A ‘dramatic truth’ in link prediction: SBM inference fails to effectively predict even the structure of synthetic networks generated with the SBM model

Umberto Michieli[1,2], Alessandro Muscoloni[1], Leonardo Badia[2] and Carlo Vittorio Cannistraci[1,3]

[1] Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, Dresden, Germany; [2] Department of Information Engineering, University of Padova, Padova, Italy; [3] Brain bio-inspired computing (BBC) lab, IRCCS Centro Neurolesi “Bonino Pulejo”, Messina, Italy

Abstract

Methods for topological link prediction are generally referred as global or local. The former exploits the entire network topology, the latter adopts only the immediate neighborhood of the link to predict. Stochastic-Block-Model (SBM) is a global method believed as one of the best link predictors and widely accepted as reference when new methods are proposed. But, our results suggest that SBM, whose computational time is high, cannot in general overcome the Cannistraci-Hebb (CH) network automaton model that is a simple local-learning-rule of topological self-organization proved by multiple sources as the current best local-based and parameter-free deterministic rule for link prediction. In addition, after extensive tests on many different real complex networks of small size, Structural-Perturbation-Method (SPM) clearly emerges as the new best global method baseline. However, even SPM overall does not outperform CH and in several evaluation frameworks (in particular on networks of large size) we astonishingly found the opposite [1]. At this point of our study, we decided to investigate better the nature of the disappointing SBM’s performance, and the extent to which SBM fails because of inference issues. The Lancichinetti-Fortunato-Radicchi (LFR) model is a special version of the degree-corrected SBM [2] and is typically adopted to generate artificial benchmarks. Interestingly, although LFR networks are generated by a model based on the SBM theory, we show that, in general, a large family of SBM link predictors cannot reach overall performances on LFR networks at the same level as SPM and CH, revealing clear inference problems of the SBM family on link prediction tasks.

Methods

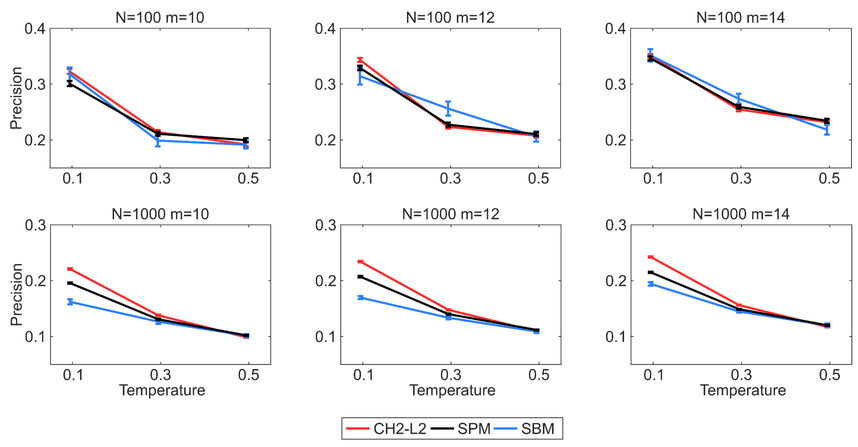
Link prediction methods

- Cannistraci-Hebb based on length-2 paths (**CH2-L2**) [3,4]
- Resource-Allocation based on length-3 paths (**RA-L3**) [5]
- Structural Perturbation Method (**SPM**) [6]
- Fast probability Block Model (**FBM**) [7]
- Stochastic Block Model (**SBM**) [8]
- Nested and Degree-Corrected Stochastic Block Model (**SBM DC N**) [9]
- Nested Stochastic Block Model (**SBM N**) [9]
- Degree-Corrected Stochastic Block Model **(SBM DC**) [9]

Evaluation procedure

For each network, 10% of links have been randomly removed (10 iterations for the SBM-based methods due to the high computational time, 100 iterations for the other methods) and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The links are ranked by likelihood scores and the precision is computed as the percentage of removed links among the top-*r* in the ranking, where *r* is the total number of links removed.

