ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Richness of Node Embeddings in Graph Echo State Networks

Domenico Tortorella and Alessio Micheli

University of Pisa - Department of Computer Science
Largo B. Pontecorvo 3, 56127 Pisa - Italy

Abstract. Graph Echo State Networks (GESN) have recently proved effective in node classification tasks, showing particularly able to address the issue of heterophily. While previous literature has analyzed the design of reservoirs for sequence ESN and GESN for graph-level tasks, the factors that contribute to rich node embeddings are so far unexplored. In this paper we analyze the impact of different reservoir designs on node classification accuracy and on the quality of node embeddings computed by GESN using tools from the areas of information theory and numerical analysis. In particular, we propose an entropy measure for quantifying information in node embeddings.

1 Introduction

Relations between entities, such as paper citations or web page networks, can be best represented by graphs. In this representation, learning tasks such as identifying the topic of papers within a citation network or the level of traffic in a set of web pages joined by hyperlinks are node classification tasks. A plethora of neural models for graphs have been proposed to solve node-level learning tasks [1], most of them sharing an architecture structured in layers that perform local aggregations of node features. This architectural bias favors graphs with a low number of inter-class edges, i.e. with an high degree of homophily. Consequently, different architectural variations have been proposed [2] to address the issue of heterophily (i.e. low homophily) in node classification tasks, often significantly increasing the computational cost of training.

Graph Echo State Network (GESN) [3] is an efficient model within the reservoir computing (RC) paradigm. In RC, input data is encoded via a randomly-initialized reservoir, while only a linear readout requires training. GESN has already proved effective in node classification tasks [4], in particular offering astounding gains in accuracy on certain heterophilic graphs. While previous literature has analyzed the design of reservoirs in Echo State Networks (ESN) for temporal sequences [5, 6, 7] and in GESN for graph-level tasks [8], the factors that contribute to the effectiveness of GESN in node-level tasks are so far unexplored. In this paper we analyze the impact of different reservoir designs on the ability of GESN to provide rich node embeddings that enable good levels of accuracy in node classification tasks. To this end, we adapt tools from the areas of information theory and numerical analysis to assess the quality of node embeddings computed by different GESN reservoir architectures. In particular, we propose an information measure based on the Renyi quadratic entropy to quantify the richness of information provided by node embeddings.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2 Reservoir computing for node classification

The goal of a supervised node classification task is to learn a model able to infer the node classes from a subset of nodes with known labels in a graph, relying both on input features and network structure to make its predictions. Most common graph neural networks (GNN) learn node embeddings for the task at hand in a fully-trained network, structured in a hierarchy of local aggregations based on graph connectivity to represent increasingly larger receptive fields of each node. The training of this class of models has posed several challenges, including a bias toward graph with high homophily [2], and the problem of node embeddings becoming indistinguishable as the number of aggregation layers increases [9]. Graph Echo State Networks (GESN) instead follow the reservoir computing paradigm, encoding the graph input data by a randomly initialized reservoir, while only the task prediction layer requires training [3]. This model has already proved surprisingly effective in solving node classification tasks, addressing the two aforementioned issues that affect fully-trained GNNs [4].

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with node feature vectors $\mathbf{x}_v \in \mathbb{R}^X$ for each node $v \in \mathcal{V}$. We also denote by $\mathcal{N}(v)$ the set of neighbors of node v, and by $\mathbf{A}$ the graph adjacency matrix. In GESN, node embeddings $\mathbf{h}_v \in \mathbb{R}^H$ are recursively computed by the dynamical system

$$\mathbf{h}_v^{(k)} = \tanh\left(\mathbf{W}_{\text{in}}\,\mathbf{x}_v + \textstyle\sum_{v' \in \mathcal{N}(v)} \hat{\mathbf{W}}\,\mathbf{h}_{v'}^{(k-1)}\right), \quad \mathbf{h}_v^{(0)} = \mathbf{0}, \tag{1}$$

