Multi-Scale Node Embedding and Network Reconstruction: A Comparative Analysis with the Single-Scale Literature

Riccardo Milocco[1] - [1,2,3] (Presenter), Fabian Jansen[2] - [2], Diego Garlaschelli[3] - [1,3]

1. IMT School for Advanced Studies, Piazza San Francesco 19, 55100 Lucca (Italy)
2. ING Bank N.V., Bijlmerdreef 106, 1102 CT Amsterdam (The Netherlands)
3. Lorentz Institute for Theoretical Physics, Leiden University, Niels Bohrweg 2, 2333 CA Leiden (The Netherlands)

Complex networks capture a variety of socially relevant processes, from economic activities to the way ecosystems respond to climate change. These phenomena naturally occur at multiple scales, but the data used to study them is often collected at convenient scale(s) where sufficient information is available. However, the properties of these networks can vary significantly at different resolution levels. This means that any models developed for a single scale (defined as Single-Scale Models or SSMs) become unreliable when applied at a different scale. This limitation underscores the need for a "multi-scale" approach that can abstract away from the specific dataset level and account for the multi-scale nature of the underlying phenomena.

Our analysis highlights the superior ability of Multi-Scale Models (MSMs), as compared to single-scale models, in capturing the binary-undirected projections of the observed graph at various levels of coarse-graining, denoted as $\{\mathbf{G}^*_\ell\}_{\ell\geq 0}$. Concretely, we apply this procedure to the World Trade Web (WTW) reported in the Gleditsch 2000 dataset and the Input-Output Network (ION) from ING Bank 2022 dataset.

To build $\{\mathbf{G}^*_\ell\}_{\ell\geq 0}$, we aggregated the observed (0-level) nodes into unique "community nodes" at higher levels $\ell \geq 0$, based on non-overlapping partitions. Specifically, for the WTW, we used geographical distances as a proxy for aggregation, whereas for the ION, we relied on NAICS (North American Industry Classification System) codes. However, any other algorithm for node aggregation can be employed.

The key distinction between the SSM and MSM is that the MSM can be both *renormalized* in the forward direction and "fine-grained" in the reverse direction. In the "bottom-up" procedure, the MSM prescribes to *sum* the microscopic parameters in order to obtain the parameters for the aggregated block-nodes where the "top-down" consists in fractioning the block-parameters into its constituents. This approach provides a robust interpretation of the *sum* of node vectors, a concept that is barely addressed in the node-embedding literature.

In this scenario, we worked with two research paradigms:

1. *Descriptive:* The objective of this research line is to describe the "scaled up" graphs $\{\mathbf{G}^*_\ell\}_{\ell\geq 0}$ by enhancing the MSM with node-embedding vectors (maxlMSM) which are obtained by maximizing the likelihood. Formally, a node embedding consists in assigning a vector to each node to encode its propensity of creating connections[4]. In the context of multi-scale unfolding, every block-node would have its vector as illustrated in Figure 1a. However, it is worth recalling that the communities have been derived by *uniquely* merging the *0-nodes*. This leads to the research question of whether the block-vectors can also be *uniquely* identified through the microscopic embeddings. Our findings indicate that the LogisticPCA model [1], regarded as a state-of-the-art machine learning model, lacks generalizability across different scales. In contrast, maxlMSM effectively captures coarser scales by enforcing that the block-vectors must be equivalent to the *sum* of the *0-vectors*. We decided to compare these two models based on the Binary Clustering Coefficient (BCC) at the "fitting" level 0, where node embeddings have been calibrated, and at level 2, which represents a coarse-grained scale (see Figure 1b). In particular, we report the expected BCC on the y-axis, corresponding to either the LogisticPCA or maxlMSM, whereas the observed BCC on the x-axis. This comparison is further enriched by insets displaying the probability matrix of the coarse-grained model (x-axis) against the one of the fitted model (y-axis) at levels 0 and 2 - the identity line represents the perfect match. Ultimately, the LPCA outperforms the maxlMSM at the fitting scale of 0; however, this relationship is reversed at level 2, where the maxlMSM overtakes the SSM performances.

[1] riccardo.milocco@imtlucca.it
[2] fabian.janses@ing.com
[3] diego.garlaschelli@imtlucca.it

[4] The choice of their dimension D is still debated and we agreed in using the dimension minimizing the *Bayesian Information Criteria* (BIC).

