University of Padova

Department of Information Engineering

Master Thesis in Telecommunication Engineering

Link Prediction on Real and Synthetic Complex Networks

Master Candidate:
Umberto Michieli

Supervisors:
Leonardo Badia
University of Padova

Carlo Vittorio Cannistraci
Technische Universität Dresden

10/09/2018
Academic Year 2017/2018

Abstract

Methods for topological link prediction are generally referred to as *global* or *local*. The former exploit the entire network topology, the latter adopt only the immediate neighborhood of the link to predict. Global methods are believed to achieve the best performance. Stochastic-Block-Model (SBM) is a global method regarded as one of the best link predictors and widely accepted as a benchmark when new methods are proposed. Several variations of SBM have been proposed throughout the years and this study represents the widest test of this theory available in the literature. The results suggest that SBM and its variations, whose computational time is high, cannot in general overcome the Cannistraci-Hebb on paths of length 2 (CH2-L2) network automaton model that is a simple local-learning-rule of topological self-organization proved by multiple sources to be the current best local-based and parameter-free deterministic rule for link prediction. In particular, SBM-based methods displayed inference problems even on Lancichinetti-Fortunato-Radicchi (LFR) networks, which are built using the SBM theory. In addition, after extensive tests, Structural-Perturbation-Method (SPM) is recommended as the new best global method baseline. However, even SPM overall does not outperform significantly CH2-L2 in all the scenarios. In particular, CH2-L2 was the best predictor for synthetic networks generated by the nonuniform Popularity-Similarity-Optimization (nPSO) model. Interestingly, when tested on non-hyperbolic synthetic networks, the performance of CH2-L2 dropped down indicating that such a self-organizational rule could be strongly correlated to the rise of hyperbolic geometry in complex networks. The superiority of global methods in link prediction, and in particular of SBM-based ones, seems then a misleading belief caused by a latent geometry bias of the few small networks used as benchmark in previous studies. Therefore, a need emerges for a latent geometry theory of link prediction in complex networks.

Link Prediction on Real and Synthetic Complex Networks

SUPERVISORS: Leonardo Badia and *Carlo Vittorio Cannistraci*
MASTER CANDIDATE: Umberto Michieli

Academic Year 2017/2018

Contents

List of Figures

List of Tables

Acronyms

AIMD	Additive Increase and Multiplicative Decrease
AS	Autonomous Systems
CH	Cannistraci-Hebb
CN	Common Neighbours
eLCLs	external Local Community Links
FBM	Fast probability Block Model
iLCLs	internal Local Community Links
L2	paths of Length 2
L3	paths of Length 3
LCP	Local Community Paradigm
LFR	Lancichinetti-Fortunato-Radicchi
LTD	Long Term Depression
LTP	Long Term Potentiation
MCMC	Monte Carlo Markov Chain
nPSO	nonuniform Popularity-Similarity Optimization
PPI	Protein-to-Protein Interaction
PSO	Popularity-Similarity Optimization

RA	Resource Allocation
RGG	Random Geometric Graph
SBM	Stochastic Block Model
SBM DC	Degree Corrected SBM
SBM DC N	Degree Corrected and Nested SBM
SBM N	Nested SBM
SPM	Structural Perturbation Method
TCP	Triadic Closure Property
WS	Watts-Strogatz

1 Introduction

During the last decades Network Science field has been rediscovered and addressed as the "new science" [1], [2]. A lot of issues have been (re-)examined thanks to these techniques, which are nowadays permeating the way we face the world as a unique interconnected component. The presence and the immediate availability of a huge amount of digital data describing every kind of network and the way in which its nodes interact, has made possible an interdisciplinary analysis of many large-scale systems.

The *United States National Research Council* defines Network Science as *"the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena"* [3].

The wide application of Network Science techniques is mainly due to their impressive abstracting power. Once a network, whose definition will be given in the following, has been identified, most of the tools to analyze it can simply be applied to the considered network with none or minimal modifications. This is the reason why Network Science field is so wide and has gained a lot of interest in the last decades: from telecommunications to computer networks, from biological to semantic networks, from social to brain networks, and so on. Again, the only prerequisite to apply Network Science techniques is to generate or identify a network representing the actors which play a role in it and their connections between each other.

The definition of a *network* (or *graph*) is then straightforward. It comes from the Mathematical field of Graph Theory and only consists of two elements: a set of *nodes* (or *vertices*) and a set of *links* (or *edges*), which are connected to the nodes and represent some kind of interaction between them. Throughout this thesis the *networks* will always been considered in the most common meaning, i.e. *undirected*, *unweighted* and *connected*. An *undirected* network is a graph in which edges do not have orientation; an *unweighted* network is a graph in which edges do not have a weight associated to them; a *connected* network is a graph where there is a path between every pair of vertices. An example of undirected, unweighted and connected network with 7 nodes and 14 edges is reported in Figure 1.1 for illustration purposes.

The aim of topological link prediction is to detect, in a given network, the non-observed

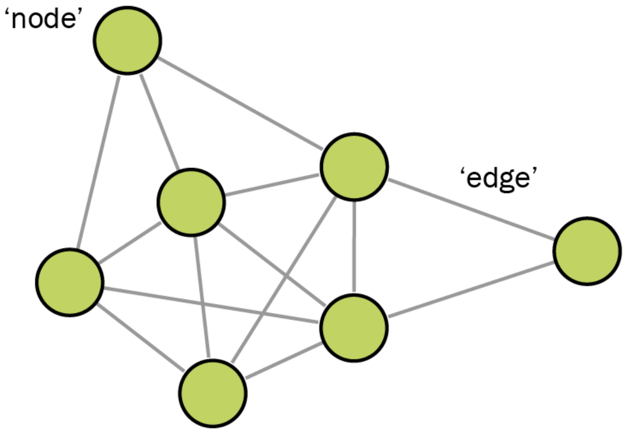

Figure 1.1: Example of network with 7 *nodes* and 14 *edges*.

links that could represent missing information or that may appear in the future, only exploiting features intrinsic to the network topology. It has a very wide range of real applications (especially in that disciplines where the discovery of new links is costly in the laboratory or in the field, such as in metabolic, biological or foodweb networks) and the three typical applications of link prediction are for *networks reconstruction*, *time-evolving networks mechanisms* and *classification of partially labeled networks*. A non-comprehensive list of applicative fields for link prediction follows [4], [5], [6]:

- suggestion of friendships in social networks [5];
- prediction of interactions in biological networks;
- prediction of being actors in acts [7];
- prediction of new collaborations in co-authorship networks [5];
- detection of hidden relationships between terrorists [8];
- recommendation of items to users [9], [10];
- personalized recommendation based on known interests [10–14];
- recommendation in e-commerce websites [15];
- prediction of protein function, because most of the interactions among proteins is still unknown [16];

- detection of anomalous emails [17];
- distintion of the research areas of scientific publications [18].

Although this study is focused on monopartite networks, link prediction has recently been successfully implemented also in different types of network topologies such as bipartite [11], [19] and multilayer networks [20].

The link prediction methods, according to the type of topological information exploited, can be broadly classified in two main categories: global and local. Global methods take advantage of the entire network topology in order to assign a likelihood score to a certain non-observed link. On the contrary, local approaches take into consideration only information about the neighborhood of the link under analysis [4], [6]. Beside them, there are also quasi-local or probabilistic approaches [4].

In 2009, R. Guimerà et al. proposed a new global inference framework based on Stochastic Block Model (SBM) in order to identify both missing and spurious interactions in noisy network observations [21]. The general idea of a block model is that the nodes are partitioned into groups and the probability that two nodes are connected depends only on the groups to which they belong. The framework introduced is a global approach where, assuming that there is no prior knowledge about which partition is more suitable for the observed network, the likelihood of a link can be computed theoretically considering all the possible partitions of the network into groups. Since this is not possible in practice, the Metropolis algorithm, which is based on a stochastic procedure, is exploited in order to sample only a subset of partitions that are relevant for the estimation of the link reliability [21]. The high computational time becomes anyway prohibitive for large networks and restricts the range of applicability to small networks up to at most a few thousand nodes [4]. However, in many recent link prediction studies where new methods are proposed, SBM is considered among the best state-of-the-art methods to adopt as a baseline for a performance comparison. For the sake of clarity, here are explicitly reported a few examples.

From [22]: "*Roger Guimerà et al. proposed a Stochastic Block Model (SBM) which can predict both missing links and spurious links and is able to give much better accuracies of prediction on various kinds of networks than current popular methods including the HRG approach [33]. Because the SBM algorithm is a state-of-the-art approach which has very outstanding accuracy performance of link prediction on undirected networks without additional node's or edge's attribute information, we mainly make performance comparisons on both accuracy of missing link prediction and computational efficiency between our algorithm and the SBM approach.*"

From [23]: "*Surprisingly, by directly applying the first-order matrix perturbation method, we achieve more-accurate link predictions than some gracefully designed methods such as HSM [18]* *and SBM [25].*"

From a recent survey on the link prediction state-of-the-art [24]: "*Stochastic block model, characteristics: outperforms at identifying both missing links and spurious links; high computation time. [. . .] probabilistic graph models achieve better performance than basic topology-based metrics, especially improve the prediction accuracy.*"

Other investigations that consider SBM a state-of-the-art global method are [25–29].

A further implementation of SBM usable also for link prediction, but to the best of our knowledge not yet tested for this task, has been developed by T. P. Peixoto in 2014 [30]. This contribution adds to the standard SBM the possibility to discriminate among nested levels of hierarchy in partitioning the network into communities, and allows to also account for small but well-defined clusters of nodes. The proposed approach could also be combined with the Degree Corrected SBM (DC SBM) already proposed in [31] and analyzed for community detection in [32–34].

These approaches are said to be also useful in order to boost link prediction tasks without empirical proofs. For example from [30]: "*We also predict that it* [the SBM N approach] *should serve as a more refined method of detecting missing information in networks, as well as for the prediction of the network evolution, determining the more salient topological features or large-scale functional summaries of the network topology.*"

Link prediction studies are not always convincing and exhaustive in the selection of the approaches to adopt as a reference for comparison. Furthermore, the malpractice to consider few (about ten) small-size (less than 1000 nodes) networks as a benchmark can lead to wrong conclusions. Two observations triggered our attention. Firstly, the remarkably high computational time of SBM and the consequent network size constraint for its application. Secondly, the fact that, as far as we are concerned, the scientific literature does not offer convincing proofs of SBM's (and other global methods) superiority with respect to the best local methods. Moving from these premises, we decided to conduct an accurate study that compares SBM and the most promising algorithms for link prediction. Hence, in this essay, it has been made a thoughtful analysis of the best global and local methods, which have been extensively (about 40 real networks used plus different models of artificial networks and almost 500 small-size brain networks) tested on many evaluation frameworks both with small (less than a thousand nodes) and large (from 3000 up to 40000 nodes) networks. As case studies, we considered evaluations on re-prediction of random removed links and on network evolution across time. performance on real and artificial networks have been compared. The overlap and the diversity between the true links predicted by global and local methods have been evaluated. Together with the study of Liben-Nowell and Kleinberg [5], this represents the largest and most recent study on testing state-of-the-art methods for topological link prediction in complex unweighted and undirected monopartite networks. In addition, we extensively test a very recent theory of network self-organization rule introduced in [35], [36] known as Cannistraci-Hebb Local Community Paradigm (CH LCP). At this stage, the link prediction field needs a reorganization of the knowledge in order to reach an agreement and set clear guidelines for the future. Here, there is an attempt to propose the baseline methods and the evaluation strategies that, for a fair comparison, should be taken into consideration and included in forthcoming studies.

2 | State-of-the-Art Methods for link prediction

The best performing and the most promising state-of-the-art methods for link prediction are now presented. Two local methods, which have already been proved to outperform other state-of-the-art local methods, have been considered. As global methods SPM has been chosen and there is a special focus on methods based on the Stochastic Block Model (SBM) theory since, as already mentioned, they are addressed as the best performing ones. Indeed, five methods based on the SBM-theory have been considered, despite the huge computational time required by this kind of methodology.