where $\mathbf{W}_{\text{in}} \in \mathbb{R}^{H \times X}$ and $\hat{\mathbf{W}} \in \mathbb{R}^{H \times H}$ are the input-to-reservoir and the reservoir recurrent weights, respectively (input bias is omitted). Reservoir weights are randomly initialized from a chosen distribution, and then rescaled to the desired input scaling and reservoir spectral radius, without requiring any training. The system evolves up until K iterations, sufficiently large for the node embeddings to capture a large enough receptive field. For graph-level tasks, equation (1) is instead iterated until the system state converges to a fixed point $\mathbf{h}_v^{(\infty)}$, which is used as the graph embedding after a global pooling operation [3]. The existence of such fixed point is guaranteed by the Graph Embedding Stability (GES) property [10]. A necessary condition [11] for the GES property is $\rho(\hat{\mathbf{W}}) < 1/\alpha$, where $\rho(\cdot)$ denotes the spectral radius of a matrix, i.e. its largest absolute eigenvalue, and $\alpha = \rho(\mathbf{A})$ is the graph spectral radius. While this property is crucial for effective graph global embeddings, node classification tasks have instead shown a stark preference for reservoirs initialized much beyond these stability constraints [4]. To solve the node classification task, we directly apply a linear readout to node embeddings $\mathbf{y}_v = \mathbf{W}_{\text{out}}\,\mathbf{h}_v^{(K)} + \mathbf{b}_{\text{out}}$, where the weights $\mathbf{W}_{\text{out}} \in \mathbb{R}^{C \times H}, \mathbf{b}_{\text{out}} \in \mathbb{R}^C$ are trained by ridge regression on one-hot encodings of target classes $y_v \in \{1, ..., C\}$ from the subset of graph nodes $\mathcal{V}_{\text{train}} \subset \mathcal{V}$ with known target labels $\{(\mathbf{x}_v, y_v)\}_{v \in \mathcal{V}_{\text{train}}}$.

3 Reservoir designs and embedding richness

We explore different designs for the recurrent weight matrix. The peculiar characteristics of $\hat{\mathbf{W}} \in \mathbb{R}^{H \times H}$, such as eigenvalue distribution and unit connectivity

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

patterns, influence the dynamical properties of the system (1), and consequently the quality of node embeddings produced by GESN. Some of the reservoir designs we analyze have been already explored in the RC field for ESN applied to temporal sequences, and include:

a) *Random ensembles.* The most simple method for a random reservoir initialization consists in drawing all matrix elements $\hat{W}_{ij}$ independently from the same distribution. If such distribution has zero mean and variance $\frac{1}{H}$, then the matrix spectrum will be asymptotically distributed uniformly on the unit disc [12]. In our experiments we will consider both uniform and normal distributions (i.e. Ginibre ensembles).

b) *Orthogonal reservoirs.* Reservoirs with orthogonal matrices, i.e. satisfying $\hat{\mathbf{W}}\hat{\mathbf{W}}^\top = \mathbf{I}_H$, have demonstrated to improve memory capacity in ESNs [5]. The eigenvalues of these matrices are distributed uniformly on a circumference of radius $\rho(\hat{\mathbf{W}})$. A simple way to sample a random orthogonal matrix is performing QR decomposition on a random ensemble matrix [12].

c) *Constrained spectra.* To explore the effects on the dynamical properties of GESN, we consider also matrices with eigenvalues constrained on the real axis (symmetric matrices) and imaginary axis (antisymmetric), obtained respectively as $\hat{\mathbf{W}}_{\text{symm}} = \hat{\mathbf{W}} + \hat{\mathbf{W}}^\top$ and $\hat{\mathbf{W}}_{\text{anti}} = \hat{\mathbf{W}} - \hat{\mathbf{W}}^\top$.

d) *Ring reservoirs.* Minimum-complexity reservoirs [6] adopt a deterministic recurrent matrix, namely a cyclic permutation matrix (itself an orthogonal matrix). This gives a ring-shaped connectivity pattern to reservoir units, making the system behave like a shift register [5]. This type of reservoirs has exhibited great performances on graph-level tasks [8].

The spectra of the reservoir matrices discussed so far are presented in Fig. 1.