Overall, the maxlMSM exhibits a better agreement at multiple scales, supporting the conclusion that it is the most appropriate model within this multi-scale framework. Comparable results were also observed for the degree and the average nearest-neighbor degree (results not shown).

At last, we evaluated the two models using AUC-ROC and AUC-PR scores, which are widely recognized metrics within the machine learning community. In particular, we plotted the performance of these scores across various levels, treating the coarser levels as test sets, and we proceeded as in the previous point: fitted the parameters at level 0, we applied the *summed* procedure to obtain the model at level 2. The plot showed how the LPCA AUCs dropped at coarser levels whereas the MSM retrieves higher scores ("single-scale overfitting") underlying the different Nature of the two models;

2. *Network Reconstruction:* The scope of this part is to reconstruct the $\{\mathbf{G}^*_\ell\}_{\ell\geq 0}$ by encoding a number of parameters equal to the constrained observables: 1 global parameter to reproduce the total number of links L^* [4, 2] or N for the $N-$degrees $\{k^*_i\}_{[0,N-1]}$ [5]. Roughly, this technique unveils the microscopic structure (on average) starting from aggregated network measurements, such as L^*, and allows for a comparison between the exogenous node-observable (e.g. the Gross-Domestic-Product GDP_i) and the fitted parameters. Interestingly, the Configuration Model (CM) and the "degree-corrected" MSM (degcMSM) have similar functional forms; a characteristic that will have two consequences. The first one is that, differently from the previous point, visualizing the BCC in Figure 1d is not sufficient to conclude the CM is affected by "single-scale overfitting". Therefore, we tested the *scale-invariance* requirement on the CM which is mathematically enforced by the MSM: in Figure 1c we plotted the *summed* VS *coarse-grained* probabilities of observing the graph at level 2. As said, the equality holds only for the MSM model but not for the CM; a distinctive proxy of "non-scale-invariance".

During the two analysis, we have also recovered that it is feasible to reproduce (multi-scale) the number of triangles with a low-dimensional embedding, contradicting the recent claim made in [3].

To summarize, this examination emphasizes the need of MSM to properly *model* the phenomenon rather than "overfitting" its features to a specific scale. Indeed, all SSMs have been shown to be biased toward the fitted scale, yielding incorrect results for the others. On the other hand, the MSM showed the possibility of *describing/reconstructing* the coarser-scales by exploiting its *summation* recipe for the block-parameters.

PS: the two studies can be considered self-contained, except for the maxlMSM with $N-$ parameters, which relates to both the research lines. Therefore, I envision the Descriptive part in Model Complexity with AI and/or the Network Reconstruction in Econophysics Session.

References

[1] Sudhanshu Chanpuriya, Ryan A. Rossi, Anup B. Rao, Tung Mai, Nedim Lipka, Zhao Song, and Cameron Musco. Exact representation of sparse networks with symmetric nonnegative embeddings. In *Not Published*, 2021.

[2] Giulio Cimini, Tiziano Squartini, Diego Garlaschelli, and Andrea Gabrielli. Systemic risk analysis on reconstructed economic and financial networks. *Scientific Reports*, 5(1), October 2015.

[3] C. Seshadhri et al. The impossibility of low-rank representations for triangle-rich complex networks. *Proceedings of the National Academy of Sciences*, 117(11):5631–5637, 2020.

[4] Elena Garuccio, Margherita Lalli, and Diego Garlaschelli. Multiscale network renormalization: Scale-invariance without geometry. *Phys. Rev. Res.*, 5:043101, Oct 2023.

[5] Juyong Park and M. E. J. Newman. Statistical mechanics of networks. *Physical Review E*, 70(6), dec 2004.