2.1 Local Methods

2.1.1 Cannistraci-Hebb for Paths of Length 2 (CH2-L2)

The Cannistraci-Hebb (CH) network automata model has been recently reformulated in a more comprehensive view in [35], [36]. It is a local-based, parameter-free and model-based deterministic rule for topological link prediction in both monopartite [6] and bipartite networks [11], [19]. It is based on the Local Community Paradigm (LCP) which is a bioinspired theory recently proposed in order to model local-topology-dependent link-growth in a class of real complex networks characterized by the development of diverse, overlapping and hierarchically organized local-communities [6]. Being a local-community-based method, it assigns to every candidate interaction a likelihood score looking only at the neighboring nodes, their cross-interactions and their interactions with the other nodes. Thus, for the CH network automaton model the likelihood of a new link to appear is function not only of the number of common neighbors but also function of the internal Local Community Links (iLCLs) and of the external Local Community Links (eLCLs) which are, respectively, the number of interactions between the common neighbors and the number of interactions of the common neighbors with nodes external to the local community. The first formalization of the CH model (called CH1-L2 in [35]) was putting more emphasis on the information content related with the common neighbors and the interactions between them (the *iLCLs*); the second formalization (i.e. the CH2-L2) takes into account that the local isolation of the operational units in the different local communities is equally important to carve the LCP architecture in the network, and this is guaranteed by the fact that the common neighbors minimize their interactions external to the local community (the *eLCLs*). It has been proved that CH2-L2 generally achieves higher performance than CH1-L2 [35]. The mathematical formalization of CH2-L2 in order to explicitly take into account also the *minimization of the external links* is then:

$$CH2\_L2(i,j) = \sum_{k \in L_2} \frac{1 + di_k}{1 + de_k}$$

Where the summation is executed over all the paths of Length 2 (L2) and:

- i and j: seed nodes of the candidate interaction
- k: intermediate node on the considered path of length two (L_2)
- di_k: internal degree of node k (number of *iLCLs*)
- de_k: external degree of node k (number of *eLCLs*)

Note that a unitary term is added to the numerator and denominator to avoid the saturation of the value in case of *iLCLs* or *eLCLs* equal to zero.

The higher the score, the higher the likelihood that the interaction exists, therefore the candidate interactions are ranked by decreasing CH2-L2 scores and the obtained ranking is the link prediction result.

The computational complexity of the CH methods, and in particular of CH2-L2, is $O(EN(1 - D))$, where N and E are the number of nodes and links in the network, and $D = \frac{2E}{N(N-1)}$ is the network density. However, in the domain of real and practical problems in which topological link prediction is applied, the complexity of CH2-L2 can be more simply expressed as $O(EN)$, and frequently approximated to $O(N^2)$, for further details please refer to the Appendix A.

Note that the link likelihoods are computed independently from each other and therefore the implementation can be easily run in parallel in order to speed up the running time.

The method has been implemented in Matlab.

2.1.2 Resource Allocation for Paths of Length 3 (RA-L3)

The concept of paths of Length 3 (L3) is very intuitive and can be summarized as: the likelihood of two nodes in the network to be connected grows with the number of L3 paths between them. This idea can be seen as a natural extension of the Common Neighbours (CN) similarities or the Triadic Closure Property (TCP) [37] on monopartite networks, and has already been exploited by some link prediction algorithms such as the Katz index [38] or the Local Path [39] metric.
TCP is historically rooted in social network analysis, namely the more friends in common, the more likely is that two individuals will know each other. This principle has turned out to be applicable to networks of various nature and is widely used also for biological ones.
However, it has been recently shown that TCP is not valid for many classes of networks and in particular for the vast majority of Protein-to-Protein Interaction (PPI) networks [40]. Motivated by this, it has been found that a modification of the L3 principle, in order to keep into account a degree normalization, is able to outperform existing link prediction algorithms on PPIs and possibly other types of networks. In this latter version, which constitutes an extension of Resource Allocation (RA) to path of length three (called RA-L3), the similarities are rescaled as:

$$RA\_L3 = \sum_{k_1 k_2 \in L_3} \frac{1}{\sqrt{d_{k_1} \cdot d_{k_2}}}$$

Where i and j are the two nodes of the candidate interaction; k_1 and k_2 are the intermediate nodes on the considered path of length 3 (L_3); d_{k_1} and d_{k_2} are the respective node degrees and the summation is executed over all the paths of length 3. The degree normalization corrects the bias induced by high degree nodes, i.e., the hubs, which are responsible of the creation of multiple short paths in the network.
The results found on PPI networks were promising but the evaluation method used should be further investigated since the authors removed 50% of the links completely destroying the network structure. This criticality has already been addressed and discussed in [35].
The code of this method has been implemented in Matlab.

2.2 Global Methods

2.2.1 Structural Perturbation Method (SPM)

SPM is a structural perturbation method that relies on a theory similar to the first-order perturbation in quantum mechanics [23]. It is a global approach, meaning that it exploits the information of the complete adjacency matrix in order to compute the likelihood score to assign to every candidate interaction. A high-level description of the procedure is the following:

1. Randomly remove a subset of the edges (usually 10%) from the network adjacency matrix ΔE, obtaining a reduced adjacency matrix x^R.

2. Compute the eigenvalues and eigenvectors of x^R.

3. Considering ΔE as a perturbation of x^R, construct the perturbed matrix $\widetilde{x}$ via a first-order approximation that allows the eigenvalues to change while keeping fixed the eigenvectors.

4. Repeat steps 1-3 for 10 independent iterations and take the average of the perturbed matrices $\widetilde{x}$.

The idea behind the method is that a missing part of the network is predictable if it does not significantly change the structural features of the observable part, represented by the eigenvectors of the matrix. If this is the case, the perturbed matrices should be good approximations of the original network [23]. The entries of the average perturbed matrix represents the scores for the candidate links. The higher the score the greater the likelihood that the interaction exists, therefore the candidate interactions are ranked by decreasing scores and the obtained ranking represents the link prediction result. The success and the feasibility of SPM is based on the strong correlation between independent perturbations, which indicates that the missing links, which are considered as unknown information in this setting, can be recovered by perturbing the network with another set of known links [23].
The computational complexity of the SPM method is $O(kN^3)$, where k is the number of iterations and N is the number of nodes in the network. In fact, every iteration is dominated by the eigen-decomposition of the perturbed adjacency matrix, which requires $O(N^3)$. However, k is usually a small constant (e.g. 10) and the iterations, since independent from each other, can be executed in parallel in order to speed up the running time.
The Matlab implementation of the method has been provided by the authors.

2.2.2 Stochastic Block Model (SBM)

The framework based on Stochastic Block Model (SBM) considered in this study has been introduced by Guimerà et al. [21] in order to identify both missing and spurious interactions in noisy network observations. The general idea of a block model is that the nodes are partitioned into groups and the probability that two nodes are connected depends only on the groups to which they belong. Assuming that there is no prior knowledge about which partition is more suitable for the observed network, the mathematical formula for obtaining the reliability of an individual link between nodes i and j is [21]:

$$R_{ij} = \frac{1}{Z}\sum_{p\in P}\left(\frac{l_{\sigma_i\sigma_j}+1}{r_{\sigma_i\sigma_j}+2}\right)\exp([-H(p)]$$

Where the sum is over every partition p in the space P of all the possible partitions of the network into groups, σ_i is the group of node i in partition p, $l_{\alpha\beta}$ is the number of links between groups α and β, $r_{\alpha\beta}$ is the maximum number of possible links between groups α and β. The function $H(p)$ is:

$$H(p) = \sum_{\alpha\leq\beta}\left[\ln(r_{\alpha\beta}+1)+\ln\binom{r_{\alpha\beta}}{l_{\alpha\beta}}\right]$$

And the normalization factor is:

$$Z = \sum_{p\in P}\exp[-H(p)]$$

However, since the exploration of all the possible partitions of the network into groups is often too computationally expensive even for small-size networks, the Metropolis algorithm, which is based on a stochastic procedure, is exploited in order to sample only a subset of partitions that are relevant for the estimation of the link reliability [21]. The higher the reliability the greater the likelihood that a non-observed interaction actually exists, therefore the candidate interactions are ranked by decreasing scores and the obtained ranking represents the link prediction result. The C code of the method has been released by the authors and can be download from the website `http://seeslab.info/downloads/network-c-libraries-rgraph/`.

2.2.3 SBM Degree Corrected and/or Nested (SBM DC and/or N)

The concept of Degree Corrected SBM (SBM DC) has been introduced for community detection tasks in [31] and for prediction of spurious and missing links in [32], in order to keep into account the variations in node degree typically exhibited in real networks.

The nested version of the DC SBM algorithm has been introduced by T. P. Peixoto in 2014 [30] in order to overcome two major limitations of simple SBM: namely the inability to separate true structures from noise and the inability to detect smaller but well-defined clusters as network size become large. The nested structure is built as a multigraph where a maximum of L hierarchical layers represent progressively lower resolution replicas of the original network, in order to detect even small well-clustered communities. Although these variations were originally proposed for community detection, it was mentioned that they should be even useful to enhance link prediction scores, but they have not been extensively tested yet for this task (most likely due to the high computational time constraints imposed).

Similarly to SBM, also in the DC and/or N SBM a network partitioning is needed and in the implementation considered in this study an optimized Monte Carlo Markov Chain (MCMC) method is used to sample the space of the possible partitions [41]. Its performance using the sampling are often indistinguishable from the original method but come at a much lower computational demand.
The C++/Python code of the methods has been released from the author and can be found at the website `http://graph-tool.skewed.de/` [42].

2.2.4 Fast Probability Block Model (FBM)

Fast probability Block Model (FBM) is a global method based on the same network partitioning theory as SBM, but it replaces the Metropolis algorithm introducing a greedy strategy for an efficient sampling over the space of the possible partitions, which leads to high improvements in the computational time [22].
For each network, 50 partitions are sampled according to the following procedure [22]

1. As first, the network is randomly partitioned in two blocks.
2. Then, for each block, until all its edges have been considered, the maximum clique is iteratively removed and it represents a group for the current partitioning.
3. At the end of the iterative removal a set of low degree nodes will remain without forming any clique, they are treated as a separate special group having low inner link density.

The procedure is explained more in detail in [22].
Given the sampled partitions, the following mathematical formula is used in order to compute the likelihood of the non-observed links [22]:

$$R_{ij} = \frac{1}{|P|} \sum_{p \in P} F(\sigma_i, \sigma_j)$$

$$F(\alpha, \beta) = \begin{cases} \dfrac{r_\alpha}{2r_\alpha - l_\alpha}, & \text{if } \alpha = \beta \\ \dfrac{l}{r_{\alpha\beta} - l_{\alpha\beta}}, & \text{if } \alpha \neq \beta \end{cases}$$

Where the sum is over every partition p in the set P of sampled partitions, σ_i is the group of node i in partition p, l_α is the number of links within group α, r_α is the maximum number of possible links within group α, $l_{\alpha\beta}$ is the number of links between groups α and β, $r_{\alpha\beta}$ is the maximum number of possible links between groups α and β.
The higher the reliability, the greater the likelihood that a non-observed interaction actually

exists, therefore the candidate interactions are ranked by decreasing scores and the obtained ranking represents the link prediction result. The discussed modifications to the standard SBM algorithm make FBM the only SBM-based method feasible to predict missing links on networks formed by up to around ten thousand nodes. The Matlab implementation of the method has been provided by the authors.

3 | Benchmark of Real and Synthetic Networks

In this Chapter the generative procedures for the synthetic benchmarks considered in this study are presented. The real networks are briefly introduced outlining the meaning of their interactions, their source and some networks measures and statistics.

