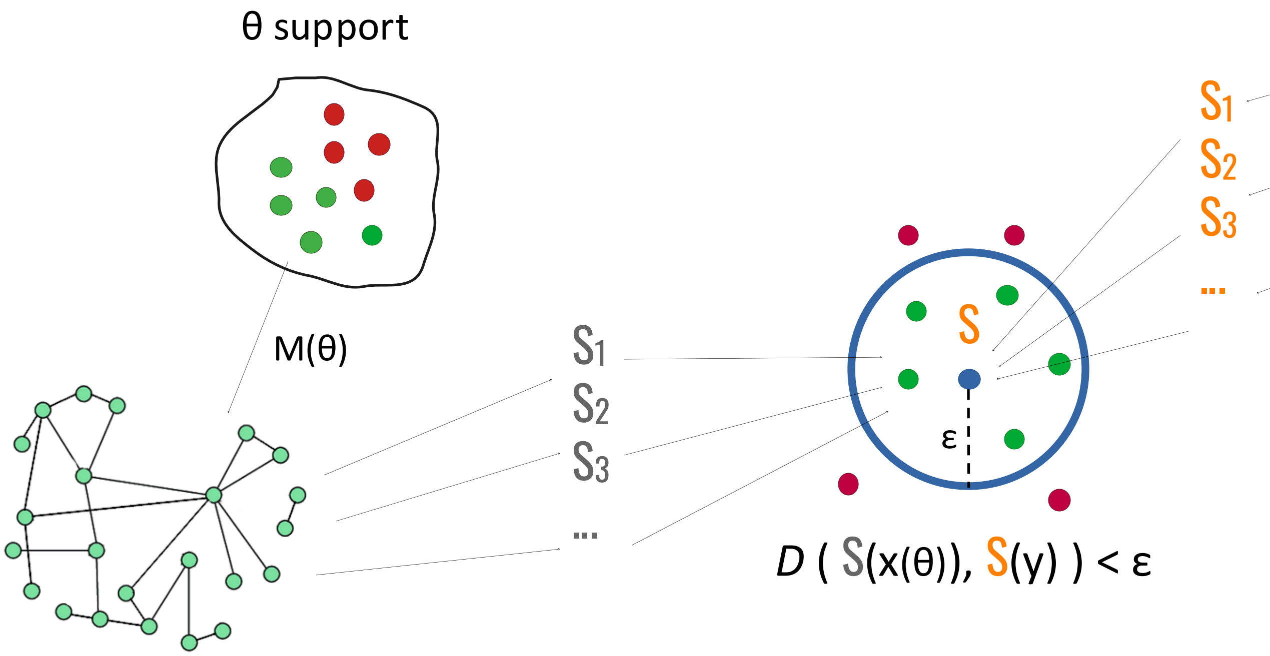
- diameter
  $\max_{ij}\{ d(i,j) \}$
- modularity

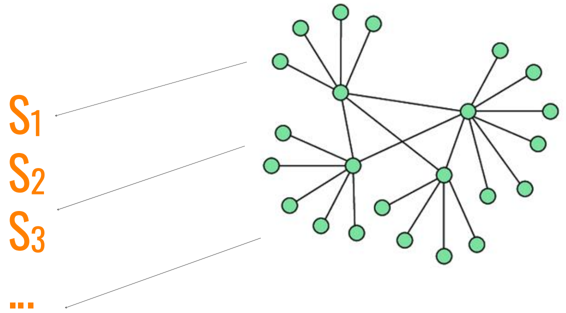
Generative models have intractable LHD → ABC