To measure the quality of node embeddings produced by the different reservoir designs, we adapt two metrics rooted in information theory and numerical analysis previously employed for this purpose in ESNs [7]:

i) *Entropy.* We consider the Renyi quadratic entropy [13] to measure the richness of information provided by node embeddings $\mathbf{h}_v$. The entropy is computed via the Parzen–Rosenblatt probability density estimation of the node embeddings distribution $\mathcal{H}_2 = -\log \frac{1}{|\mathcal{V}|^2} \sum_{i=1}^{|\mathcal{V}|} \sum_{j=1}^{|\mathcal{V}|} G_{\mathbf{S}}(\mathbf{h}_i - \mathbf{h}_j)$, where $G_{\mathbf{S}}(\cdot)$ is the Gaussian kernel with covariance matrix $\mathbf{S}$ estimated by Silverman's rule. Higher values correspond to more distinguishable node representations, potentially offering also a metric for quantifying over-smoothing [9] not susceptible to affine transformations such as rescaling.

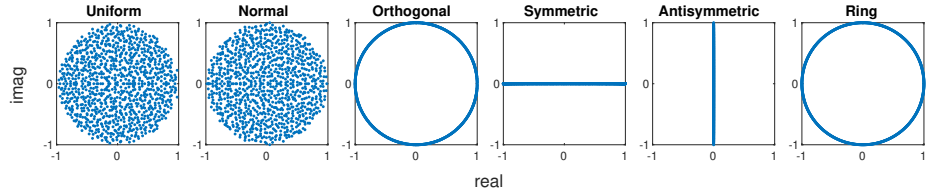

Fig. 1: Spectra of different designs of reservoir recurrent matrices.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

ii) *Uncoupled features.* To quantify the redundancy in node embeddings we measure the number of uncoupled features, that is the number of principal components sufficient to capture a fraction $q \in [0, 1]$ of total embedding variability, as analogously done in [7]. Given a matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}| \times H}$ of all node embeddings and its singular values $\sigma_i(\mathbf{H})$, we consider the metric $\mathcal{U}_q = \min_\mu \{\mu : \sum_{i=1}^{\mu} \sigma_i(\mathbf{H}) \geq q \sum_{i=1}^{H} \sigma_i(\mathbf{H})\}$; in our experiments we set $q = 0.9$.

4 Experiments and discussion

We evaluate the different GESN reservoir designs on seven node classification tasks, adopting the same experimental setting of our previous work [4]. We select the number of units H in the range $[2^4, 2^{12}]$, reservoir radius $\rho \in [0.1/\alpha, 50/\alpha]$, input scaling in $[0, 1]$, and readout regularization in $[10^{-5}, 10^2]$. We set the number of iterations of equation (1) to $K = 30$, which is comfortably larger then all graph diameters. Results are reported in Tab. 1, along with a graph-agnostic baseline ($\mathbf{h}_v = \tanh(\mathbf{W}_{\text{in}} \mathbf{x}_v)$) for reference, while reservoir radii chosen by model selection and the richness metrics of node embeddings obtained by selected reservoirs are reported in Fig. 2. We notice three different behaviors:

Actor and Wisconsin are heterophlic tasks where the accuracy of GESN is in line with the graph-agnostic baseline. Reservoir radii are selected within the stability region $\rho < 1/\alpha$, meaning node input features are more relevant then

