Limits of graph neural networks on large random graphs

GSP workshop 2023

Matthieu Cordonnier[1]
Nicolas Keriven[2], Nicolas Tremblay[1], Samuel Vaiter[3]

[1]GIPSA-lab, Grenoble
[2]IRISA, Rennes
[3]Labiratoire J-A Dieudonné, Nice

GNN for large graphs

Theoretical properties of GNNs remain largely misunderstood.

Expressivity : what class of functions can GNNs approximate. Azizian et al.

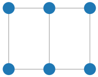
GNN for large graphs

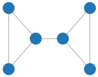
Theoretical properties of GNNs remain largely misunderstood.

Expressivity : what class of functions can GNNs approximate. Azizian et al.

Most studies have a combinatorial approach and focus on graph-level prediction.

Graph isomorphism problem and comparison to WL test. Xu et al.





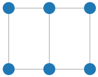
GNN for large graphs

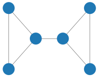
Theoretical properties of GNNs remain largely misunderstood.

Expressivity : what class of functions can GNNs approximate. Azizian et al.

Most studies have a combinatorial approach and focus on graph-level prediction.

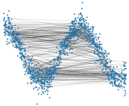
Graph isomorphism problem and comparison to WL test. Xu et al.

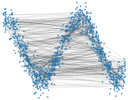




This is irrelevant on large graphs as they may have similar structure but are never isomorphic. Moreover, we rather focus on node-level prediction.

Different number of nodes, edges, etc..





Some classical problem related to GNN on large graphs

Stability to deformation, transferability.

Ruiz et al. 2020, Levie et al. 2021, Keriven et al. 2020

Some classical problem related to GNN on large graphs

Stability to deformation, transferability.
Ruiz et al. 2020, Levie et al. 2021, Keriven et al. 2020

Expressive power : What class of function can GNNs approximate
Keriven et al. 2021, 2023

Some classical problem related to GNN on large graphs

Stability to deformation, transferability.
Ruiz et al. 2020, Levie et al. 2021, Keriven et al. 2020

Expressive power : What class of function can GNNs approximate
Keriven et al. 2021, 2023

Generalisation : How well GNNs perform on unknown data
Maskey et al. 2022

Latent space random graph models

Let $\mathcal{X} \subset \mathbb{R}^d$, P a probability measure on $\mathcal{X}$ and $W : \mathcal{X}^2 \to [0, 1]$ a connectivity kernel. A random graph drawn from the model (W, P) has nodes :

$$X_1, \dots, X_n \overset{iid}{\sim} P.$$

Latent space random graph models

Let $\mathcal{X} \subset \mathbb{R}^d$, P a probability measure on $\mathcal{X}$ and $W : \mathcal{X}^2 \to [0, 1]$ a connectivity kernel. A random graph drawn from the model (W, P) has nodes :

$$X_1, \ldots, X_n \overset{iid}{\sim} P.$$

1. Weighted sampled graph: fully connected and e_{ij} is given a weight $w_{ij} = W(X_i, X_j)$.
   Lovasz 2010.

Latent space random graph models

Let $\mathcal{X} \subset \mathbb{R}^d$, P a probability measure on $\mathcal{X}$ and $W : \mathcal{X}^2 \to [0,1]$ a connectivity kernel. A random graph drawn from the model (W, P) has nodes :

$$X_1, \ldots, X_n \overset{iid}{\sim} P.$$

1. Weighted sampled graph: fully connected and e_{ij} is given a weight $w_{ij} = W(X_i, X_j)$.
   Lovasz 2010.
2. Graphon model: $e_{ij} \sim \mathcal{B}(W(X_i, X_j))$
   Lovasz 2010.