Generative models have intractable LHD → ABC

Observed Network y





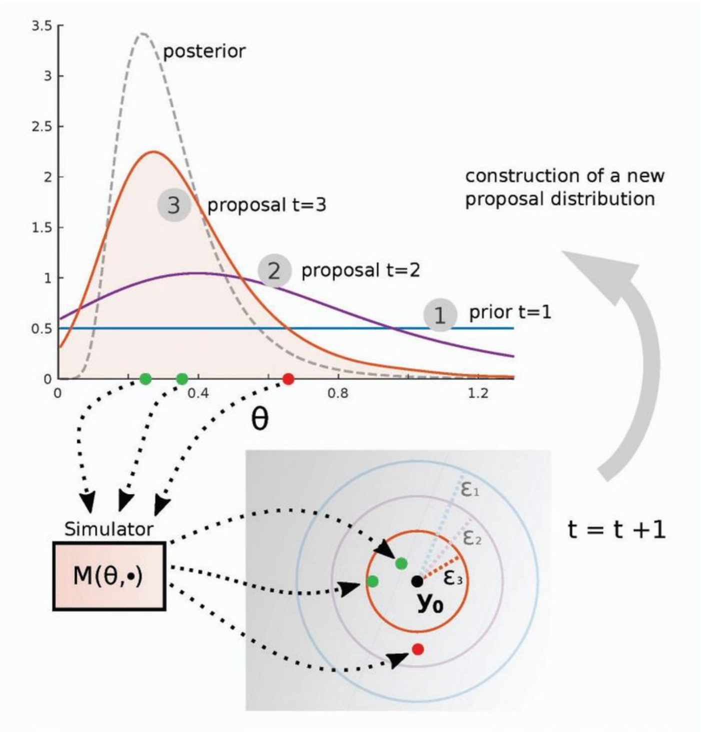
Synthesized Network x

$D\ (\ S(x(\theta)),\ S(y)\ )\ <\ \varepsilon$

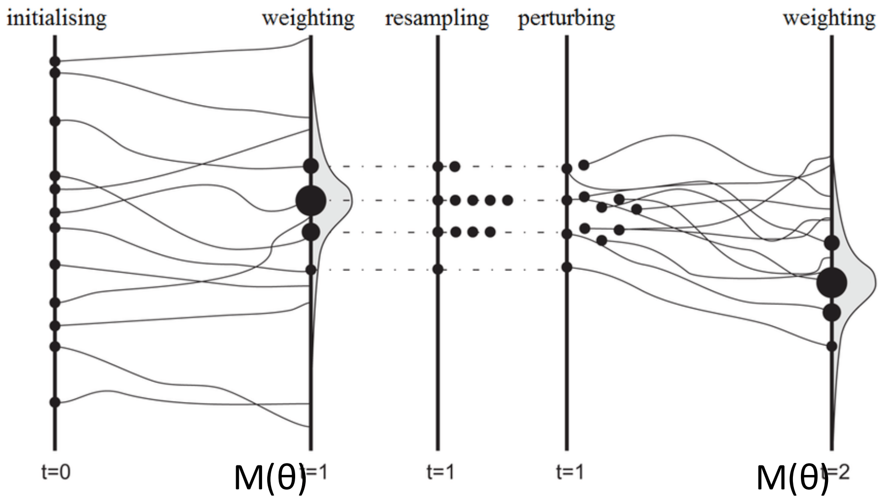
- degree distribution
- local clustering coeff
- closeness
- ...

$$D\ (\ S(x(\theta)),\ S(y)\ )\ =\ \max_{R\in\{1,\ldots,\mathrm{diam}(y)\}} \left|\ ID_R(x(\theta))\ -\ ID_R(y)\ \right|$$

Sequential Monte Carlo - ABC



prior t=0 posterior t=1 prior t=1 posterior t=2

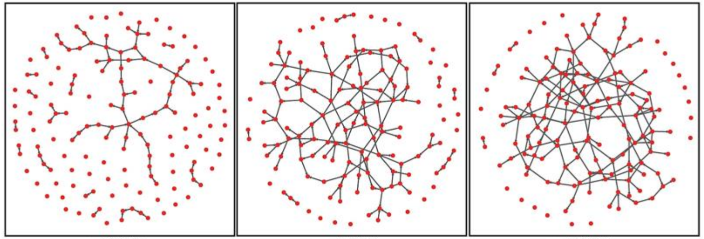


M(θ) + accept/reject M(θ) + accept/reject

[23] Sisson, Scott A., Yanan Fan, and Mark M. Tanaka. "Sequential monte carlo without likelihoods." PNAS 104.6 (2007) [24] Beaumont, Mark A., et al. "Adaptive approximate Bayesian computation." Biometrika 96.4 (2009)