Reservoir	Uniform	Normal	Orthog.	Symm.	Antisymm.	Ring	*Baseline*
Actor	$34.4_{\pm 0.6}$	$34.4_{\pm 0.9}$	$34.4_{\pm 0.9}$	$34.4_{\pm 0.9}$	$34.4_{\pm 0.8}$	$\mathbf{34.5}_{\pm 0.9}$	$34.3_{\pm 0.8}$
Wisconsin	$83.8_{\pm 3.9}$	$83.4_{\pm 3.5}$	$83.5_{\pm 3.4}$	$83.3_{\pm 4.1}$	$82.9_{\pm 3.4}$	$\mathbf{84.0}_{\pm 4.0}$	$83.8_{\pm 3.8}$
Squirrel	$73.3_{\pm 1.7}$	$73.5_{\pm 1.6}$	$\mathbf{73.6}_{\pm 1.7}$	$53.1_{\pm 1.0}$	$53.9_{\pm 1.5}$	$38.9_{\pm 1.0}$	$34.7_{\pm 1.2}$
Chameleon	$76.7_{\pm 1.6}$	$76.7_{\pm 1.4}$	$\mathbf{76.9}_{\pm 1.5}$	$65.8_{\pm 1.3}$	$66.5_{\pm 1.8}$	$53.7_{\pm 1.8}$	$49.0_{\pm 1.5}$
Cora	$86.1_{\pm 0.9}$	$86.0_{\pm 1.3}$	$\mathbf{86.2}_{\pm 1.0}$	$\mathbf{86.2}_{\pm 1.5}$	$85.5_{\pm 1.2}$	$85.6_{\pm 1.4}$	$74.1_{\pm 2.3}$
Citeseer	$74.4_{\pm 2.3}$	$74.4_{\pm 2.2}$	$\mathbf{74.6}_{\pm 2.1}$	$74.0_{\pm 2.2}$	$74.1_{\pm 4.6}$	$74.2_{\pm 2.1}$	$71.4_{\pm 2.1}$
Pubmed	$\mathbf{89.2}_{\pm 0.4}$	$89.1_{\pm 0.4}$	$\mathbf{89.2}_{\pm 0.3}$	$88.9_{\pm 0.3}$	$88.8_{\pm 0.4}$	$88.8_{\pm 0.5}$	$86.7_{\pm 0.4}$

Table 1: Node classification accuracy on low and high homophily graphs for different reservoir matrices. Best results are highlighted in bold.

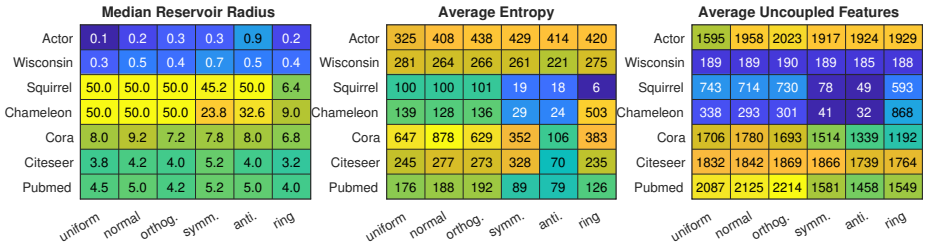

Fig. 2: Reservoir radii, entropy and uncoupled features for reservoirs chosen by model selection. (The reservoir units H are generally selected in $[2^{11}, 2^{12}]$.)

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

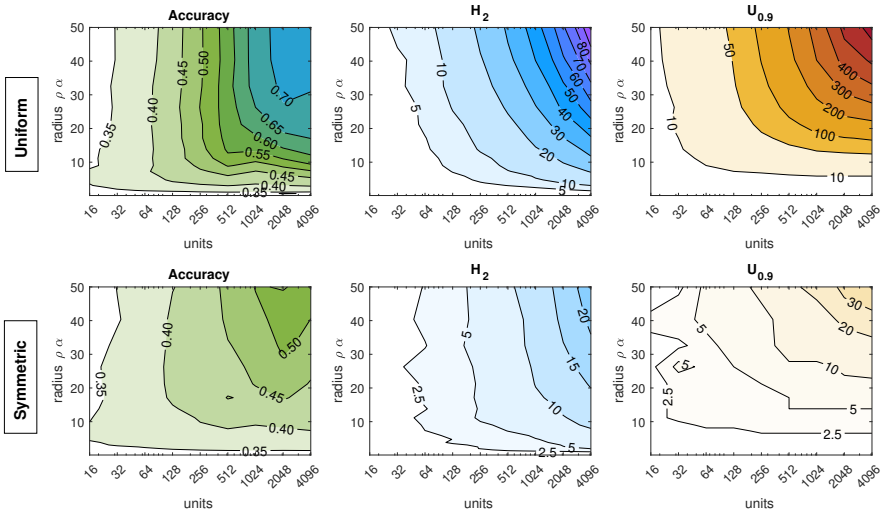

Fig. 3: Impact of reservoir radius and units on classification accuracy, entropy and uncoupled features for Squirrel with reservoirs of 4096 units.

graph connectivity. In this regime ring reservoirs perform slightly better, but the differences are not significant in terms of accuracy and node embedding richness.