3.1 Generation of Synthetic Networks

3.1.1 Generation of nonuniform Popularity-Similarity Optimization (nPSO) Networks

The nonuniform Popularity-Similarity Optimization (nPSO) model [43] is a variation of the Popularity-Similarity Optimization (PSO) model [44] introduced in order to confer to the generated networks an adequate community structure, which is lacking in the original model. Since the connection probabilities are inversely proportional to the hyperbolic distances, a uniform distribution of the nodes over the hyperbolic disc (as in the PSO model) does not create agglomerates of nodes that are concentrated on angular sectors and that are more densely connected between each other than with the rest of the network. A nonuniform distribution (as in the nPSO model), instead, allows to do it by generating heterogeneity in the angular node arrangement. Given the parameters of the PSO model (number of nodes N, half of the average degree m, temperature T inversely related to the clustering, power-law exponent γ) and a nonuniform probability distribution defined in $[0, 2\pi[$, the procedure to generate a network is the same as for the PSO case, with the only difference that the angular coordinates of the nodes are not sampled uniformly but according to the given nonuniform probability distribution.
Therefore building a network in the hyperbolic disc requires the following steps:

1. Initially the network is empty;

2. At time $i = 1, 2, ..., N$ a new node i appears with radial coordinate $r_i = 2\ln(i)$ and angular coordinate θ_i sampled in $[0, 2\pi[$ accordingly to a desired nonuniform probability distribution; all the existing nodes $j < i$ increase their radial coordinates according to $r_j(i) = \beta r_j + (1-\beta)r_i$ in order to simulate popularity fading;

3. If $T = 0$, the new node connects to the m hyperbolically closest nodes; if $T > 0$, the new node picks a randomly chosen existing node $j < i$ and, given that it is not already connected to it, it connects to it with probability

$$p(i,j) = \frac{1}{1 + \exp\left(\dfrac{h_{ij} - R_i}{2T}\right)}$$

   repeating the procedure until it becomes connected to m nodes.
   Note that

$$R_i = r_i - 2\ln\left[\frac{2T(1 - \exp(-(1-\beta)\ln(i)))}{\sin(T\pi)m(1-\beta)}\right]$$

   is the current radius of the hyperbolic disc, and

$$h_{ij} = \cosh^{-1}(\cosh r_i \cosh r_j - \sinh r_i \sinh r_j \cos\theta_{ij})$$

   is the hyperbolic distance between node i and node j, where

$$\theta_{ij} = \pi - \mid \pi - \mid \theta_i - \theta_j \mid\mid$$

   is the angle between these nodes.

4. The growing process stops when N nodes have been introduced.

In this study, without loss of generality, Gaussian mixture distributions will be considered, with communities that emerge in correspondence of the different components. A Gaussian mixture distribution is characterized by the following parameters [45]:

- $C > 0$, which is the number of components, each one representative of a community;

- $\mu_{1...C} \in [0, 2\pi[$, which are the means of the components, representing the central locations of the communities in the angular space;

- $\sigma_{1...C} > 0$, which are the standard deviations of the components, determining how much the communities are spread in the angular space; a low value leads to isolated communities, a high value makes the adjacent communities to overlap;

- $\rho_{1...C}(\sum_i \rho_i = 1)$, which are the mixing proportions of the components, determining the relative sizes of the communities.

Note that, although the means of the components are located in $[0, 2\pi[$, the sampling of the angular coordinate θ can fall out of this range. In this case, it has to be shifted within the original range using the modulo operator: $\theta = \text{modulo}(\theta, 2\pi)$.
Although the parameters of the Gaussian mixture distribution allow for the investigation of disparate scenarios, this study applies the most straightforward setting. For a given number of components C, their means are considered to be equidistantly arranged over the angular space, the standard deviation and the mixing proportions are set equal for every component:

- $\mu_i = \frac{2\pi}{C} \cdot (i-1) \quad i = 1...C$
- $\sigma_1 = \sigma_2 = ... = \sigma_C = \sigma$
- $\rho_1 = \rho_2 = ... = \rho_C = \frac{1}{C}$

In particular, in the simulations presented in this study the standard deviation is fixed to 1/6 of the distance between two adjacent means ($\sigma = \frac{1}{6} \cdot \frac{2\pi}{C}$), which allowed for a reasonable isolation of the communities independently from their number. The community memberships are assigned considering for each node the component whose mean is at the lowest angular distance.

3.1.2 Generation of Random Geometric Graph (RGG) Networks

The basic version in two dimensions of the Random Geometric Graph (RGG) model dates back to 1961 [46] by Gilbert, and it has been subsequently extended to a generic number of dimensions in 2002 [47]. It has two input parameters: the number of nodes N and the threshold distance (i.e. the radius of the neighborhood) $r \in [0, 1]$. In the proposed simulations N points are placed uniformly at random in a unitary disc in the Euclidean space. A link between two points exists if their relative Euclidean distance is at most r. Naturally, this model generates networks with underlying Euclidean geometry.

3.1.3 Generation of Watts-Strogatz (WS) Networks

The Watts-Strogatz (WS) model [48], proposed in 1998, introduced the concept of small-world networks, strongly clustered as regular lattices and with a small characteristic path length like random graphs, arguing that many real networks are somewhere between these two extreme topological configurations.

It has three input parameters: N, which is the number of nodes; $m > 0$, representing half of the average node degree and therefore defining the number of edges $E = mN$; $\beta \in [0, 1]$, which is the rewiring probability.
The procedure to generate a network requires the following steps:

1. Create a ring lattice of N nodes, assuming the nodes ordered in a circular list and connecting each of them to its m next and previous neighbors;

2. For every node, consider each edge to the m next neighbors and rewire it with probability β. The new target node is chosen uniformly at random, avoiding self-loops and link duplication.

Tuning the parameter β allows to generate networks with characteristics between regularity ($\beta = 0$, no links rewired) and randomness ($\beta = 1$, all the links rewired). In particular, Watts and Strogatz [48] showed how, starting from a ring lattice, the introduction of even a few (small β) short-cuts leads to an immediate drop in the characteristic path length, whereas the high clustering coefficient remains practically unchanged. They are random networks with non-scale-free node distribution.

3.1.4 Generation of Lancichinetti-Fortunato-Radicchi (LFR) Networks

The Lancichinetti-Fortunato-Radicchi (LFR) model [49], proposed in 2008, tries to correct the heterogeneity of both node degree and community size distributions typically displayed in real-world networks and not accounted by previous generative models. Such heterogeneity has been proven to be responsible of many important network features and it is accounted by assuming that both the degree and the community size are power-laws with exponents γ and β respectively. The LFR benchmark can be seen as a special version of the degree-corrected stochastic block model [31], with the degree and the block size distributed according to truncated power laws [50]. This model has in total eight input parameters: the number of nodes N, the average node degree $k > 0$, the maximum node degree $maxk$, the mixing parameter μ which represents the fraction of links shared by a node with other nodes outside its community, the aforementioned power law exponents γ and β, the minimal and maximal community sizes $minc$ and $maxc$ respectively. As an optional parameter a desired value for the average clustering coefficient C can be specified. The iterative procedure to generate a network is briefly reported in the following and it is explained in detail in [49]:

1. first, a degree from a truncated power law distribution with exponent γ is assigned to each node. The value for $mink$ is chosen so that to satisfy the constraints on the average node

degree k and on the maximum node degree $maxk$. The nodes are connected keeping their degree sequence;

2. each node shares a fraction $1 - \mu$ of its links with the other nodes of its community and a fraction on μ links with the other nodes of the network;
3. the sizes of the communities are taken from a truncated power law with parameter β such that the sum of all sizes is equal to the number N of nodes;
4. at the beginning all the nodes are not assigned to any community. At every iteration, until all the nodes are inspected, a node is assigned to a randomly picked community until the maximum size of the community is reached;
5. optional rewiring steps may be needed in order to satisfy the wished values of μ and C, such that the degrees of all the nodes remain the same and only the split between internal and external degree is affected.

The C++ code of the method has been released by the authors and can be downloaded from the website `http://sites.google.com/site/santofortunato/inthepress2`.

3.2 Real Networks Dataset

All the real networks have been transformed into undirected and unweighted, self-loops have been removed and the largest connected component has been considered. A brief description of the small, large and brain real-networks used throughout the thesis follows.

Mouse neural: in-vivo single neuron connectome that reports mouse primary visual cortex (layers 1, 2/3 and upper 4) synaptic connections between neurons [51].

Karate: social network of a university karate club collected by Wayne Zachary in 1977. Each node represents a member of the club and each edge represents a tie between two members of the club [52].

St. Marks foodweb: carbon-flow network of a seagrass ecosystem constructed from a comprehensive database collected at three different sites during January and February 1994 from the St. Marks Wildlife Refuge, situated in Apalachee Bay in the north-eastern Gulf of Mexico. The network was then constructed by averaging the food web for each month and the overall winter. The network consists of 51 compartments [53].

Dolphins: a social network of bottlenose dolphins. The nodes are the bottlenose dolphins (genus Tursiops) of a bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand. An edge indicates a frequent association. The dolphins were observed between 1994 and 2001 [54].

Ythan foodweb: The food web for Ythan Estuary on the North Sea near Aberdeen, Scotland [55]. Nodes are autotrophs, herbivores, carnivores and decomposers; links represent food sources.

Macaque neural: a macaque cortical connectome, assembled in previous studies in order to merge partial information obtained from disparate literature and database sources [56].

Polbooks: nodes represent books about US politics sold by the online bookseller Amazon.com. Edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon. The network was compiled by V. Krebs and is unpublished, but can be found at `http://www-personal.umich.edu/ mejn/netdata/`.

SEA terrorist: a terrorist network of Southeast Asian Aggregate Attack Series. This is a single collapsed network of all the individual Indonesian cases (before 2005). The network was significantly enriched by Sidney Jones and Ken Ward and is unpublished, but can be found at `http://doitapps.jjay.cuny.edu/jjatt/data.php`.

ACM2009_ contacts: network of face-to-face contacts (active for at least 20 seconds) of the attendees of the ACM Conference on Hypertext and Hypermedia 2009 [57].

Football: network of American football games between Division IA colleges during regular season Fall 2000 [58].

Physicians innovation: the network captures innovation spread among physicians in the towns in Illinois, Peoria, Bloomington, Quincy and Galesburg. The data were collected in 1966. A node represents a physician and an edge between two physicians shows that the left physician told that the right physician is his friend or that he turns to the right physician if he needs advice or is interested in a discussion [59].

AQ terrorist: this is the largest connected component of the aggregated Al Qaeda Operations Attack Series. The network represents the relations of individuals associated with over 10 attacks teams deployed by Al Qaeda over a decade, from 1993 to 2003 [60].

Manufacturing email: email communication network between employees of a mid-sized manufacturing company [61].

Jazz: collaboration network between Jazz musicians. Each node is a Jazz musician and an edge denotes that two musicians have played together in a band. The data were collected in 2003 [62].

Residence hall friends: friendship network between residents living at a residence hall located on the Australian National University campus [63].

Rhesus brain: Network of interactions among cortical regions in the macaque brain (genus Rhesus), extracted from 410 tract tracing studies collated in the CoCoMac database [64].

Van der Waals: van der Waals contact network of human TriosephosphateIsoMerase (TIM) barrel [65].

Haggle contacts: contacts between people measured by carried wireless devices. A node represents a person and an edge between two persons shows that there was a contact between them [66].

Worm nervous: a C. Elegans connectome representing synaptic interactions between neurons [48].

US Air: Network of air flights in the US as it was in 1997. The network is unpublished and it is available in the Pajek database `http://vlado.fmf.uni-lj.si/pub/networks/data/`.

Netsci: a co-authorship network of scientists working on networks science [67].

Infectious contacts: network of face-to-face contacts (active for at least 20 seconds) of people during the exhibition 'INFECTIOUS: STAY AWAY' in 2009 at the Science Gallery in Dublin [57].

Flightmap: a network of flights between American and Canadian cities [68].

Email: email communication network at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. Nodes are users and each edge represents that at least one email was sent between each other [69].

Polblog: a network of front-page hyperlinks between blogs in the context of the 2004 US election. A node represents a blog and an edge represents a hyperlink between two blogs [70].

Odlis: Online Dictionary of Library and Information Science (ODLIS): ODLIS is designed to be a hypertext reference resource for library and information science professionals, university students and faculty, and users of all types of libraries. Version December 2000 [71].

Advogato: a trust network of the online community platform Advogato for developers of free software launched in 1999. Nodes are users of Advogato and the edges represent trust relationships [72].

Arxiv astroph: collaboration graph of authors of scientific papers from the arXiv's Astrophysics (astro-ph) section. An edge between two authors represents a common publication [73].

Thesaurus: this is the Edinburgh Associative Thesaurus. Nodes are English words and a link denotes that one word was given as a response to the other stimulus word in user experiments [74].

Arxiv hepth: this is the network of publications in the arXiv's High Energy Physics – Theory (hep-th) section. The links that connect the publications are citations [73].

Facebook: a network of a small subset of posts to user's walls on Facebook. The nodes of the network are Facebook users, and each edge represents one post, linking the users writing a post to the users whose wall the post is written on [75].

ARK200909-ARK201012: six Autonomous Systems (AS) Internet topologies extracted from the data collected by the Archipelago active measurement infrastructure (ARK) developed by CAIDA, from September 2009 up to December 2010 at timesteps of three months. The connections in the topology are not physical but logical, representing AS relationships [76].