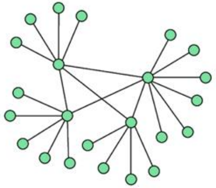
Erdös-Rényi

Each edge is established with probability p



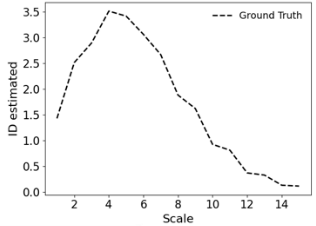
N=300 p_{gt}=0.01

ER



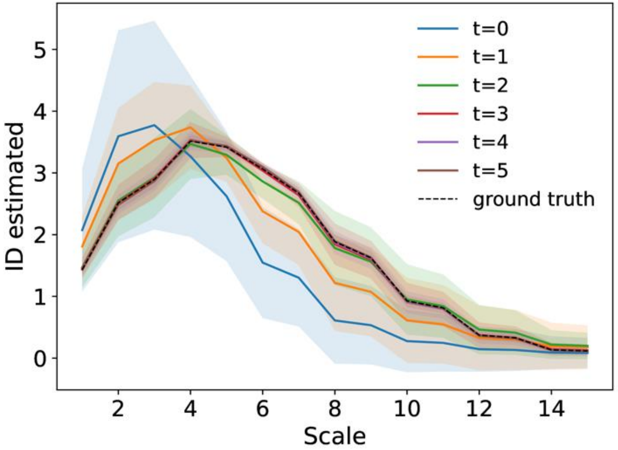
observed network

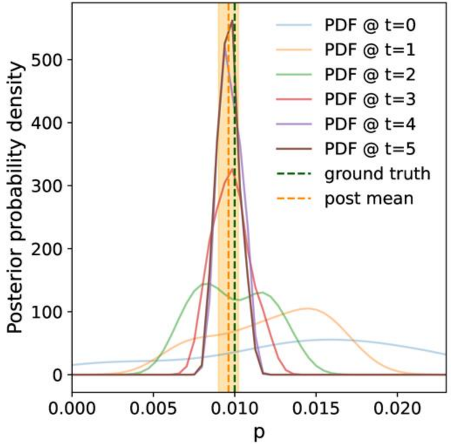
I3D

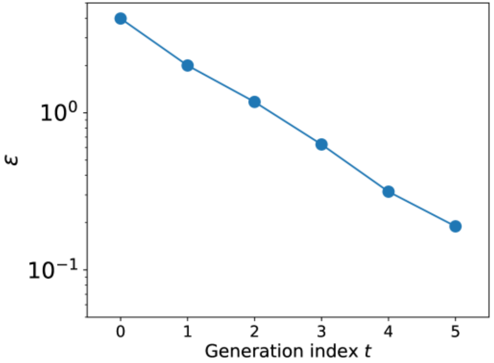


reference ID

π(p)=U(0, 0.025)

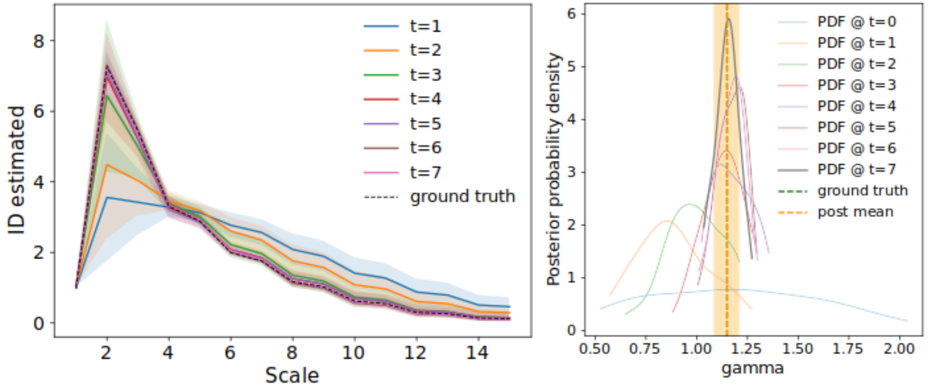




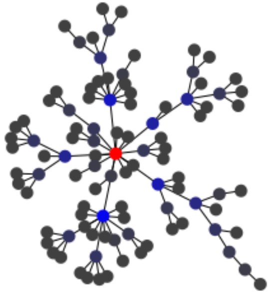


[25] P. Erdős, A. Rényi, et al., “On the evolution of random graphs,” Publ. Math. Inst. Hung. Acad. Sci, vol. 5, no. 1, pp. 17–60, 1960.