Latent space random graph models

Let $\mathcal{X} \subset \mathbb{R}^d$, P a probability measure on $\mathcal{X}$ and $W : \mathcal{X}^2 \to [0, 1]$ a connectivity kernel. A random graph drawn from the model (W, P) has nodes :

$$X_1, \ldots, X_n \overset{iid}{\sim} P.$$

1. Weighted sampled graph: fully connected and e_{ij} is given a weight $w_{ij} = W(X_i, X_j)$.
   Lovasz 2010.
2. Graphon model: $e_{ij} \sim \mathcal{B}(W(X_i, X_j))$
   Lovasz 2010.
3. Latent position model: $e_{ij} \sim \mathcal{B}(\alpha_n W(X_i, X_j))$ where α_n is a sparsity factor.
   Lei et al. 2015.

Main idea of GNN on large random graphs

A GNN on a random graph drawn from (W, P) has a “continuous” counterpart c-GNN on the random graph model (W, P).

GNN

Propagates a **graph signal** on a sampling of $\mathcal{X}$

$$X_i \mapsto f(X_i),\ i = 1 \dots, n$$

w.r.t the adjacency

$$W(X_i, X_j),\ i, j = 1 \dots, n.$$

c-GNN

Propagates a **function** on the latent space $\mathcal{X}$

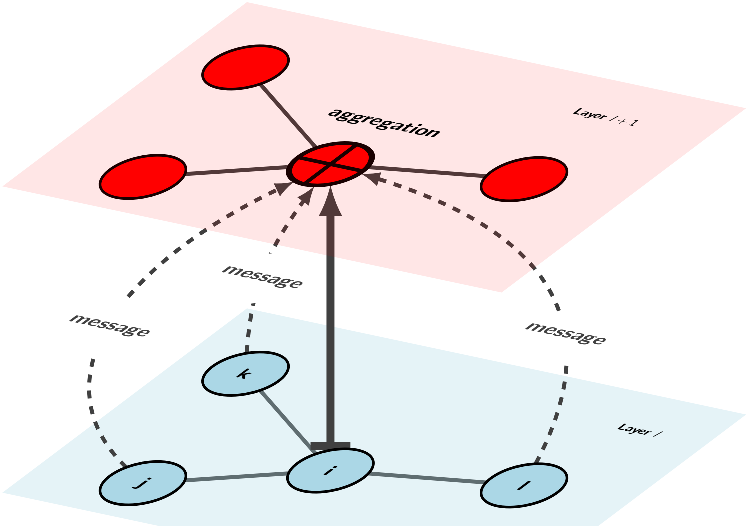
$$x \mapsto f(x),\ x \in \mathcal{X},$$

w.r.t the connectivity kernel

$$W(x, y),\ x, y \in \mathcal{X}.$$

Message Passing GNN (MPGNN)

Gather, Transform, Aggregate.



Topic of this work : Convergence

Our Problem

Given a GNN structure and a random graph model (W, P), does the GNN on random graphs *converge* to its c-GNN counterpart as n tends to infinity ? If yes, at which rate ?

Topic of this work : Convergence

Our Problem

Given a GNN structure and a random graph model (W, P), does the GNN on random graphs *converge* to its c-GNN counterpart as n tends to infinity ? If yes, at which rate ?

Some existing related results

Keriven et. al 2021

For Latent position model (model 3).

For Spectral GNN

Maskey et al. 2022

For weighted sampled graph model (model 1).

For Message Passing GNN with degree normalized mean aggregation.

[redacted]

Our results

Message passing with generic aggregation function.

Weighted sampled graph model (model 1.)

Under some regularity conditions on the aggregation: convergence with rate at least $O\left(\left(\frac{\ln n}{n}\right)^{1/2}\right)$.

Particular case of max aggregation: convergence with rate $O\left(\left(\frac{\ln n}{n}\right)^{1/d}\right)$ (recall $\mathcal{X} \subset \mathbb{R}^d$).

Formulation of message passing GNN

$\Theta_G : \mathbb{R}^{n\times d_0} \to \mathbb{R}^{n\times d_L}$ is a L layers MPGNN.
$Z^{(0)} \in \mathbb{R}^{n\times d_0}$ is the input graph signal.