Link prediction on synthetic nPSO networks



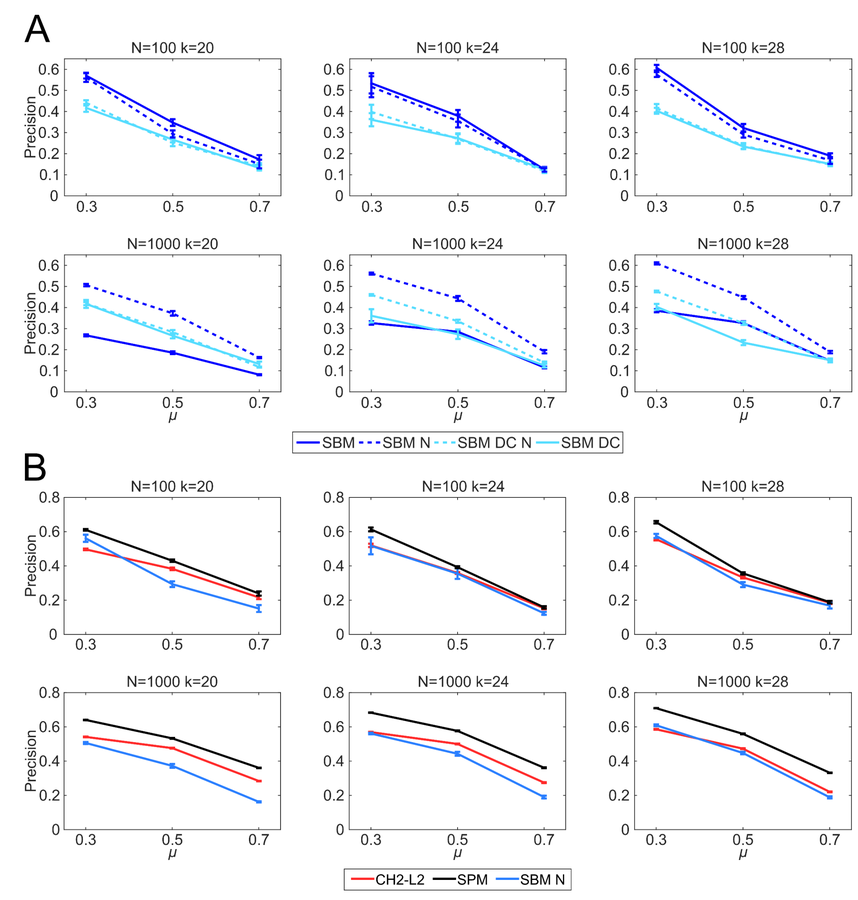
Synthetic networks have been generated using the nonuniform PSO model [10] with parameters N = [100, 1000] (network size), γ = 3 (power-law degree distribution exponent), m = [10, 12, 14] (half of average degree), T = [0.1, 0.3, 0.5] (temperature, inversely related to the clustering coefficient) and 8 communities. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Only the best SBM variant is shown.