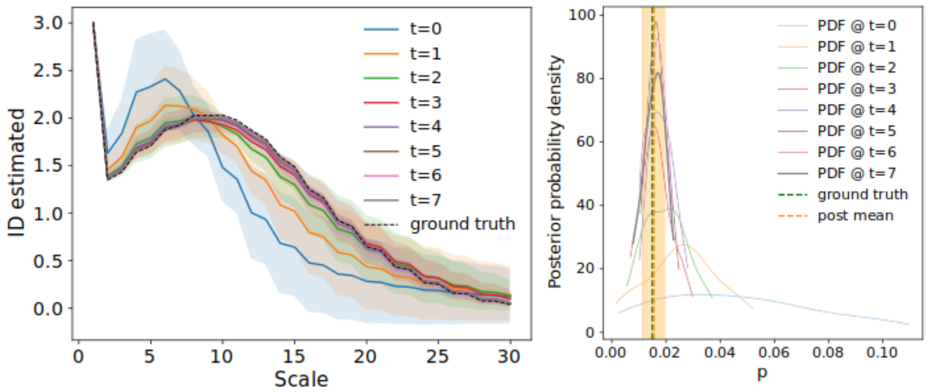
Non-Linear Preferential Attachment

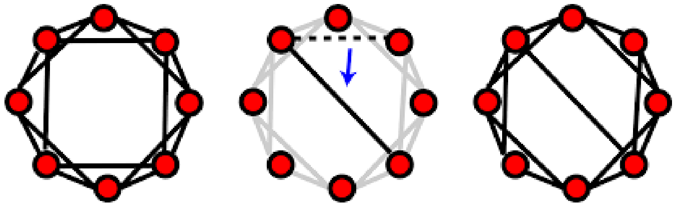


The new node is wired to
m = int(E/N) existing nodes
according to $p_i = k^{\gamma}_i / \Sigma_j k_j^{\gamma}$