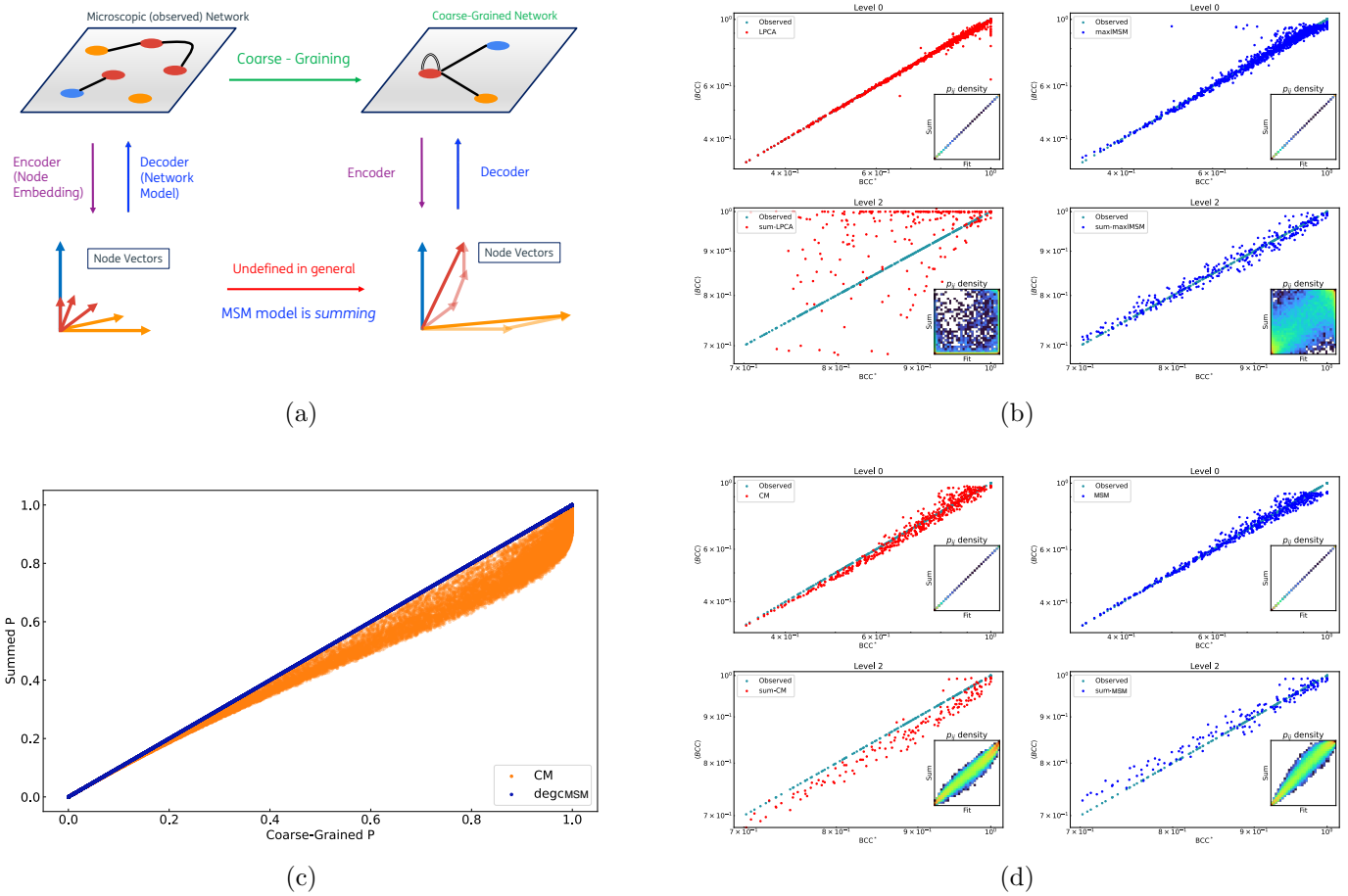

Figure 1: The graphical abstract in Figure 1a depicts the node-embedding procedure for two levels and the *coarse-graining* flow at the "network level". Generally, there is no established method for *renormalizing* the single-scale node-embedding to obtain a block-embedding. In Figure 1b, 1d we plotted the BCC for LPCA/maxlMSM and CM/degcMSM at the levels 0 (fitted), 2 (summed). In the insets compare the *summed* VS *fitted* probabilities at level 0 or 2. Interestingly, CM maintains a relationship with the observed BCC at level 2 differently from LPCA. In Figure c, the comparison between the *summed* (SP) VS *coarse-grained* probabilities reveals that for the CM, the *summed* probability underestimates the *coarse-grained* one. In contrast, the degcMSM theoretically enforces the identity among the two.