Definition

$Z^{(1)}, \dots, Z^{(L)}$ are recursively computed by :

$$z_i^{(l+1)} = F^{(l+1)}\left(z_i^{(l)}, \left\{\!\!\left\{ \left(z_j^{(l)}, w_{i,j} \right) \right\}\!\!\right\}_{v_j \in \mathcal{N}(v_i)} \right) \in \mathbb{R}^{d_{l+1}}. \tag{1}$$

The output is :

$$\Theta_G(Z^{(0)}) = Z^{(L)}.$$

The $F^{(l)}$ are called the **aggregations**.

Major property : permutation equivariance

Equivariance to graph isomorphism

Let $\sigma \in S_n$. If $\sigma \cdot G$ and $\sigma \cdot Z$ are the **isomorphic graph** and **graph signal** where nodes have been relabeled *w.r.t* σ. Then

$$\Theta_{\sigma \cdot G}(\sigma \cdot Z) = \sigma \cdot \Theta_G(Z)$$

To that extent, the **aggregation** ignores ordering of the neighborhood.

Examples

1. **Convolutional** Bronstein et al. 2021

$$z_i^{(l+1)} = \frac{1}{|\mathcal{N}(v_i)|} \sum_{v_j \in \mathcal{N}(v_i)} w_{i,j} \psi^{(l+1)} \left(z_j^{(l)} \right).$$

2. **GAT** Velickovic et al. 2018

$$z_i^{(l+1)} = \sum_{j \in \mathcal{N}(v_i)} \frac{c_{ij}^{(l+1)}}{\sum_{k \in \mathcal{N}(v_i)} c_{ik}^{(l+1)}} \psi^{(l+1)}(z_j^{(l)}).$$

   Where $c_{ij}^{(l+1)} = c^{(l+1)}(z_i^{(l)}, z_j^{(l)}, w_{ij})$.

3. **Max Convolutional** Hamilton et al. 2018

$$z_i^{(l+1)} = \max_{v_j \in \mathcal{N}(v_i)} w_{i,j} \psi^{(l+1)} \left(z_j^{(l)} \right).$$

If $c_{ij}^{(l+1)} = w_{ij}$, we obtain the Degree Normalized MPGNN , Maskey et al. 2022

On random graphs

We recall the random graph model 1:

$\mathcal{X} \subset \mathbb{R}^d$ is **compact**.
P is a Borel **probability measure** on $\mathcal{X}$.
$W : \mathcal{X}^2 \to [0, 1]$ is a **measurable symmetrical kernel**.

Definition (Random Graph Model)

(W, P) is a **Random Graph Model** on $\mathcal{X}$. $G \sim \mathcal{G}_n(W, P)$ if :

- ⋆ $V(G) = \{X_i, \ldots, X_n\}$, where $X_i \overset{iid}{\sim} P$.
- ⋆ G is complete.
- ⋆ $w_{i,j} = w_{j,i} = W(X_i, X_j)$.

GNN on large random graph : intuition

Let $f : \mathcal{X} \to \mathbb{R}^{d_0}$, $Z = (f(X_i), \dots, f(X_n))$ be a **input signal**.

1. **Convolutional**

$$\frac{1}{n}\sum_i W(X_i, X_j)\psi(f(X_i)) \rightsquigarrow \int_{y\in\mathcal{X}} W(x,y)\psi(f(y))dP(y).$$

2. **GAT**

$$\sum \frac{c_{ij}}{\sum c_{ij}}\psi(f(X_i)) \rightsquigarrow \int_{\mathcal{X}} \frac{c(x,y)}{\int_{\mathcal{X}} c(x,t)dP(t)}\psi(f(y))\,dP(y).$$

3. **Max Convolutional**

$$\max_i W(X_i, X_j)\psi(f(X_i)) \rightsquigarrow \operatorname*{ess\,sup}_{y\in\mathcal{X}} W(x,y)\psi(f(y)).$$

Formulation of c-GNN

Propagates a **signal** over a **graph** $\rightsquigarrow$ Propagates a **function** over a **latent space**.

