From Degree Distribution to Graphs

Graph sampling review

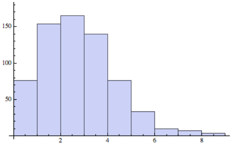
Mhamed Hajaiej[1] Nicolas Rahmouni[1]

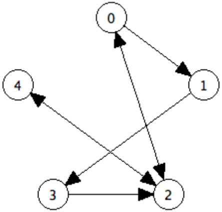
[1]Master MVA
ENS Paris-Saclay

Graphs in Machine Learning

Problematic

- How to construct a **directed** graph without any **self-loop** nor **multiple edge** with the exact degree sequence $D = (d^+, d^-)$? (If at least one graph exists, D is called a bi-graphical sequence)
- How to construct the set $\mathcal{G}(D)$ of **all graphs** having those properties?





Questions we adress:

- Computational efficiency?
- Ability to sample graphs from $\mathcal{G}(D)$ independently? Uniformly?
- "Differences" between those graphs

Sampling one graph

A Havel-Hakimi condition for bigraphicality

Condition for having a BDS

Assume that the BDS (d^+, d^-) is in **normal order**, and that $d_n^+ > 0$ (n is the last vertex, which is not ordered). Then (d^+, d^-) is bi-graphical if and only if the residual BDS

$$\Delta_k^+ = \begin{cases} d_k^+ & \text{if } k \neq n \\ 0 & \text{if } k = n \end{cases}$$

$$\Delta_k^- = \begin{cases} d_k^- - 1 & \text{if } k \leq n \\ d_k^- & \text{if } k > n \end{cases}$$

is bi-graphical.

Sampling one graph

The algorithm

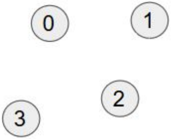
Algorithm

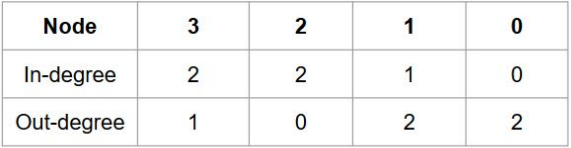
While there is a positive out-degree d^+ in the residual BDS.

- Choose a node v_n.
- Check if the d_n^+ first in degrees in the normal ordered BDS are strictly positive.
    - If not, the BDS is not bi-graphical → STOP
    - Otherwise, connect the corresponding nodes to v_n
- Compute the residual BDS $\Delta = (\Delta^+, \Delta^-)$ and order it in normal order.

Sampling one graph

A simple example

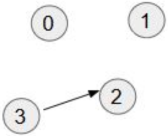


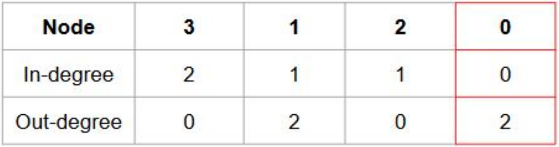


Node	3	2	1	0
In-degree	2	2	1	0
Out-degree	1	0	2	2

Sampling one graph

A simple example

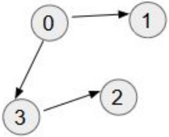


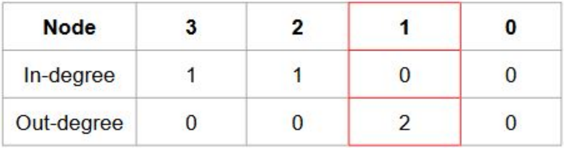


Node	3	1	2	0
In-degree	2	1	1	0
Out-degree	0	2	0	2

Sampling one graph

A simple example

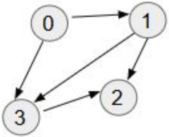


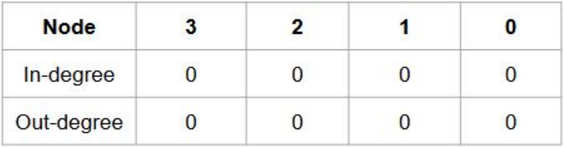


Node	3	2	1	0
In-degree	1	1	0	0
Out-degree	0	0	2	0

Sampling one graph

A simple example



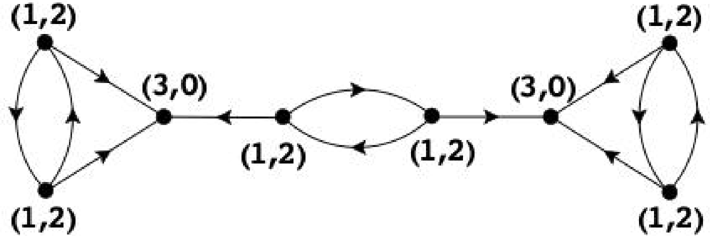


Node	3	2	1	0
In-degree	0	0	0	0
Out-degree	0	0	0	0

Sampling one graph

Remarks on the algorithm

- The **complexity** is a priori in $\mathcal{O}(N^2 \log(N))$
- Can be reduced to approximately $\mathcal{O}(N)$ if the re-ordering is done carefully.
- We can introduce a random choice for v_n to **sample independent** realization in $\mathcal{G}(D)$.
- **BUT** this algorithm can only sample graphs from a **subset** of $\mathcal{G}(D)$...