Squirrel and Chameleon are heterophlic tasks where information needed for correct predictions is provided exclusively by graph connectivity, as evidenced by the large improvements over the baseline. In this case the reservoir designs present strikingly different behaviors, with more distinguishable node embeddings (as evidenced by higher entropy) corresponding to better classification accuracy. In that respect orthogonal matrices and random ensembles are the best performing, while constraining reservoir spectra appear to impact negatively on the dynamics of GESN, causing an accuracy drop of 10%–20% accompanied by a five-fold to ten-fold reduction in entropy and uncoupled features. Ring reservoirs select a reservoir radius much closer to the stability bound compared to the others, and provide a small but still significant improvement over baselines; in this case, the high values of entropy and uncoupled features may indicate distinguishable but misleading node embeddings, as graph topology may not be sufficiently taken into account due to the small selected reservoir radius (see [4]).

The remaining tasks are high homophily graphs, with reservoir radii selected beyond the stability region but not as much as the two previous tasks. Still, GESN improves significantly over graph-agnostic baselines. Orthogonal reservoirs perform slightly better, while all other designs are able to provide roughly the same level of accuracy and embedding richness.

In Fig. 3 we show an example of how richness metrics can be employed as a guide for the selection of reservoir parameters within a certain design. Notice indeed how the best hyper-parameter region for accuracy is the same as for entropy and uncoupled features. As the number of uncoupled features is considerably smaller than the number of reservoir units, it is evident that the

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

potential variability of node embeddings is not completely exploited.

5 Conclusion

In this paper we have for the first time analyzed several reservoir designs for GESN applied to node classification. Our experiments have shown that orthogonal matrices and random ensembles perform generally best, while constraints on spectra or unit connectivity can cause significant degradation on tasks more dependent on network rather than input features. In this case, the entropy of node embeddings can be used to guide the design choice. The general redundancy shown in node embeddings suggests introducing reservoir sparsity designs such as layering, which will be explored in future works. We will also explore the application of entropy as a scale-invariant measure for quantifying the degradation of node representations due to over-smoothing in deep graph neural networks.

Acknowledgments Research partly supported by PNRR, PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1, funded by European Commission under the NextGeneration EU programme.

References

[1] D. Bacciu, F. Errica, A. Micheli, and M. Podda. A gentle introduction to deep learning for graphs. *Neural Networks*, 129:203–221, 2020.

[2] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, volume 33, pages 7793–7804, 2020.

[3] C. Gallicchio and A. Micheli. Graph echo state networks. In *The 2010 International Joint Conference on Neural Networks*, pages 3967–3974, 2010.

[4] A. Micheli and D. Tortorella. Addressing heterophily in node classification with graph echo state networks. *Neurocomputing*, 550:126506, 2023.

[5] O. L. White, D. D. Lee, and H. Sompolinsky. Short-term memory in orthogonal neural networks. *Physical Review Letters*, 92(14):148102, 2004.

[6] A. Rodan and P. Tiňo. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1):131–144, 2011.

[7] C. Gallicchio and A. Micheli. Richness of deep echo state network dynamics. In *IWANN 2019: Advances in Computational Intelligence*, volume 11506 LNCS, pages 480–491, 2019.

[8] C. Gallicchio and A. Micheli. Ring reservoir neural networks for graphs. In *The 2020 International Joint Conference on Neural Networks*, 2020.

[9] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3438–3445, 2020.

[10] C. Gallicchio and A. Micheli. Fast and deep graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3898–3905, 2020.

[11] D. Tortorella, C. Gallicchio, and A. Micheli. Spectral bounds for graph echo state network stability. In *The 2022 International Joint Conference on Neural Networks*, 2022.

[12] E. Meckes. The eigenvalues of random matrices. *IMAGE: The Bulletin of the International Linear Algebra Society*, (65):9–22, 2020.

[13] D. Xu and D. Erdogmuns. Renyi's entropy, divergence and their nonparametric estimators. In *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, pages 47–102, New York, NY, 2010. Springer New York.