Watts-Strogatz

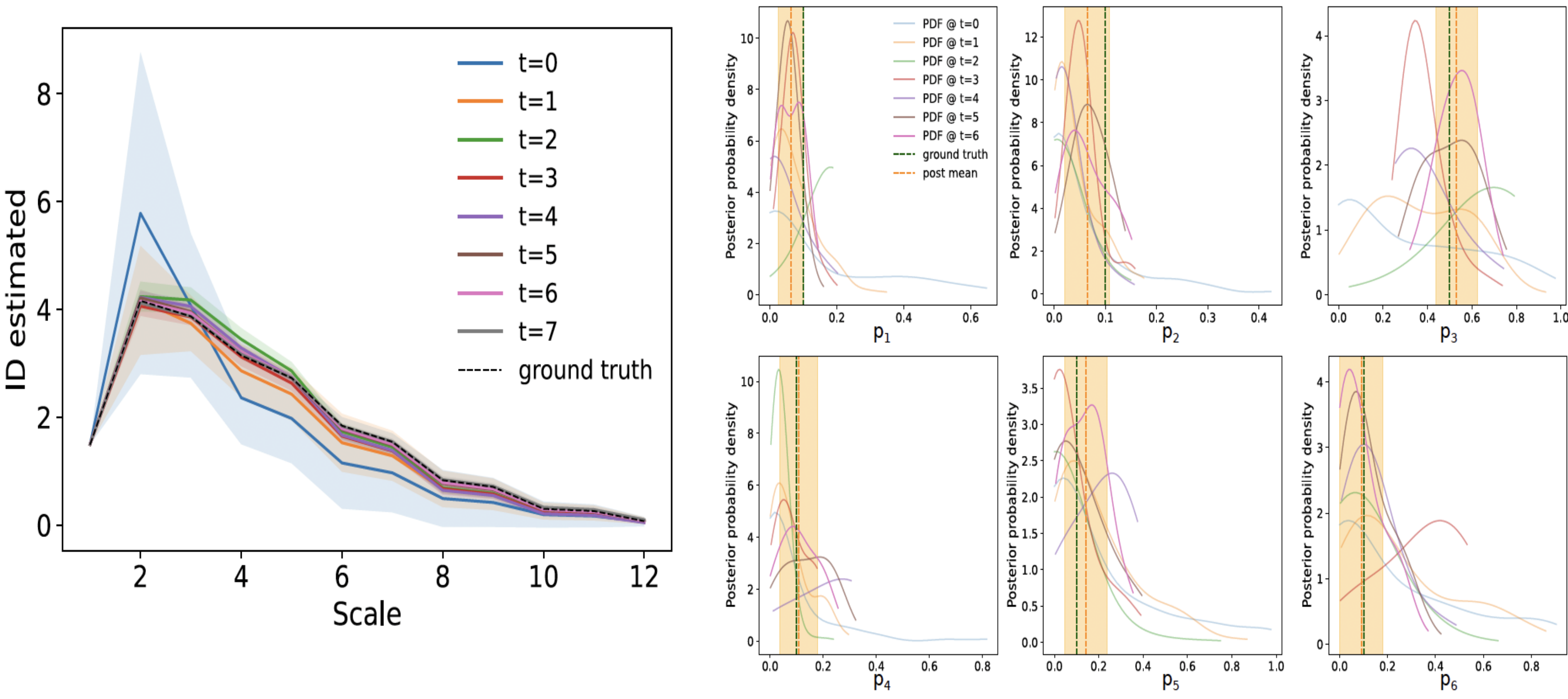




Create a ring network with k = int(2*E/N) nearest connections
Rewire each edge with prob p

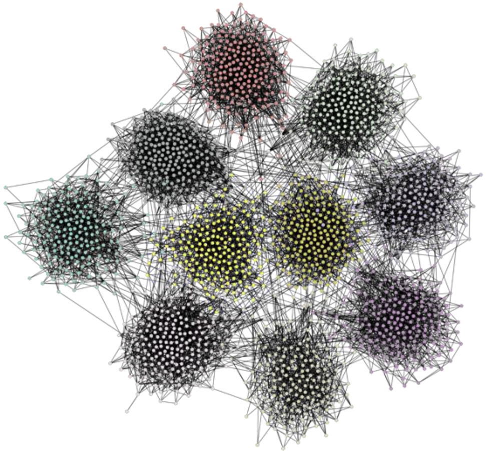
[26] L. Krapivsky, S. Redner, and F. Leyvraz, “Connectivity of growing random networks,” Phys. Rev. Lett., vol. 85, pp. 4629–4632, 21 Nov. 2000.
[27] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’networks,” Nature, vol. 393, no. 6684, pp. 440–442, 1998

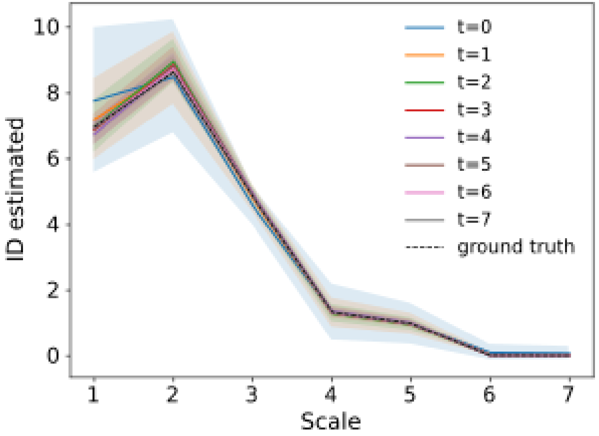
Action Based Network Generator: 6 parameters