(W, P) is a random graph model.
$f = f^{(0)} : \mathcal{X} \to \mathbb{R}^{d_0}$ is an input signal.
$\Theta_{W,P} : (\mathcal{X} \to \mathbb{R}^{d_0}) \longrightarrow (\mathcal{X} \to \mathbb{R}^{d_L})$ is a L-layers c-GNN.

Definition

$f^{(1)}, \dots, f^{(l)}$ are recursively computed by :

$$f^{(l+1)}(x) = \mathcal{F}_P^{(l+1)} \left(f^{(l)}(x), \left(f^{(l)}, W(x, \cdot) \right) \right) \in \mathbb{R}^{d_{l+1}}$$

Final output : $\Theta_{W,P}(f) = f^{(L)} : \mathcal{X} \to \mathbb{R}^{d_L}$.

Required property : "continuous permutation" equivariance

Equivariance to random graph model isomorphism

$\phi : \mathcal{X} \to \mathcal{X}$ is a **bimeasurable bijection**.

$\phi \cdot (W, P) = (W(\phi^{-1}(\cdot), \phi^{-1}(\cdot)), \phi_\# P)$ is an **isomorphic random graph model.**

$\phi \cdot f = f \circ \phi^{-1}$ is the **isomorphic signal** on $\phi \cdot (W, P)$.
Then,

$$\Theta_{\phi \cdot (W,P)}(\phi \cdot f) = \phi \cdot \Theta_{W,P}(f)$$

$\phi_\# P$ is the pushforward measure of P through ϕ : $\phi_\# P(A) = P(\phi^{-1}(A))$ for any measurable set A.

countinuous counterparts of examples

1. **Convolutional**

$$f^{(l+1)}(x) = \int_{y\in\mathcal{X}} W(x,y)\psi^{(l+1)}\left(f^{(l)}(y)\right) dP(y)$$

2. **GAT**

$$f^{(l+1)}(x) = \int_{y\in\mathcal{X}} \frac{c^{(l+1)}(x,y)}{\int_{t\in\mathcal{X}} c^{(l+1)}(x,y)dP(t)}\psi^{(l+1)}\left(f^{(l)}(y)\right) dP(y).$$

Where $c^{(l+1)}(x,y) = c^{(l+1)}(f^{(l)}(x), f^{(l)}(y), W(x,y))$.

3. **Max Convolutional**

$$f^{(l+1)}(x) = \operatorname*{ess\,sup}_{y\in\mathcal{X},\ P} W(x,y)\psi^{(l+1)}\left(f^{(l)}(y)\right).$$

Convergence

We want to characterize convergence of a GNN on large random graphs. For $G_n \sim \mathcal{G}_n(W, P)$ do $\Theta_{G_n} \to \Theta_{W,P}$? At which speed ?

Convergence

We want to characterize convergence of a GNN on large random graphs. For $G_n \sim \mathcal{G}_n(W, P)$ do $\Theta_{G_n} \to \Theta_{W,P}$? At which speed ?

Define S_x the sampling operator at $(X_1, \dots, X_n)$.

$$\begin{tikzcd}
 & (f(X_i))_{i=1\dots,n} = Z \arrow[rr, "\Theta_{G_n}"] & & Z^{(L)} \in \mathbb{R}^{n \times d_L} \\
f \arrow[ur, dashed, "S_X"] \arrow[dr, "\Theta_{W,P}"] & & & \\
 & f^{(L)} \arrow[rr, dashed, "S_X"] & & \left(f^{(L)}(X_i)\right)_{i=1\dots,n} \in \mathbb{R}^{n \times d_L}
\end{tikzcd}$$

Error: $\max_{1 \le i \le n} \|Z_i^{(L)} - f^{(L)}(X_i)\|_\infty$
Main tool: Concentration inequalities

Convergence for "smooth aggregation"

Main tool : McDiarmid concentration inequaltiy.