Van den Heuvel: this is a dataset of 486 structural human brain networks coming from healthy-controls. It was constructed from the T1 and diffusion weighted imaging data of the Human Connectome Project (WU-Minn HCP Data - 1200 Subjects release) [77], [78]. Individual connectomes have been provided by the Dutch Connectome Lab, Utrecht, Netherlands and have been created following a procedure as explained in [79], [80].

All the networks in the dataset, unless explicitly mentioned the source, can be downloaded from the Koblenz Network Collection at `http://konect.uni-koblenz.de`.

Several statistics of these networks are shown in Table 3.1 divided for small-size networks (up to *polblog*), for the brain networks (*van den Heuvel* dataset) and for large-size networks (further subdivided into internet networks, named with the *ARK* prefix, and into other large-size real networks). For each network, several statistics have been computed. N is the number of nodes. E is the number of edges. The parameter k refers to the average node degree. D is the network density. C is the average clustering coefficient, computed for each node as the number of links between its neighbors over the number of possible links [48]. L is the characteristic path length of the network [48]. *LCP-corr* is the Local-Community-Paradigm correlation [6], representing the correlation between the number of common-neighbors and the number of links between them, looking at each pair of connected nodes in the network. *Struct-cons* is the structural consistency [23], a quantitative index that estimates the link predictability of the network. *Power-law* is the exponent γ of the power-law distribution estimated from the observed degree distribution of the network using the maximum likelihood procedure described in [81].
The measures related to the *van den Heuvel* dataset (i.e. brain networks) are expressed as averaged values, but the number of nodes is fixed across the networks of the dataset.

Table 3.1: Statistics of real-world networks.

Several statistics of the real networks organized as follows: the first 25 entries are small-size real networks (up to *polblog*), then the averaged statistics for the 486 brain networks (*van den Heuvel* dataset) are reported. From *ARK200909* to the end are large-size real networks, further subdivided into internet networks, named with the *ARK* prefix, and into other large networks.

	N	E	k	D	C	L	LCP corr	Struct cons	Power law
mouse_ neural	18	37	4.1	0.24	0.22	2.0	0.91	0.41	4.0
karate	34	78	4.6	0.14	0.57	2.4	0.76	0.42	2.1
stmarks_ foodweb	54	350	13.0	0.24	0.41	1.8	0.91	0.40	4.0
dolphins	62	159	5.1	0.08	0.26	3.4	0.91	0.37	7.0
ythan_ foodweb	92	414	9.0	0.10	0.22	2.3	0.90	0.38	3.0
macaque_ neural	94	1515	32.2	0.35	0.77	1.8	0.97	0.76	4.5
polbooks	105	441	8.4	0.08	0.49	3.1	0.94	0.31	2.6
SEA_ terrorist	108	565	10.5	0.10	0.71	2.6	0.97	0.57	2.9
ACM2009_ contacts	113	2196	38.9	0.35	0.53	1.7	0.97	0.33	3.7
football	115	613	10.7	0.09	0.40	2.5	0.89	0.45	9.1
physicians_ innovation	117	465	7.9	0.07	0.22	2.6	0.79	0.21	4.5
AQ_ terrorist	125	312	5.0	0.04	0.55	4.6	0.91	0.52	4.5
manufacturing_ email	167	3250	38.9	0.23	0.59	2.0	0.99	0.55	3.1
jazz	198	2742	27.7	0.14	0.62	2.2	0.95	0.70	4.5
residence_ hall_ friends	217	1839	16.9	0.08	0.36	2.4	0.90	0.35	6.3
rhesus_ brain	242	3054	25.2	0.10	0.45	2.2	0.96	0.38	4.2
vanderwaals	248	1003	8.1	0.03	0.48	4.5	0.87	0.39	10.0
haggle_ contacts	274	2124	15.5	0.06	0.63	2.4	0.99	0.60	1.5
worm_ nervoussys	297	2148	14.5	0.05	0.29	2.5	0.91	0.23	3.3
USAir	332	2126	12.8	0.04	0.63	2.7	0.98	0.49	1.8
netsci	379	914	4.8	0.01	0.74	6.0	0.92	0.59	3.4
infectious_ contacts	410	2765	13.5	0.03	0.46	3.6	0.95	0.41	6.4
flightmap	456	37947	166.4	0.37	0.81	1.6	0.99	0.78	1.7
email	1133	5451	9.6	0.01	0.22	3.6	0.85	0.18	4.9
polblog	1222	16714	27.4	0.02	0.32	2.7	0.93	0.25	2.4
van den Heuvel	82	1164	28.4	0.35	0.65	1.7	0.98	0.56	4.3
ARK200909	24091	59531	4.9	0.0002	0.36	3.5	0.95	0.10	2.1
ARK200912	25910	63435	4.9	0.0002	0.36	3.5	0.94	0.10	2.1
ARK201003	26307	66089	5.0	0.0002	0.37	3.5	0.94	0.10	2.3
ARK201006	26756	68150	5.1	0.0002	0.37	3.5	0.95	0.09	2.1
ARK201009	28353	73722	5.2	0.0002	0.37	3.5	0.94	0.10	2.2
ARK201012	29333	78054	5.3	0.0002	0.38	3.5	0.95	0.10	2.2
odlis	2898	16376	11.3	0.0039	0.30	3.2	0.93	0.10	2.6
advogato	5042	39227	15.6	0.0031	0.25	3.3	0.90	0.16	2.7
arxiv astroph	17903	196972	22.0	0.0012	0.63	4.2	0.95	0.67	2.8
thesaurus	23132	297094	25.7	0.0011	0.09	3.5	0.87	0.07	2.8
arxiv hepth	27400	352021	25.7	0.0009	0.31	4.3	0.92	0.27	2.9
facebook	43953	182384	8.3	0.0002	0.11	5.6	0.87	0.09	3.7

4 | Results

In this link prediction investigation we decided to focus the attention on four state-of-the-art approaches to be compared with SBM and its variations. They are based on completely different theories that will be now concisely introduced together with the explanation of their choice (for further details please refer to Chapter 2).

The first method we considered is the Structural Perturbation Method (SPM), a global approach that relies on a theory similar to the first-order perturbation in quantum mechanics [23]. It implements a perturbation procedure based on the idea that a missing part of the network is predictable if it does not significantly change the structural features of the observable part, represented by the eigenvectors of the matrix. Therefore, assuming the perturbed matrices to be good approximations of the original adjacency matrix, they are exploited for assigning likelihood scores to the non-observed interactions [23]. The original publication [23] already suggested SPM to be a promising method able to clearly outperform SBM. A recent study [82] has confirmed that SPM is actually one of the best performing state-of-the-art global approaches for topological link prediction.

The second method we considered is the CH2-L2 network automaton model described in [35], [36]. This local-learning rule can also be exploited for link prediction and it states that the local-community organization (i.e. aggregation of linked common neighbors) increases the likelihood that a set of nodes connects together because they are confined in the same local community, consequently also the likelihood that they will create new connections inside the community is increased by the mere structure of the network topology.

The third method we considered is the extension to path of length three of Resource Allocation (we will call it as RA-L3) which has been recently shown to outperform other link prediction methods for protein-to-protein interactions, foodwebs and trade networks [35], [40].

The fourth method we considered is Fast probability Block Model (FBM), a global method based on the same network partitioning theory as SBM, but it replaces the Metropolis algorithm introducing a greedy stochastic strategy for an efficient sampling over the space of the possible partitions, which leads to high improvements in the computational time [22]. This is the only

way to have a SBM-based method somehow scalable for larger networks.
Other methods belonging to the SBM family we exploited are the Degree Corrected and/or Nested SBM (SBM DC/N). The SBM DC introduces a penalization for high-degree nodes through a normalization over the expected degrees of the vertices [32], [34]. The Nested SBM (SBM N) considers as vertices the blocks of nodes, which, in the SBM, have the same probabilities to connect. In this way more hierarchical resolutions of the network are constructed and those can be nested in a multigraph [30]. Also the Degree Corrected and Nested SBM (SBM DC N) variation is possible and should incorporate the degree variability inside each block [30].
To sum up, we focused our comparison to SBM choosing the two methods that recent studies have demonstrated to be the best performing respectively for the global and the local approaches (i.e. SPM and CH2-L2), the promising local approach of RA-L3 already successful in some fields of study, the faster variant of SBM (FBM) and other three variations of SBM which could improve the accuracy results. All the methods have been tested on both artificial and real complex networks and the results will be now discussed.

When no information is available about missing or future interactions, the standard procedure adopted for evaluating the link prediction performance on a given network is the following:

1. a certain number r of links are randomly removed from the network;
2. the algorithm is executed in order to obtain a ranking of the non-observed links in the reduced network by decreasing likelihood scores;
3. the precision is computed as the percentage of removed links among the top-r in the ranking;
4. the previous steps are repeated for several iterations and the average precision is reported as measure of performance for the algorithm on the given network.

A common and accepted practice that have also been adopted in this study is to set r equal to 10% of the links in the network. A 10% removal is commonly accepted [6], [23] because it is proven to generate missing interactions in the network without significantly affecting the main topological properties. Larger removal percentages can cancel important topological information such as local-community organization [6]. For the methods SPM, CH2-L2, RA-L3 and FBM the evaluation procedure has been repeated for 100 iterations, whereas for the other SBM-based methods it has been limited to 10 iterations due to the high computational time.
Another evaluation metric widely used is the area under the ROC curve (AUC-ROC or, short name, just AUC). However, a recent comprehensive investigation focused on the problem of link prediction evaluation [83], followed and supported by successive link prediction studies [11], [82],

has pointed out how AUC can be deceptive and it has strongly advised against the adoption of this metric. The first reason is that AUC should be used for the evaluation of a classification problem in which a positive and a negative set are present, but in a link prediction problem it is not appropriate to consider the non-observed links as negative links, therefore a negative set cannot be well defined. Secondly, even if it were considered a classification problem, it would be characterized by an extreme imbalance between the positive and negative sets, since most of the real networks are sparse. In this situation ROC curves and their areas fail to honestly convey, represent and quantify the difficulty of the prediction problem, leading to exceptionally high scores even when the precision would be particularly low [83], therefore we further discourage its usage.

4.1 Evaluation on Real Complex Networks

In order to compare the performance of the link prediction methods analyzed in this study, a dataset of 25 small-size networks has been collected from different real-world domains and a set of 486 structural brain networks has been adopted. Due to the computational time constraints imposed by the SBM family, only networks of size up to around one thousand nodes have been considered for all the methods. Several statistics of the real networks are shown in Table 3.1 and an explanation of the meaning of the networks is reported in Section 3.2. The dataset is intended to cover topologies having as much as possible different characteristics, in order to avoid to favor methods tending to perform better in presence of particular structural properties.

Since to the best of our knowledge this is the first study to exploit the degree corrected and/or nested version of the SBM for link prediction tasks, the influence on precision and on computational time of the number of sampling from the distribution of all the possible partitions of the networks into groups (called *sweeps*) has been widely tested.
A detailed comparison ran on the same workstation (see Appendix B for further details) is reported in Table 4.1, 4.2 and 4.3 respectively for SBM DC N, SBM DC and SBM N.
In these tables, for each network, 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. In order to evaluate the performance, 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. The table reports for each network the mean precision and the mean computational time (in hours) over the random iterations for the entire dataset. The real networks are sorted by increasing number of nodes N. Note that the values are not reported for the cases in which the computation was too expensive.

First of all, it is immediate to notice that the degree correction operation is not time demand-

Table 4.1: Precision-time evaluation of link prediction on small-size real networks using the Degree Corrected and Nested SBM.