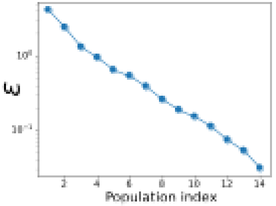


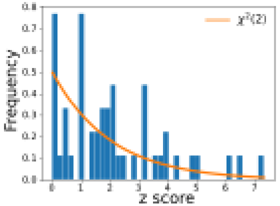
Planted Partition (PP): building communities

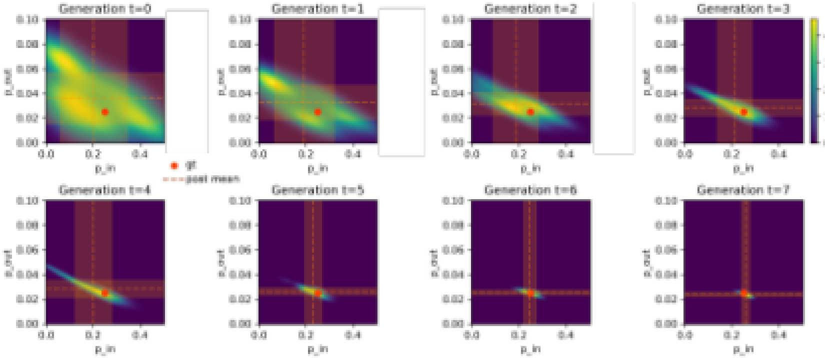
- number of communities *L*
- nodes per community *k*
- conn prob within community: p_{in}
- conn prob outside community: p_{out}











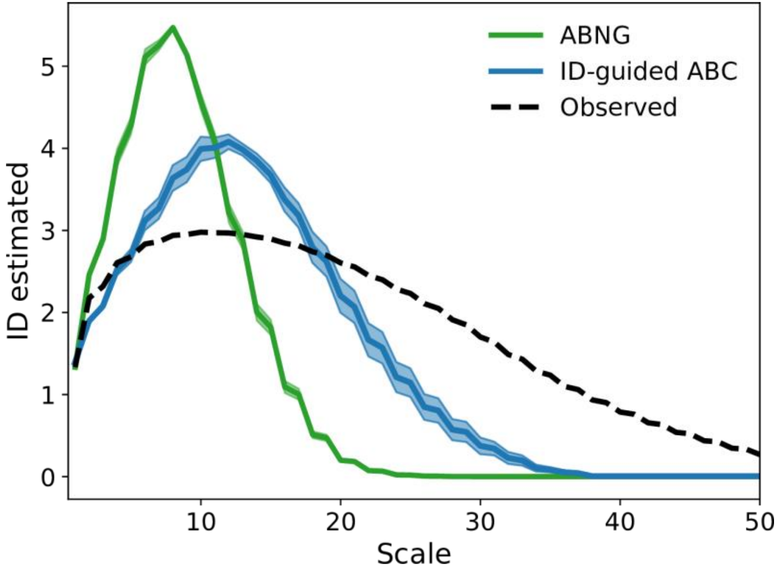
[28] A. Condon and R. Karp, "Algorithms for graph partitioning on the planted partition model," Random Structures & Algorithms, vol. 18, no. 2, pp. 116–140, 2001

model struggles with large-diameter real networks

US power stations: N=4941 E=6594

obtained by fitting betweennes, page rank degree distribution and clustering coeff

our procedure, trying to reproduce only the ID



!! It is far from trivial to devise growth mechanisms based that preserve the large scale structure !!

[27] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’networks,” Nature, vol. 393, no. 6684, pp. 440–442, 1998

ID-based generative model

Algorithm 4 Metropolis-Hastings accept/reject of single edge addition

Input: Starting network x, reference network's $\mathrm{ID}_R(x_0)$

Output: Evolved network x

1: Select the putative new edge following ABNG algorithm

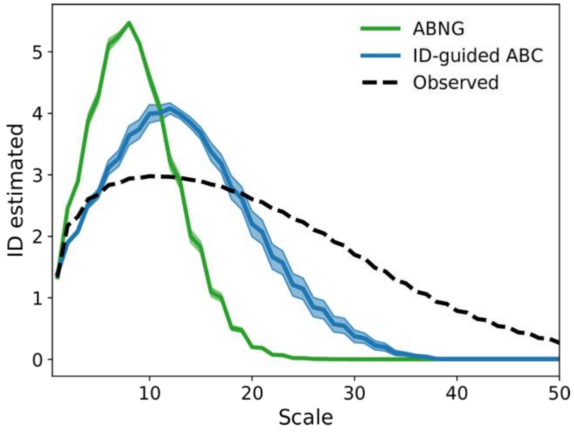
2: Store x'