Sampling all graphs in $\mathcal{G}(D)$

Beginning from one realization

Edge swaps theorem

Let's G_1 and G_2 be to graphs in $\mathcal{G}(D)$. Then, there exists a sequence of edges swaps of length at most $2|E|$ that permits to transform G_1 into G_2.

Sampling graphs in a MCMC-style:

- Start from one realization
- Do k random edge swaps (until memory is forgotten)

Produces **pseudo-random samples** at the limit (ergodicity of the Markov chain) but the behavior is not well controlled.

Sampling all graphs in $\mathcal{G}(D)$

Configuration model

Algorithm (Brute-force with rejection)

For a BDS $D = (d^+, d^-)$

- For each node v_n, create d_n^+ out-stubs and d_n^- in-stubs
- For all out-stubs, choose a random unconnected in-stub and connect them
- If the obtained graph has self-loop or multiple edges, restart.

- Samples are independent and uniformly distributed.
- Rejection scheme → Very long to converge
- Not usable in practice.

Sampling all graphs in $\mathcal{G}(D)$

Direct independent sampling: Algorithm

What is wrong with the Havel-Hakimi algorithm?

At each iteration of the Havel-Hakimi algorithm the node v_n can only be connected to the d_n^+ first nodes in normal order.
BUT maybe there are other connection that do not break graphicality.

Idea: Find **the set of all nodes** that are suitable to be connected at each iteration $\mathcal{A}_n$ instead of a subset. ($\{d_n^+$ first nodes in normal order$\} \subset \mathcal{A}_n$)

Algorithm

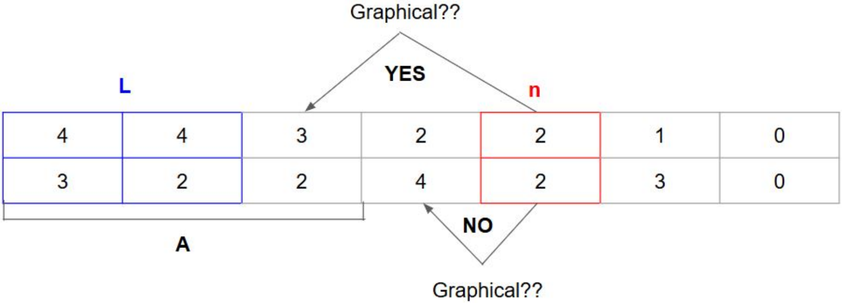
While there is a positive out-degree d^+ in the residual BDS.

- Choose a work-node v_n with non-zero out-degree and compute χ_n.
- Find the acceptable set $\mathcal{A}_n$.
- Choose at random a node $v_m \in \mathcal{A}_n$, connect it to v_n and add it to χ_n.
- Compute the residual BDS $\Delta = (\Delta^+, \Delta^-)$ and order it.
- If v_n still has a positive out-degree, restart at step 2.

Sampling all graphs in $\mathcal{G}(D)$

Direct independent sampling: How to find the acceptable set?

- Initialize $\mathcal{A}_n = \{d_n^+ \text{ first nodes in normal order}\}$
- Sequentially check the graphicality of the sequence where we connect the next node in the normal order. Add all nodes to $\mathcal{A}_n$ until graphicality is broken.



4	4	3	2	2	1	0
3	2	2	4	2	3	0

Sampling all graphs in $\mathcal{G}(D)$

Remarks on the algorithm

- Checking the graphicality of a sequence is in $\mathcal{O}(N)$ so the global complexity for sampling one graph is $\mathcal{O}(N^3 \log(N))$ which makes it impossible to sample large graphs in real-time.
- The samples are independent but non uniformly distributed.

BUT

- If the degree distribution is known in advance, all the graphs are computed and stored. Then, one can easily sample uniformly from $\mathcal{G}(D)$.
- Otherwise, if the issue is to compute the mean of a function over this graph set, the probability of obtaining the output graph is easily computed during the algorithm. So weighted average can be performed.

Graph similarity and metrics

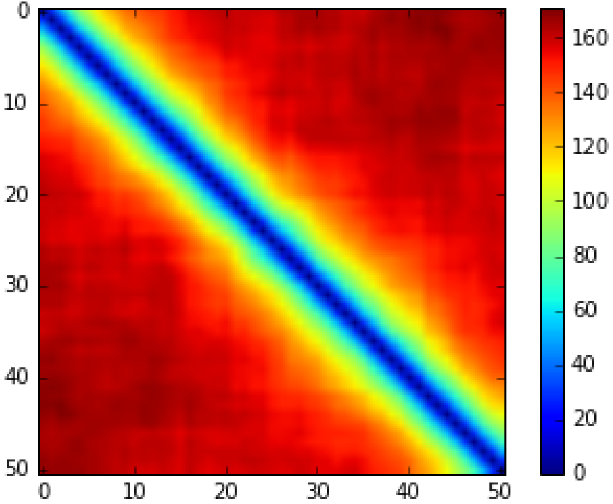
- Difficult problem and open question.
- Multiple metrics.