	Precision					**Time [h]**				
# of sweeps:	**25**	**100**	**500**	**1000**	**5000**	**25**	**100**	**500**	**1000**	**5000**
mouse neural	0.05	0.05	0.08	0.05	0.05	0	0	0	1	3
karate	0.13	0.16	0.18	0.18	0.19	0	0	1	2	10
stmarks_foodweb	0.19	0.19	0.21	0.21		0	0	2	4	
dolphins	0.08	0.08	0.11	0.09		0	1	4	7	
ythan_foodweb	0.20	0.22	0.22	0.22		0	2	8	16	
macacque	0.38	0.38	0.38	0.39		0	1	7	14	
polbooks	0.15	0.15				1	2			
SEA_terroris	0.29	0.29				1	2			
ACM2009_contacts	0.19	0.19				0	2			
football	0.28	0.27				1	2			
physicians_innovation	0.02	0.02				1	3			
AQ_terrorist	0.16	0.17				1	3			
manufacturing_email	0.37	0.37				1	5			
jazz	0.37	0.36				2	9			
residence_hall_friends	0.15	0.14				2	10			
haggle_contacts	0.23	0.24				3	13			
rhesus_brain	0.09	0.10				3	15			
vanderwaals	0.45	0.44				4	18			
worm_nervoussys	0.15	0.15				6	21			
USAir	0.39	0.38				8	27			
netsci	0.25	0.25				8	36			
infectious_contacts	0.19	0.21				10	47			
flightmap	0.55					118				
email	0.10					220				
polblog	0.20					252				

ing, while, instead, the nested procedure is heavily time consuming due to the construction of the hierarchical representation of the networks. Furthermore we can see that varying the number of sweeps from 25 to e.g. 5000 does not significantly affect the overall precision score but it influences dramatically the computational time required (making the computation always not feasible). Some fluctuations on the precision results are due to the stochastic sampling over the set of all the possible networks' partitions. These results can be appreciated by looking at all the three tables presented and comparing them one versus the other. For these considerations we decided to always set the number of sweeps to 25 for all the evaluation frameworks considered in this study.

Table 4.4 reports the precision evaluation of the eight methods considered for each real network. The maximum level of precision reached on the different networks is quite variable, going from 0.08 in *physicians innovation* up to 0.75 in *flightmap*. Looking at the best methods

Table 4.2: Precision-time evaluation of link prediction on small-size real networks using the Degree Corrected SBM.

	Precision					**Time [h]**				
# of sweeps:	**25**	**100**	**500**	**1000**	**5000**	**25**	**100**	**500**	**1000**	**5000**
mouse neural	0.03	0.05	0.05	0.05	0.05	0	0	0	0	0
karate	0.14	0.14	0.16	0.15	0.16	0	0	0	0	1
stmarks_foodweb	0.18	0.19	0.20	0.20	0.21	0	0	0	0	2
dolphins	0.08	0.10	0.08	0.08	0.09	0	0	0	1	3
ythan_foodweb	0.19	0.21	0.21	0.21	0.21	0	0	1	1	7
macacque	0.36	0.36	0.37	0.37	0.37	0	0	1	2	8
polbooks	0.14	0.15	0.15	0.15		0	0	1	2	
SEA_ terroris	0.25	0.26	0.27	0.25		0	0	1	2	
ACM2009_ contacts	0.19	0.19	0.19	0.19		0	0	1	3	
football	0.26	0.27	0.26	0.27		0	0	1	3	
physicians_ innovation	0.04	0.04	0.06	0.04		0	0	1	3	
AQ_ terrorist	0.12	0.11	0.11	0.12		0	0	1	3	
manufacturing_ email	0.37	0.37	0.37	0.38		0	1	2	8	
jazz	0.35	0.33	0.34			0	1	4		
residence_ hall_friends	0.15	0.15	0.15			0	1	4		
haggle_ contacts	0.22	0.23				0	2			
rhesus_ brain	0.05	0.07				0	1			
vanderwaals	0.44	0.45				1	2			
worm_ nervoussys	0.12	0.11				1	2			
USAir	0.37	0.38				1	3			
netsci	0.15	0.17				1	3			
infectious_ contacts	0.16	0.16				1	4			
flightmap	0.56	0.56				10	15			
email	0.08	0.08				18	47			
polblog	0.18	0.18				21	111			

for each network, highlighted in bold, it is evident that SPM obtains the highest performance in 12 out of 25 networks, followed by CH2-L2 which wins in 9 networks, whereas SBM reaches the best prediction in only 3 networks, all of them very small (the largest has only 54 nodes). It is immediate to notice that the degree correction and the nested hierarchical structure do not boost the accuracy of the plain SBM and should be discarded for making inference on real networks structure. Removing the three variations of SBM (namely SBM DC N, SBM DC and SBM N), the gap between the best and the worst method for each network is in general contained within a level up to 0.25, however, a few outliers can be noticed. The first one is *netsci* with a divergence of 0.41 between CH2-L2, the best method, and SBM, the worst. But much more relevant are the cases of the *foodwebs*, where it occurs an atypical discrepancy between SPM, SBM and RA-L3, generally high in precision, versus FBM and CH2-L2, generally low in precision. The low performance of L2-based methods with respect to the performance of L3-based methods have

Table 4.3: Precision-time evaluation of link prediction on small-size real networks using the Nested SBM.

	Precision					**Time [h]**				
# of sweeps:	**25**	**100**	**500**	**1000**	**5000**	**25**	**100**	**500**	**1000**	**5000**
mouse neural	0.00	0.00	0.00	0.00	0.00	0	0	0	1	2
karate	0.01	0.01	0.01	0.01	0.01	0	0	1	2	10
stmarks_foodweb	0.07	0.05	0.07	0.07		0	0	2	5	
dolphins	0.00	0.03	0.06	0.04		0	1	4	8	
ythan_foodweb	0.06	0.06	0.06	0.06		0	2	8	18	
macacque	0.12	0.12	0.12	0.12		0	1	7	14	
polbooks	0.01	0.03				1	2			
SEA_ terroris	0.08	0.07				1	2			
ACM2009_ contacts	0.05	0.04				0	2			
football	0.06	0.08				0	2			
physicians_ innovation	0.01	0.00				0	3			
AQ_ terrorist	0.03	0.03				0	3			
manufacturing_ email	0.09	0.08				1	5			
jazz	0.09	0.09				1	8			
residence_ hall_ friends	0.04	0.04				2	10			
haggle_ contacts	0.03					3				
rhesus_ brain	0.02					4				
vanderwaals	0.10					4				
worm_ nervoussys	0.03					5				
USAir	0.07					9				
netsci	0.05					8				
infectious_ contacts	0.07					9				
flightmap	0.12					98				
email	0.02					185				
polblog	0.18					225				

recently been addressed and explained in [35], where it has been shown that this is due to intrinsic networks' organizational rules. The *foodweb* networks offer a clear example for disproving the suitability of the mean precision as an overall metric of best performance across several networks. The mean precision, even if reported for the sake of completeness, is particularly sensitive to the presence of such networks in which certain topological properties favor the prediction only for some methods, creating a huge gap between the various performance. The introduction of a few of this kind of networks would certainly bias the comparison toward the methods that fit with them. In fact, if the mean precision is computed excluding some selected networks, the order of overall performance for the methods could be changed with respect to the one that includes the networks. However, we clarify that we are not suggesting to exclude in future studies peculiar networks that lead to an anomalous divergence in performance between the methods, in fact, they still represent real-world topologies and for a fair comparison the dataset should be as rich

Table 4.4: Precision evaluation of link prediction on small-size real networks. For each network, 10% of links have been randomly removed (10 iterations for SBM, SBM DC N, SBM DC and SBM N 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. In order to evaluate the performance, 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. The table reports for each network the mean precision over the random iterations, the mean precision and the mean ranking over the entire dataset. For each network the best method (or methods) is highlighted in bold. The networks are sorted by increasing number of nodes N.

	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**

and diverse as possible. On the contrary, we want to encourage to include such networks and use a more robust and reliable metric for assessing the overall performance, in order to establish a final ranking of the methods.

The evaluation we propose, already adopted in two recent link prediction studies [82], [84], is the precision-ranking. After the computation of the precision as previously described and shown in Table 4.4, the methods are ranked for each network by decreasing precision, considering an average rank in case of ties. The mean ranking of the methods over all the networks represents the final evaluation score, the values are reported as last row of Table 4.4. The best performing approach, as already deducible from the precision, results to be SPM, with an average ranking of 2.1. The second method is CH2-L2 with 2.9; the third approach is plain SBM with 3.6, then FBM with 4.1, RA-L3 with 4.3 and as lasts the SBM variations.

The introduction of the ranked values has attenuated the big gap of performance in the network and consequently the final scores appeared to be robust. This is actually the goal of the precision-ranking evaluation, when multiple methods are compared across several networks it prevents that a unique but consistent alteration in the set of networks will substantially subvert the overall evaluation. We stress that the ability of a method to obtain higher performance in multiple networks is an indicator of great robustness and adaptability of the approach to diverse topologies: we believe that these principles should obtain a higher consideration with respect to a method that offers a lower performance on many networks and rare peaks of outperformance in a few networks, which may be even due to overfitting toward certain structural features.

In order to check the statistical significance of the difference in performance between the methods, pairwise permutation tests (over 10000 iterations) for the mean have been performed using for each pair of methods the pairwise ranking values computed for each network. Table 4.5 reports for each pair of methods the p-value of the test, adjusted for multiple hypothesis comparison by the Benjamini–Hochberg correction. Considering a significance level of 0.05, the pairs whose mean performance are not significantly different are SPM and CH2-L2, CH2-L2 and SBM, SBM and FBM, SBM and RA-L3, FBM and RA-L3 which are actually the ones with a difference in the mean ranking constrained within around one ranking-position. This result corroborates the classification of SPM as the best state-of-the-art global method being statistically different from SBM, the second global method. It has to be noticed that CH2-L2, the higher local approach, obtained an overall score of performance higher than two global methods like SBM (not statistically significant) and FBM (statistically significant), and its gap with SPM is not statistically significant. This is the first important finding of this thesis and confirms the result showed in previous studies about the effective prediction capabilities of CH2-L2, despite exploiting a restricted amount of topological information with respect to the other approaches here considered.

Table 4.5: Permutation test for the mean ranking in link prediction on small-size real networks.
For each pair of methods, a permutation test for the mean has been applied to the two vectors of pairwise link prediction rankings on the small-size real networks, using 10000 iterations. The table reports the pairwise p-values, adjusted for multiple hypothesis comparison by the Benjamini–Hochberg correction. The p-values lower than the significance level of 0.05 are highlighted in bold.

	CH2-L2	**SBM**	**FBM**	**RA-L3**	**SBM DC N**	**SBM DC**	**SBM N**
SPM	0.5973	**0.0003**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**
CH2-L2		0.5973	**0.0006**	**0.0044**	**0.0002**	**0.0002**	**0.0002**
SBM			10,000	0.1070	**0.0060**	**0.0002**	**0.0002**
FBM				0.3569	**0.0057**	**0.0013**	**0.0002**
RA-L3					**0.0018**	**0.0002**	**0.0002**
SBM DC N						**0.0002**	**0.0002**
SBM DC							**0.0002**

Table 4.6 reports the computational time required by the methods in order to perform the link prediction. The algorithms have been run in the workstations specified in detail in the Appendix B. The table highlights that SPM, CH2-L2, RA-L3 and FBM are quite fast as methods and respectively required from a few seconds up to one minute and a half for the link prediction on a network of around one thousand nodes. SBM and its variations, instead, required from around one day to almost 11 days for the same task. In particular, the nested hierarchical procedure is highly time consuming and it is absolutely not suitable for link prediction task. This table, together with Table 4.5, advocates the second crucial finding of this study: SBM displays huge computational time in comparison to the best state-of-the-art approaches, without any overall significant gain in link prediction performance even versus the best local method.

For a deeper investigation, the correct predictions shared by the different methods have been also analyzed (only CH2-L2, SPM and SBM have been considered in this analysis being respectively the best local approach, the best global approach and the best performing algorithm among the SBM family over the small-size real networks). For each small-size real network (and for each of 10 iterations), considering the entire set of links that have been correctly predicted by a pair of methods, the percentage of these links that are shared or not between the two methods is computed. The mean of the percentages taken over all the networks and iterations are reported as a Venn diagram for each pair of considered methods in Figure 4.1. It is possible to notice that on average the different approaches share around half of the correctly predicted links. The remaining part is distributed in an almost balanced way, with a few percentage

Table 4.6: Computational time on small-size real networks.
For each network, 10% of links have been randomly removed (10 iterations for SBM, SBM DC N, SBM DC and SBM N 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 table reports for each network the mean computational time over the random iterations and the mean time over the entire dataset. For each network the best method is highlighted in bold. The networks are sorted by increasing number of nodes *N*. Note that CH2-L2, SPM, RA-L3 and FBM are Matlab implementations, SBM is a C library, whereas the SBM variations are Python modules with core algorithms written in C++. The methods have been run in the workstations reported in Appendix B.