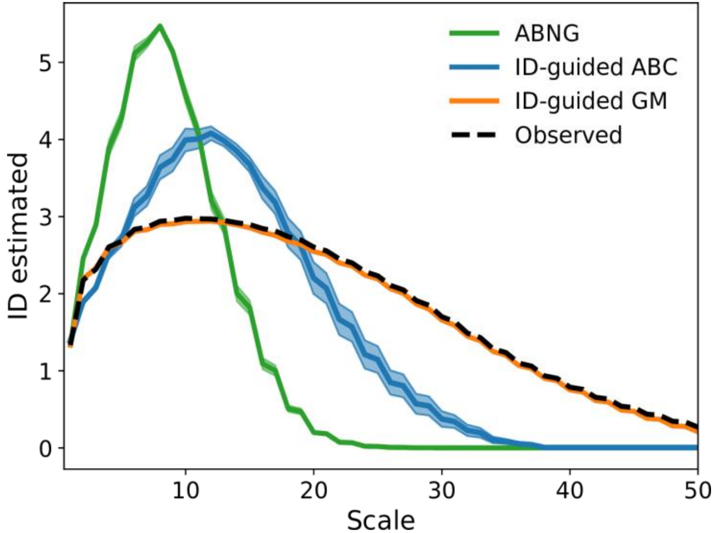
3: Compute $\mathcal{D}(x', x_0)$

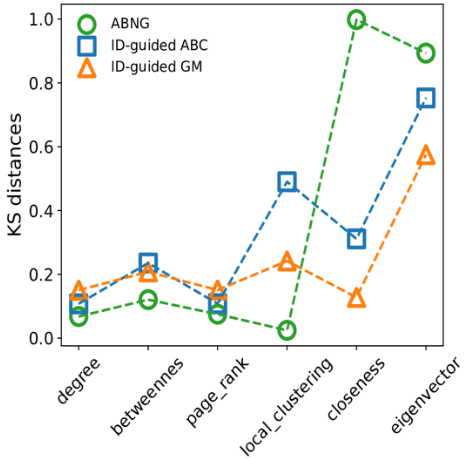
4: **if** $\exp\left[\beta\left(\mathcal{D}(x', x_0) - \mathcal{D}(x, x_0)\right)\right] < \mathcal{U}(0,1)$ **then**

5: $\quad x \leftarrow x'$

return x







Code in github repositories

IntRinsic: R package by F. Denti:
https://github.com/Fradenti/intRinsic
Implements: TWO-NN, GRIDE, Hidalgo

Phython code: https://github.com/sissa-data-science/DADApy
Implements: TWO-NN, GRIDE, Adaptive ID, I3D

References

1 Facco, et al. *Estimating the intrinsic dimension of datasets by a minimal neighborhood information*. Scientific reports 2017

2 Ansuini, et al. *Intrinsic dimension of data representations in deep NN*, Advances in NIPS, 2019

3 Allegra, et al. *Data segmentation based on the local ID*, Scientific Reports 2020

4 Denti, Doimo, Laio, Mira, *The generalized ratios intrinsic dimension estimator*, Scientific Reports 2022

5 Santos-Fernandez et al. The role of ID in high-resolution player tracking data, The Annals of Applied Statistics 2022

6 Denti, *intRinsic: an R package for model-based estimation of the ID of a dataset*, J Stat Software 2023

7 Macocco, Glielmo, Grilli, Laio, ID estimation for discrete metrics, Physical Review Letters, 2023

8 Varghese et al. A global perspective on the intrinsic dimensionality of COVID-19 data, Scientific Reports, 2023

9 I. Macocco et al. ID as a multi-scale summary statistics in ABC network parameter inference, R&R, Scientific Reports