Link prediction on real networks

	SPM	CH2 L2	SBM	FBM	RA L3	SBM DC N	SBM DC	SBM N
mouse_neural	0.02	0.09	**0.10**	0.01	0.03	0.05	0.03	0.00
karate	0.17	0.19	**0.28**	0.27	0.15	0.13	0.14	0.01
stmarks_foodweb	0.26	0.13	**0.29**	0.14	0.25	0.19	0.18	0.07
dolphins	0.13	0.14	0.16	**0.19**	0.10	0.08	0.08	0.00
ythan_foodweb	**0.27**	0.08	0.24	0.09	0.26	0.20	0.19	0.06
macaque_neural	**0.72**	0.56	0.68	0.55	0.64	0.38	0.36	0.12
polbooks	0.17	**0.18**	0.15	**0.18**	0.16	0.15	0.14	0.01
SEA_terrorist	0.45	**0.46**	0.29	0.33	0.25	0.29	0.25	0.08
ACM2009_contacts	0.26	**0.27**	0.25	0.26	**0.27**	0.19	0.19	0.05
football	0.31	**0.36**	0.34	0.25	0.21	0.28	0.26	0.06
physicians_innovation	0.07	**0.08**	0.06	**0.08**	0.05	0.02	0.04	0.01
AQ_terrorist	0.36	**0.42**	0.22	0.35	0.26	0.16	0.12	0.03
manufacturing_email	**0.51**	0.42	0.47	0.39	0.39	0.37	0.37	0.09
jazz	**0.65**	0.58	0.47	0.45	0.40	0.37	0.35	0.09
residence_hall_friends	**0.28**	0.25	0.18	0.24	0.15	0.15	0.15	0.04
rhesus_brain	**0.31**	0.25	0.21	0.24	0.19	0.23	0.22	0.03
vanderwaals	**0.29**	0.19	0.08	0.17	0.13	0.09	0.05	0.02
haggle_contacts	0.62	0.57	0.62	0.57	**0.63**	0.45	0.44	0.10
worm_nervoussys	**0.16**	0.12	0.15	0.11	0.12	0.15	0.12	0.03
USAir	**0.46**	0.43	0.38	0.38	0.40	0.39	0.37	0.07
netsci	0.41	**0.54**	0.13	0.33	0.29	0.25	0.15	0.05
infectious_contacts	**0.37**	0.35	0.30	0.33	0.26	0.19	0.16	0.07
flightmap	**0.75**	0.56	0.64	0.56	0.58	0.55	0.56	0.12
email	0.16	**0.17**	0.09	0.16	0.12	0.10	0.08	0.02
polblog	**0.23**	0.17	0.19	0.17	0.18	0.20	0.18	0.18
mean precision	**0.34**	0.30	0.28	0.27	0.26	0.22	0.21	0.06
mean ranking	**2.1**	2.9	3.6	4.1	4.3	5.2	6.1	7.9
mean time	1.4 s	**1.0 s**	2.6 h	8.4 s	2.8 s	1.1 d	2.2 h	22.3 h

The table reports for each network the mean precision over the random iterations, as well as the mean precision, mean ranking and mean time over the entire dataset. For each network the best method (or methods) is highlighted in bold.

Link prediction on synthetic LFR networks



Synthetic networks have been generated using the LFR model [11] with parameters N = [100, 1000] (networks size), k = [20, 24, 28] (average degree) and μ = [0.3, 0.5, 0.7] (mixing parameter). The minimum and the maximum of the community sizes have been fixed respectively to *minc* = N/20 and *maxc* = 4**minc*. The maximum degree of a node has been set to *maxk* = 3**k*. Low values of μ generate strong clustering and a desired clustering coefficient C = [0.7, 0.5, 0.3] is attempted to be satisfied respectively to the values of μ. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. The plots in **(A)** compare the methods of the SBM family, whereas the best of them overall, i.e. SBM N, is compared in **(B)** with the state-of-the-art methods CH-L2 and SPM.

References:

[1] A. Muscoloni, U. Michieli, and C. V. Cannistraci, “Local-ring network automata and the impact of hyperbolic geometry in complex network link-prediction”, arXiv:1707.09496, 2018.
[2] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” Physics Reports, vol. 659. pp. 1–44, 2016.
[3] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, “From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks,” Sci. Rep., vol. 3, no. 1613, pp. 1–13, 2013.
[4] A. Muscoloni, I. Abdelhamid, and C. V. Cannistraci, “Local-community network automata modelling based on length- three-paths for prediction of complex network structures in protein interactomes, food webs and more.” bioRxiv, 2018.
[5] I. A. Kovács et al., “Network-based prediction of protein interactions,” bioRxiv, p. 275529, 2018.
[6] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, “Toward link predictability of complex networks,” Proc. Natl. Acad. Sci., vol. 112, no. 8, pp. 2325–2330, 2015.
[7] Z. Liu, J. L. He, K. Kapoor, and J. Srivastava, “Correlations between Community Structure and Link Formation in Complex Networks,” PLoS One, vol. 8, no. 9, 2013.
[8] R. Guimerà and M. Sales-Pardo, “Missing and spurious interactions and the reconstruction of complex networks,” Proc. Natl. Acad. Sci. U. S. A., vol. 106, no. 52, pp. 22073–22078, 2009.
[9] T. P. Peixoto, “The Graph-tool Python Library,” Figshare, 2014.
[10] A. Muscoloni and C. V. Cannistraci, “A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities,” New J. Phys., vol. 20, p. 052002, 2018.
[11] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms”, Physical Review E, 2008.