	CH2-L2	SPM	RA-L3	FBM	SBM DC	SBM	SBM N	SBM DC N
mouse_neural	0.4 s	**0.1 s**	0.3 s	0.2 s	4.1 s	2.0 s	46.9 s	43.7 s
karate	0.2 s	**0.1 s**	0.3 s	0.2 s	14.5 s	3.8 s	3.0 min	3.0 min
stmarks_foodweb	0.2 s	**0.1 s**	0.4 s	0.2 s	35.8 s	14.4 s	6.7 min	6.7 min
dolphins	0.2 s	**0.1 s**	0.4 s	0.2 s	51.3 s	10.8 s	10.8 min	10.4 min
ythan_foodweb	0.2 s	**0.1 s**	0.4 s	0.3 s	2.1 min	44.2 s	24.1 min	23.6 min
macaque_neural	0.2 s	**0.1 s**	0.4 s	0.3 s	2.2 min	1.9 min	20.1 min	20.2 min
polbooks	0.2 s	**0.1 s**	0.4 s	0.4 s	2.7 min	55.0 s	31.2 min	31.6 min
SEA_terrorist	0.2 s	**0.1 s**	0.4 s	0.4 s	3.0 min	1.5min	34.7 min	34.5 min
ACM2009_contacts	0.2 s	**0.1 s**	0.5 s	0.5 s	3.8 min	2.5 min	27.4 min	25.9 min
football	0.2 s	**0.1 s**	0.4 s	0.5 s	3.5 min	1.3 min	29.0 min	34.6 min
physicians_innovation	0.2 s	**0.1 s**	0.4 s	0.5 s	3.5 min	1.2 min	29.9 min	36.8 min
AQ_terrorist	0.2 s	**0.1 s**	0.4 s	0.5 s	3.9 min	1.2 min	30.7 min	45.8 min
manufacturing_email	0.3 s	**0.2 s**	0.7 s	0.7 s	11.4 min	10.2 min	31.0 min	1.1 h
jazz	0.3 s	**0.2 s**	0.7 s	1.0 s	18.3 min	15.2 min	46.3 min	2.0 h
residence_hall_friends	0.4 s	**0.2 s**	0.7 s	1.3 s	18.1 min	15.1 min	2.4 h	2.5 h
rhesus_brain	0.4 s	**0.3 s**	0.9 s	1.3 s	27.5 min	22.7 min	3.3 h	3.1 h
vanderwaals	0.5 s	**0.2 s**	0.7 s	1.6 s	19.7 min	11.1 min	3.6 h	3.3 h
haggle_contacts	0.5 s	**0.3 s**	0.9 s	0.5 s	31.4 min	11.4 min	4.2 h	3.8 h
worm_nervoussys	0.6 s	**0.4 s**	1.0 s	2.3 s	38.8 min	30.0 min	5.3 h	6.0 h
USAir	0.7 s	**0.4 s**	1.0 s	2.8 s	48.1 min	34.1 min	8.6 h	8.1 h
netsci	0.9 s	**0.5 s**	1.2 s	3.6 s	47.4 min	27.6 min	8.4 h	8.1 h
infectious_contacts	1.0 s	**0.6 s**	1.6 s	6.1 s	1.4 h	1.4 h	9.2 h	9.8 h
flightmap	**1.4 s**	13.1 s	18.7 s	6.9 s	9.6 h	18.9 h	4.1 d	4.9 d
email	7.2 s	**7.0 s**	12.2 s	1.5 min	18.4 h	15.7 h	7.7 d	9.2 d
polblog	**8.6 s**	10.3 s	257 s	1.5 min	21.3 h	1.0 d	9.4 d	10.5 d
mean	**1.0 s**	1.4 s	2.8 s	8.4 s	2.2 h	2.6 h	22.3 h	1.1 d

points more for the better performing method among the two. The couple having the highest overlap is SPM and CH2-L2 with 56.6% although based on very different theories, the one with the lowest overlap is CH2-L2 and SBM with 43.8%. From this last analysis emerges that a proper combination of even only two of these methods would significantly increase the number of correctly predicted links. Therefore, in order to exploit this link prediction heterogeneity across methods, we advance the idea to build hybrid methods that should potentially lead to higher performance.

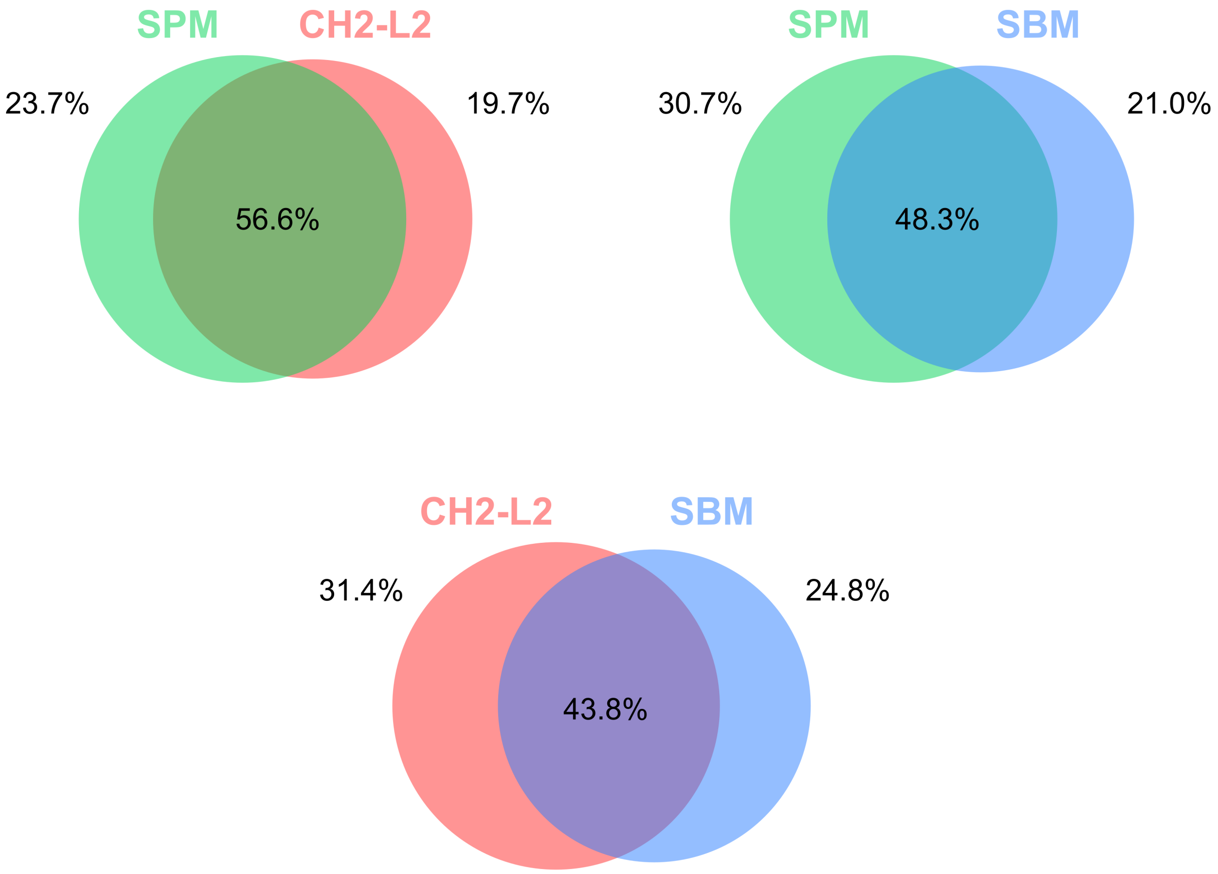

Figure 4.1: Pairwise Venn diagrams of correctly predicted links on small-size real networks.
For each pair of the methods SPM, CH2-L2 and SBM, the overlap of the correctly predicted links has been analyzed. For each small-size real network (and for each of 10 iterations), considering the entire set of links that have been correctly predicted by two methods, the percentage of these links that are shared or not is computed and reported in the corresponding Venn diagram, as average over all the networks and iterations.

Another important evaluation we propose is the analysis of link prediction methods on a set of 486 structural human brain connectomes. Because of the high computational time and the low performance, the SBM variations have not been considered. Link prediction problem in brain connectomes has gained much interest among the scientific community because several neuroscientific studies have demonstrated that certain forms of learning consist of synaptic modifications, while the number of neurons remains basically unaltered [6], [85], [86]. The results are depicted in Table 4.7, where it is immediate to verify that SPM is the best link prediction approach followed by CH2-L2 and then SBM. This is confirmed by both the mean precision and the ranking (left side of Table 4.7) and it is validated by the statistical significance between each pair of methods computed as described above (right side of Table 4.7). Here, SPM performs clearly better than the other methods and, again, it is interesting to notice that the local method CH2-L2 outperforms the global method SBM.

Table 4.7: Precision-ranking evaluation of link prediction on van den Heuvel brain networks.
This dataset is composed by 486 small-size structural brain networks acquired from healthy-controls. For each network, 10% of links have been randomly removed for 100 iterations and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The left side of the table reports the ranking of the methods by decreasing mean precision (over the random iterations and over the healthy-controls). The best method is highlighted in bold. For each pair of methods, a permutation test for the mean has been applied to the two vectors of pairwise link prediction rankings on all the networks, using 10000 iterations. The right side of the table reports the pairwise p-values, adjusted for multiple hypothesis comparison by the Benjamini–Hochberg correction. The p-values lower than the significance level of 0.05 are highlighted in bold.

	mean precision	**ranking**	**p-values**				
SPM	**0.48**	**1.0**		**CH2-L2**	**SBM**	**RA-L3**	**FBM**
CH2-L2	0.42	2.0	**SPM**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
SBM	0.36	3.0	**CH2-L2**		**0.0001**	**0.0001**	**0.0001**
RA-L3	0.31	4.0	**SBM**			**0.0001**	**0.0001**
FBM	0.15	5.0	**RA-L3**				**0.0001**

4.2 Evaluation on Synthetic Networks

It was surprising to discover that a local method such as CH2-L2 is comparable (because its performance is not always significantly different) to SPM - that is the best global method. Possibly, the previous results could have been only part of the picture and biased by the selection of small-size real networks available in the literature, therefore we decided to extend the link prediction evaluation considering also artificial networks. A more detailed explanation on the procedure of the artificial networks generation can be found in Section 3.1.

4.2.1 Framework of RGG Networks

A Random Geometric Graph (RGG) is an artificial model to generate networks in the Euclidean space. The RGG model has been used to generate networks with parameters N = [100, 500, 1000] (networks size) and, respectively, r = [0.25, 0.15, 0.10] (threshold neighborhood radius). The values chosen for N are intended to cover the range of networks observed in the dataset of small-size real networks (the first 25 entries of Table 3.1); the values of r are chosen in order to have a connected network with an average degree comparable to the mean average degree displayed by the real networks considered.

Table 4.8 collects the results of precision averaged over the link removal iterations for each network (10 iterations for SBM due to the high computational time, 100 iterations for the other methods). Additionally, it shows the mean precision and the mean ranking over the different parameter settings (the last two columns of Table 4.8). This is a further suggestion that SPM is the best state-of-the-art method for link prediction followed by CH2-L2. It is interesting to verify the critical drop in performance of the SBM algorithm which cannot compete with the others methodologies.

Table 4.8: Precision-ranking evaluation of link prediction on synthetic RGG networks.
For each combination of parameters, 100 networks have been generated. For each network, 10% of links have been randomly removed (10 iterations for SBM 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. For each network the table reports the mean precision over the random iterations for the different configurations. For each setting, the best method is highlighted in bold. The mean ranking of the methods over all the networks represents the final evaluation for a proper comparison of the performance. The methods are ordered by decreasing mean precision and ranking.

	N=100	**N=500**	**N=1000**	**mean precision**	**mean ranking**
SPM	**0.41**	**0.58**	**0.72**	**0.57**	**1.0**
CH2-L2	0.35	0.43	0.55	0.44	2.0
FBM	0.30	0.37	0.50	0.39	3.0
RA-L3	0.25	0.32	0.43	0.33	4.0
SBM	0.16	0.18	0.26	0.20	5.0

4.2.2 Framework of nPSO Networks

The nonuniform Popularity-Similarity-Optimization (nPSO) model is a network model recently proposed in [43], [87] as a development of the simpler PSO model [44]. The PSO model is a generative network model which describes how random geometric graphs grow in the hyperbolic space. PSO networks evolve in time optimizing a trade-off between node popularity, represented by the radial coordinate, and similarity, symbolized by the angular coordinate distance. In the PSO model many common structural and dynamical characteristics of real networks were taken into account, but an adequate community structure was lacking. This has been solved in the nPSO model allowing to set a nonuniform distribution of the nodes over the hyperbolic distance, being the connection probability a decreasing function of the hyperbolic distance.