Definition (Bounded differences)

Let $f : \mathcal{E}^n \to \mathbb{R}^d$. $C_1, \dots, C_n$ are the bounded differences of f if :

$$|f(x_1 \dots, x_i, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)| \leq C_i,$$

$\forall x_1, \dots, x_n, x_i' \in \mathcal{E}$.

Theorem

If f has finite bounded differences, then for any independent random variables $X_1, \dots, X_n$, $f(X_1, \dots, X_n)$ has a sub-Gaussian concentration around its expected value.

Suppose $F^{(l)}$ is Lipschitz continuous for a well chosen metric. Then it has finite bounded differences and they can be chosen all equal by symmety: $D_n^{(l)} = C_1^{(l)} = \cdots = C_n^{(l)}$. Then

Theorem (informal)

Let $\rho > 0$, with probability at least $1 - \rho$:

$$\max_{1 \le i \le n} \|z_i^{(L)} - f^{(L)}(X_i)\|_\infty \lesssim L D_n \sqrt{n \ln \left(\frac{n 2^L d_m}{\rho} \right)} + L r_n \qquad (2)$$

$D_n = \max_l D_n^{(l)}$, $d_m = \max d_l$, r_n *is a remainder that is specific to the network.*

Corollary (sufficient condition of convergence)

If $D_n = o\left(1/\sqrt{n \ln n}\right)$ then (2) tends to 0.

Example	D_n	r_n	$D_n = o\left(1/\sqrt{n \ln n}\right)$
Conv	$O(1/n)$	0	✓
GAT [1]	$O(1/n)$	$O(1/\sqrt{n})$	✓
Max. Conv	$\Omega(1)$	–	✗

Max does not have sharp bounded differences $\Longrightarrow$ we need other concentration inequality.

[1] Under some Lipschitz regularity condition on the attention coefficients $c^{(l)}$.

Convergence for Max Conv GNN

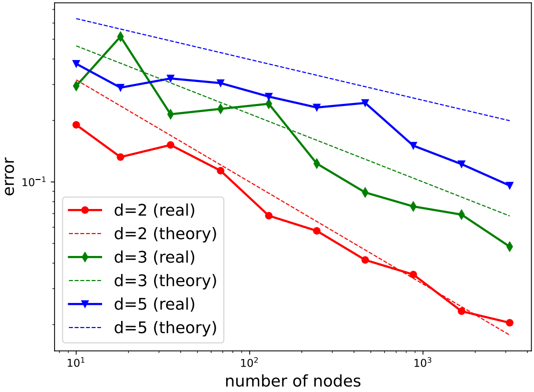
Theorem

Upon some regularity condition on $(\mathcal{X}, P)$, let $\rho > 0$, then with probability at least $1 - \rho$:

$$\max_{1 \leq i \leq n} \|z_i^{(L)} - f^{(L)}(X_i)\|_\infty \lesssim L \left(\frac{1}{n} \ln \left(\frac{2^L n d_{\max}}{\rho} \right) \right)^{1/d}, \quad (3)$$

Remark

For max aggregation, convergence speed depends on input dimension.



Conclusion

We have proven convergence of GNNs to their countinuous counterparts when the aggregation has a “Lipschitz-type” smoothness.

We have proven it too for max aggregation and observed a different behaviour.

McDiarmid concentration inequality may be suboptimal ? Can we find an aggregation with intermediate convergence speed bewteen $(1/n)^{1/2}$ and $(1/n)^{1/d}$?

Can we extend to more realistic random graph models (model 3.) ?

How to use c-GNNs to understand GNNs.

Thank you for your attention.