We have tried

- Hamming distance on Adjacency matrix.
- Average of the max eigenvector of Adjacency matrix.
- Degree correlation between neighboring nodes. Suggested to have a potential to classify different types of graphs.
- Histograms of graphs Dijkistra distances.
- Histograms of graphs page ranks.

Swaps for sampling

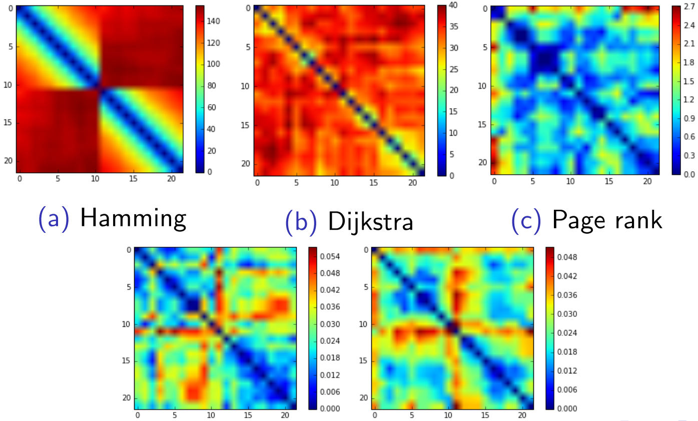
- MCMC sampling : How many swaps we need ?



Swaps for sampling

100 swaps

- MCMC sampling : How many swaps we need ? Depends on the graph dimension.
- Difficulties in graph similarity measurement



(a) Hamming

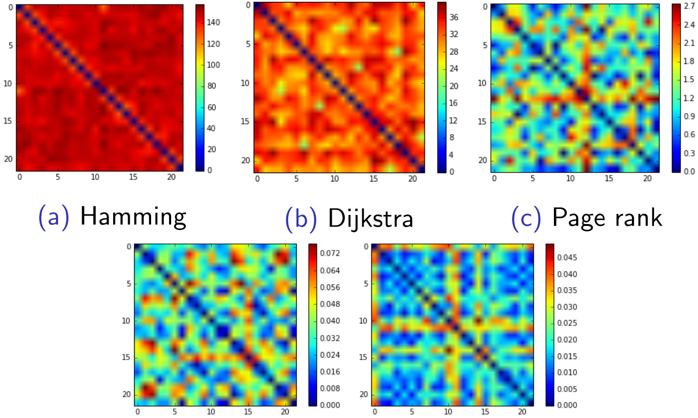
(b) Dijkstra

(c) Page rank

Swaps for sampling

1000 swaps

- More swaps and saturation.



(a) Hamming

(b) Dijkstra

(c) Page rank

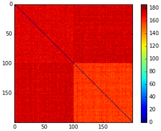
(d) Average of the max eigenvector

(e) Degree correlation

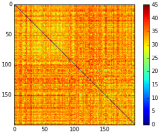
Methods comparison

MCMC vs independent sampling

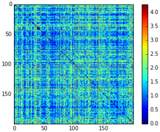
- Hard to evaluate : Similarity measures ambiguity again.
- MCMC seems to be better according to Hamming.
- Independent sampling is better according to Dijkstra and Degree.
- The two techniques samples different graphs from each other according to Hamming and Degree correlation.



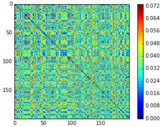
(a) Hamming



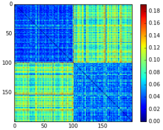
(b) Dijkstra



(c) Page rank



(d) Max

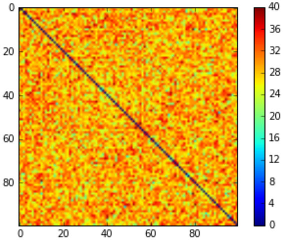


(e) Degree

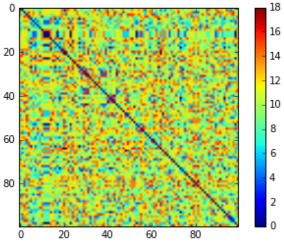
Comparing different graph types .

small world / preferential attachment / Erdos-Renyi

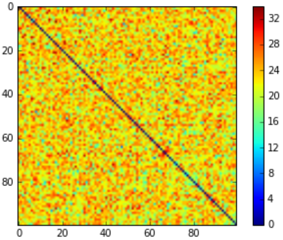
- We need more experiment to generalize the results.



(a) Hamming distance for small world graph.



(b) Hamming distance for preferential attachment graph.



(c) Hamming distance for Erdos-Renyi graph.