It has already been shown that, on the PSO model, the CH1-L2 method outperforms all the others methods (this will be conserved also for the CH2-L2 index) and that SBM is the method performing worst [36]. Since SBM might be sensitive to the organization of the network in blocks and the PSO artificial networks do not have communities, we repeated the same simulations using the nPSO model.

Here, the nPSO model has been used to generate networks with parameters $\gamma = 3$ (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), $N = [100, 500, 1000]$ (network size) and 8 communities. The values chosen for N and T are intended to cover the range of network size and clustering coefficient C observed in the dataset of small-size real networks (the first 25 entries of Table 3.1). Since the average γ estimated on the dataset of small-size real networks is higher than the typical range $2 < \gamma < 3$ [48], $\gamma = 3$ has been selected.

Figure 4.2 reports for each parameter combination the average link prediction precision and the respective standard error computed over 100 networks for SPM, CH2-L2 and FBM and over 10 networks for the SBM family, due to the high computational time. Note that only the best method among the SBM family (i.e. SBM) is reported in Figure 4.2 for clarity reasons and a comparison of the SBM family methods (i.e. SBM, SBM N, SBM DC and SBM DC N) is shown in Figure 4.3 for the sake of completeness. Also, in Figure 4.4 it could be noticed that between the two local link prediction methods considered, CH2-L2 achieves always better performance than RA-L3 when the difference is significant and, for this reason, the latter it has not been reported in Figure 4.2.

The first fact to highlight is that the methods generally obtain different performance for low temperature (high clustering), similar results for medium temperature (medium clustering) and almost the same performance for high temperature (low clustering). Furthermore, they all exhibit a decreasing behavior going from low to high temperature. This is expected since,

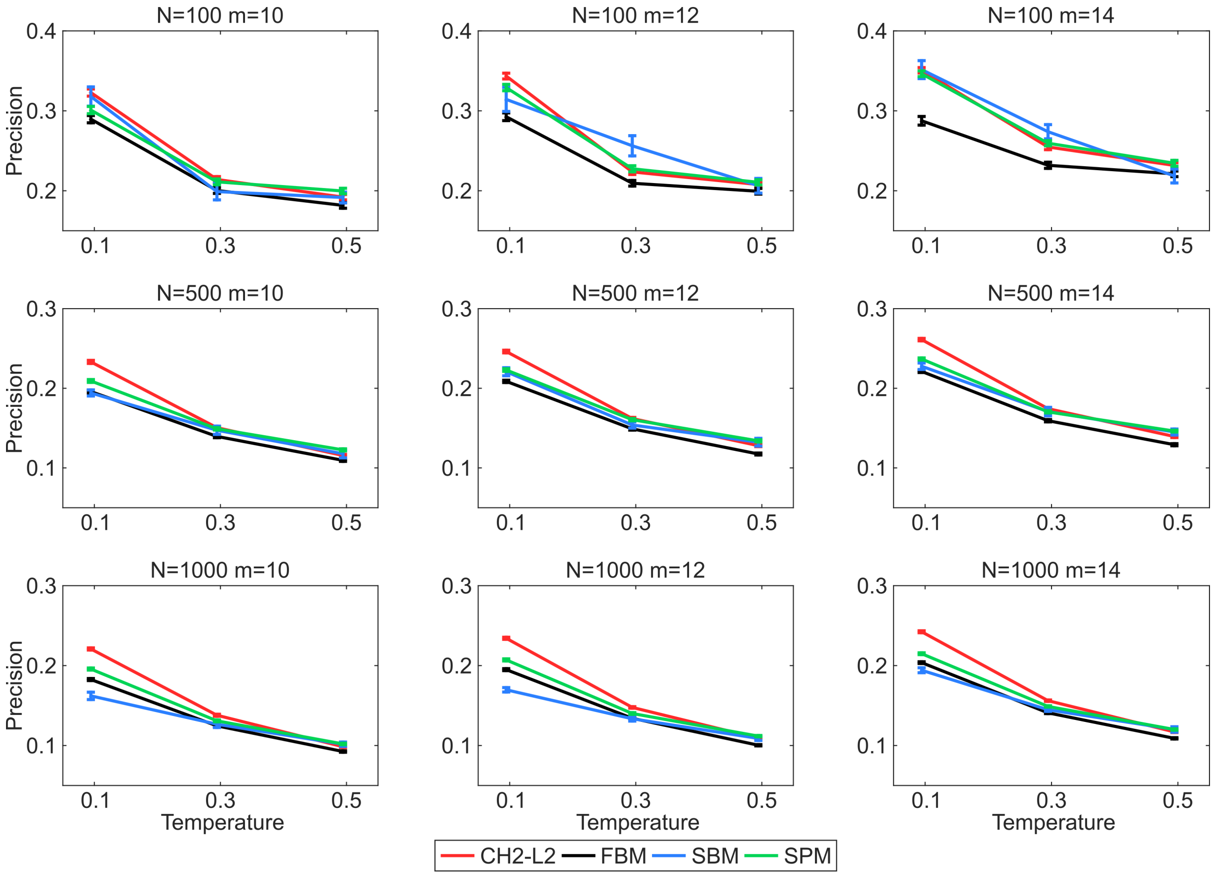

Figure 4.2: Precision evaluation of link prediction on nPSO networks with 8 communities.

For each combination of parameters, 100 networks have been generated. For each network, 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Note that only 10 networks have been considered in case of SBM, Nested SBM (SBM N), Degree Corrected and Nested SBM (SBM DC N) and Degree Corrected SBM (SBM DC) due to the high computational time. Only the best method among the SBM family, i.e. SBM, is reported here for clarity reasons.

according to the PSO model theory (and inherited by the nPSO model), for increasing temperature the network tends to assume a more random and degenerate topology, which makes the link predictability harder.

From the link prediction performance in Figure 4.2 it is possible to see that, while for networks of size $N = 100$ the ranking of the methods is variable and the precisions are on average comparable for CH2-L2, SPM and SBM; for networks of increasing size CH2-L2, which is the only

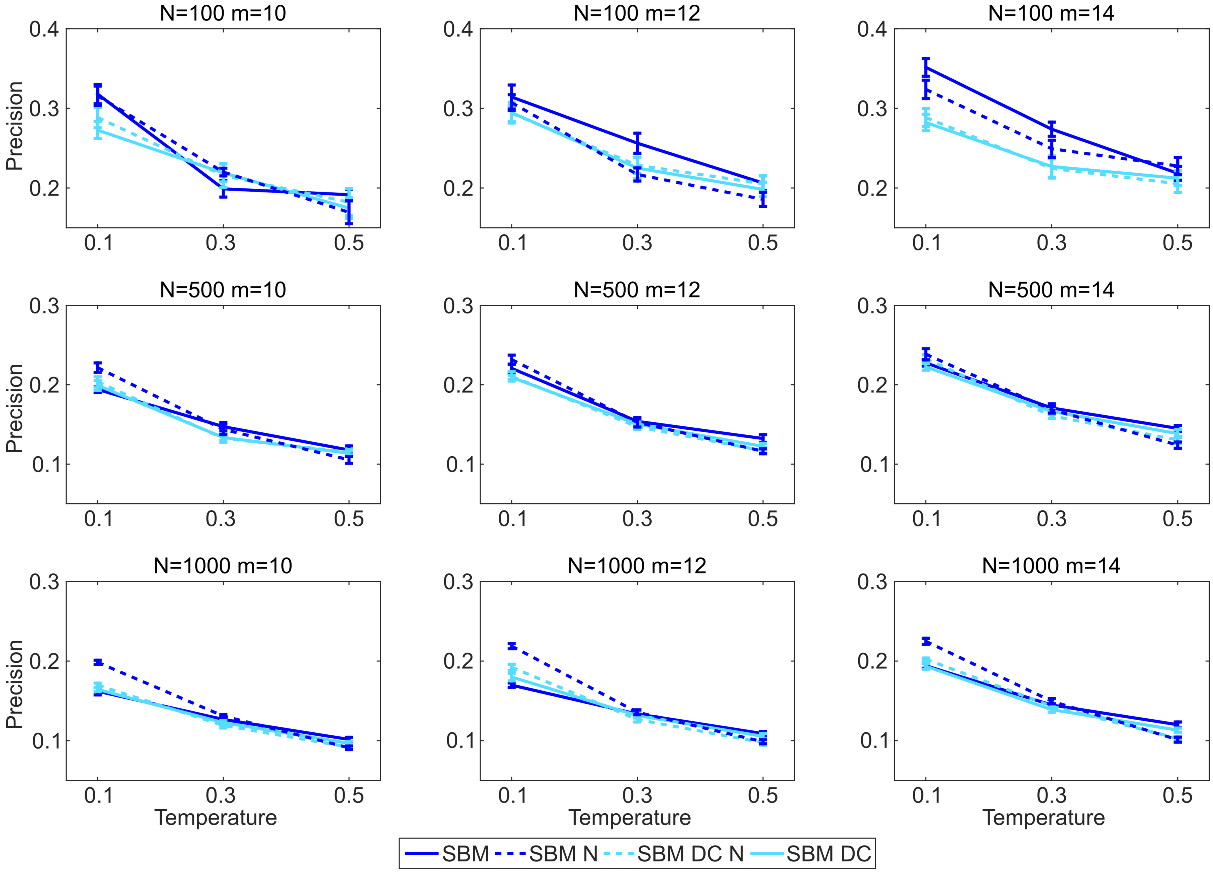

Figure 4.3: Precision evaluation of link prediction of SBM variations on nPSO networks with 8 communities.

For each combination of parameters, 100 networks have been generated. For each network, 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Note that only 10 networks have been considered due to the high computational time.

local approach shown in the figure, tends to significantly surpass the global methods for low temperature $T = 0.1$ followed by SPM. For low temperature, SBM is higher than or equal to FBM for networks of size $N = 100$ and 500, whereas it is the worst performing for networks of bigger size. For higher temperatures, even though the difference in performance becomes thinner and often vanishes (within overlapping standard errors), the same trend is preserved or at least not significantly inverted. In Table 4.9 the permutation tests of the pairwise ranking for all the network configurations for each pair of methods is reported, together with the average ranking: overall, CH2-L2 and SPM outperform the other methods and their relative difference

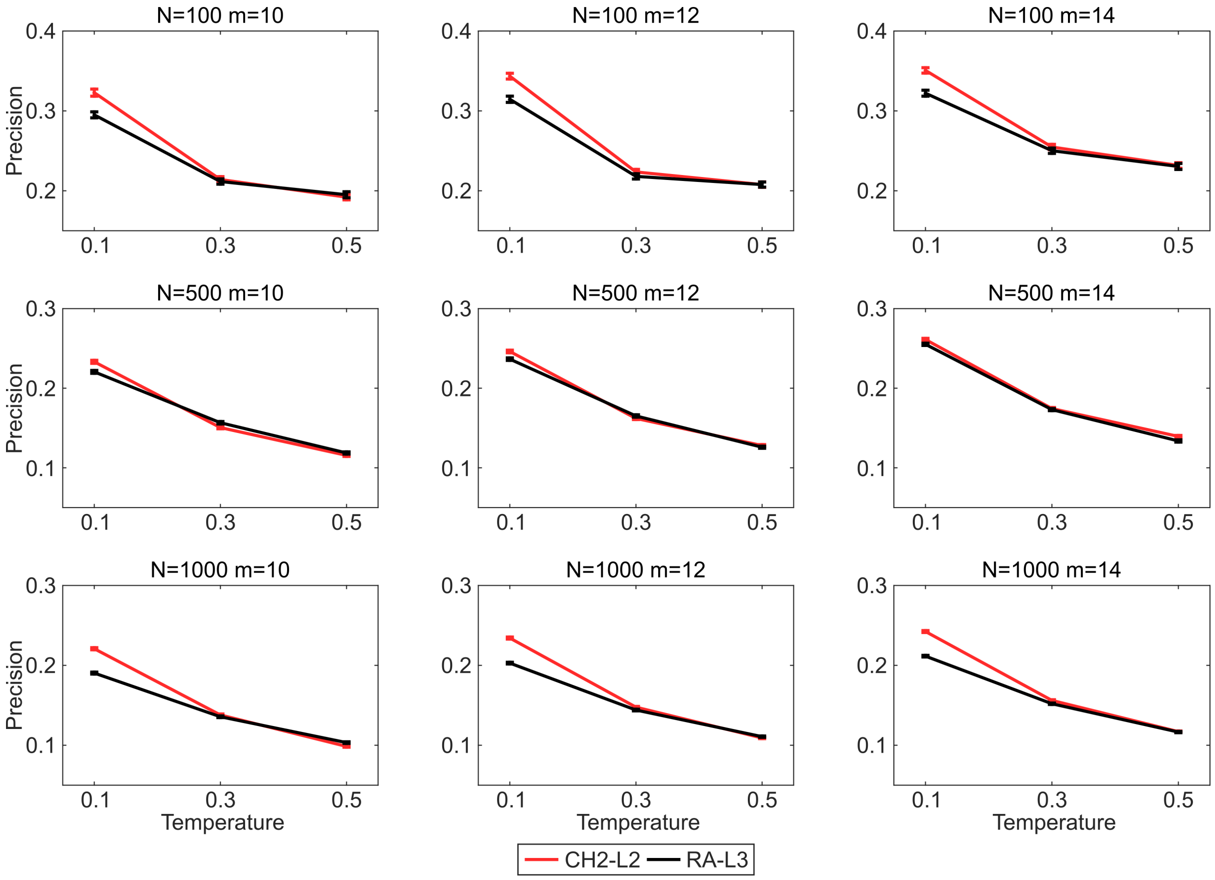

Figure 4.4: Precision evaluation of link prediction of CH2-L2 and RA-L3 on nPSO networks with 8 communities.

For each combination of parameters, 100 networks have been generated. For each network, 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations.

is not statistically significant.

Even with the introduction of the communities (differently from the standard PSO model), CH2-L2 obtains the best mean ranking with 2.1 with respect to the global approaches (the best global method is SPM with 2.7) and in particular to SBM with 4.3, which should have performed better in this scenario. This suggests that the hyperbolicity of the networks might be the main cause of this result and this point will be better analyzed in Chapter 5. As additional comment, we let notice that the nPSO model captured the quite comparable performance of the methods on the smallest networks as observed on real topologies, offering a more realistic framework with respect to the original PSO model. Finally, for the nPSO model the gain of performance of

Table 4.9: Permutation test and mean ranking of link prediction on nPSO networks with 8 communities.
For each pair of methods, a permutation test for the mean has been applied to the two vectors of pairwise link prediction rankings on the nPSO networks with 8 communities, using 10000 iterations. The table reports the pairwise p-values, adjusted for multiple hypothesis comparison by the Benjamini–Hochberg correction. The p-values lower than the significance level of 0.05 are highlighted in bold. The methods are ordered by overall average ranking, which is shown as the last column.

	CH2-L2	SPM	RA-L3	SBM	SBM N	SBM DC	SBM DC N	FBM	ranking
CH2-L2		0.3099	**0.0011**	**0.0075**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	2.1
SPM			0.3099	**0.0002**	0.1278	**0.0002**	**0.0002**	**0.0002**	2.7
RA-L3				**0.0021**	**0.0016**	**0.0002**	**0.0002**	**0.0002**	3.0
SBM					0.3099	**0.0002**	**0.0002**	**0.0006**	4.3
SBM N						**0.0336**	**0.0088**	**0.0075**	4.9
SBM DC							1	1	6.3
SBM DC N								1	6.3
FBM									6.4

CH2-L2 and SPM is evident for networks of size 500 and 1000 nodes. This result suggests that, if the nPSO is well-designed to be realistic, CH2-L2 and SPM should outperform also the other methods in link prediction on large size real networks with hyperbolic geometry.

The ranking of the methods is not exactly the same as in the small-size real networks dataset, which - as a speculation - might suggest that, although some structural properties are reproduced by the model, it does not cover the whole variability present in the real network topologies. Conversely, it might be true also the opposite, that the selection of real complex networks we used is biased towards network topologies that favor global models, while the artificial networks not. However, the two separate evaluations are still in agreement on one point: the two methods that recent studies have demonstrated to be among the best performing for the global and the local approaches, respectively SPM and CH2-L2, obtained a higher overall performance with respect to the other methods especially the ones based on the stochastic block model theory.

4.2.3 Framework of WS Networks

The Watts-Strogatz (WS) model [48] is a network generative model which introduced the concept of small-world networks with strong clustering coefficient and with small characteristic path length, as usually exhibited by most of real world networks. By properly tuning the parameters of the model it is possible to obtain random-like networks with non-scale-free node distribution. We performed the simulation using the Watts-Strogatz model considering parameters $N = [100,$

500, 1000] (network size), m = [10, 12, 14] (half of average degree) and β = [0.001, 0.01, 0.1] (rewiring probability). The values chosen for the parameters N and m are the same used for the synthetic networks generated using the RGG and the nPSO model. The values chosen for β are intended to produce networks with different properties mainly in terms of clustering coefficient and characteristic path length. Figure 4.5 reports the link prediction results from which it is evident that SPM largely outperforms the other link prediction methods followed by RA-L3 and CH2-L2. The two local methods perform very similarly and a comparison between them is

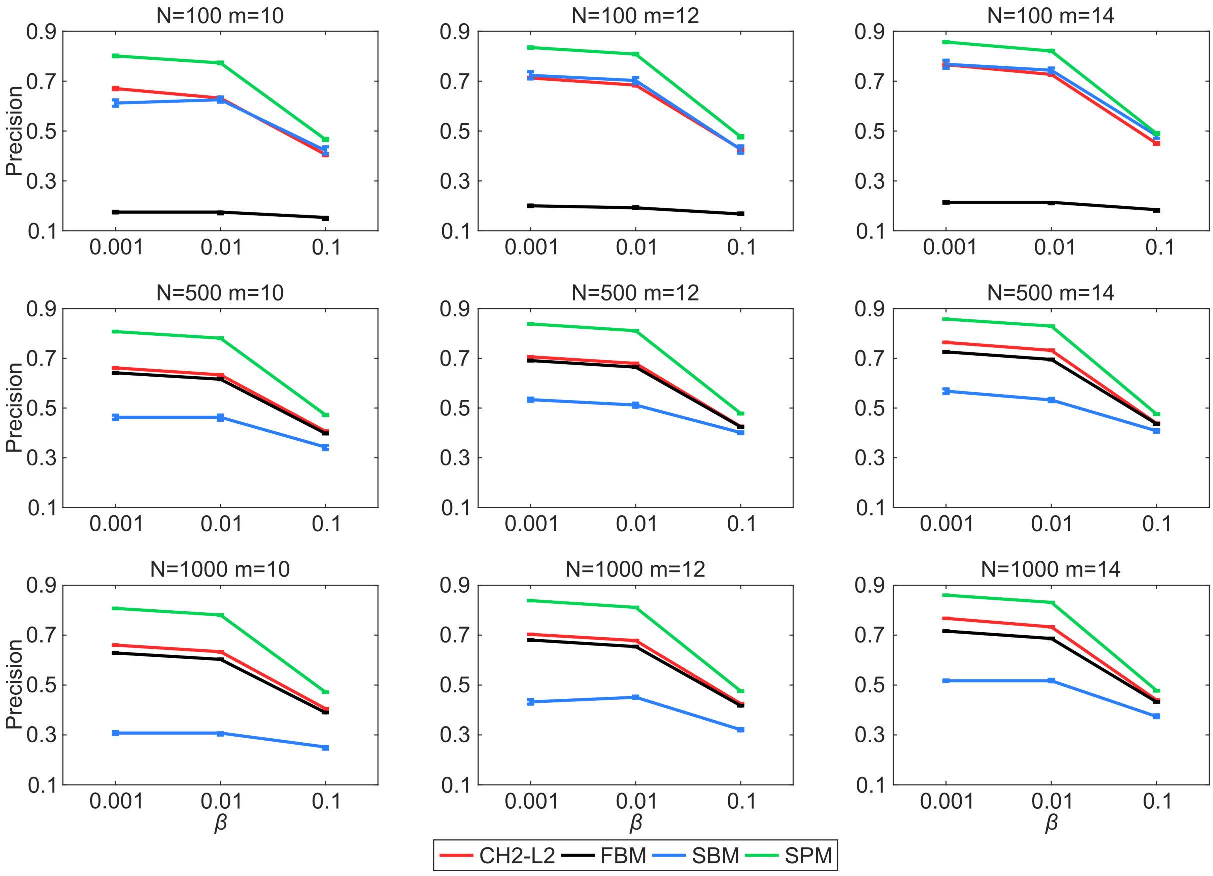

Figure 4.5: Precision evaluation of link prediction on Watts-Strogatz networks with 8 communities.

For each combination of parameters, 100 networks have been generated. For each network, 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Note that for SBM only 10 networks have been considered due to the high computational time.

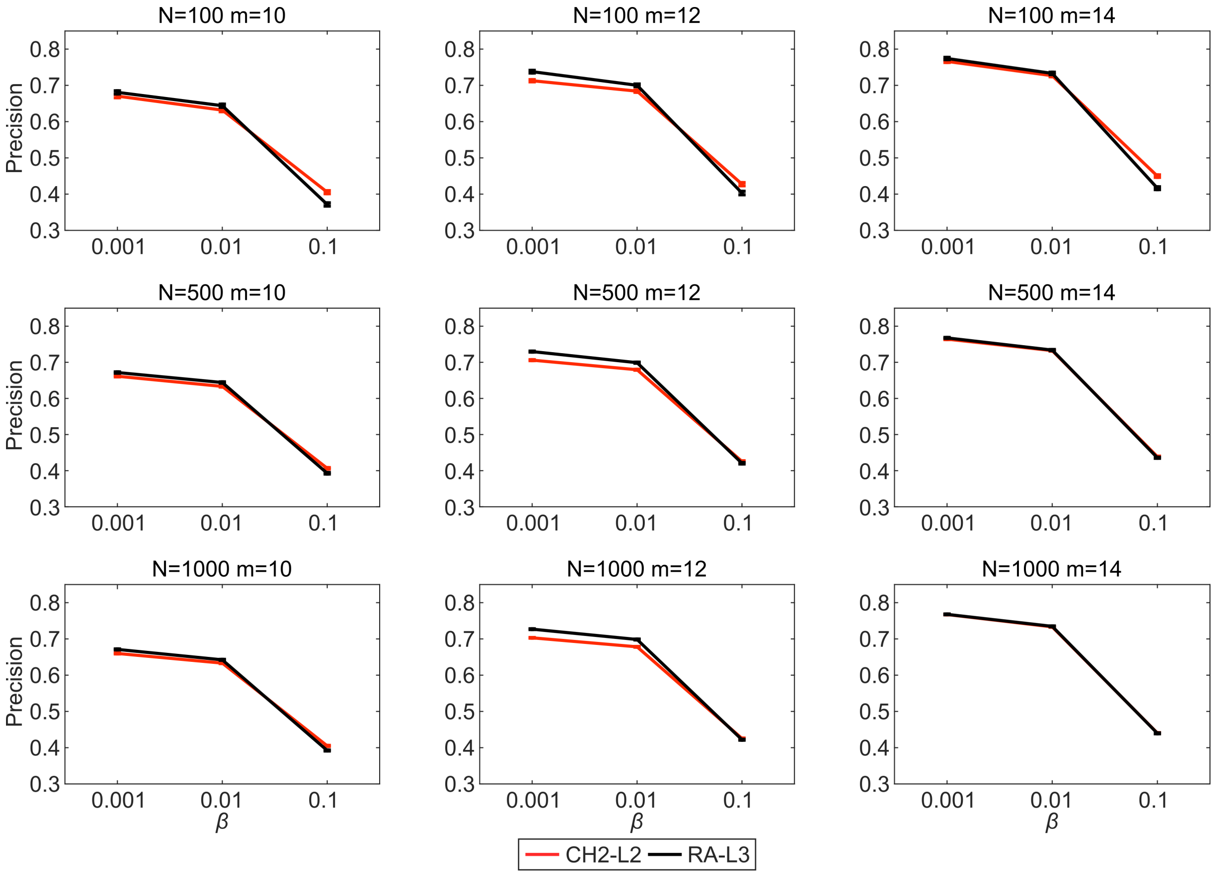

Figure 4.6: Precision evaluation of link prediction of CH2-L2 and RA-L3 on Watts-Strogatz networks with 8 communities.
For each combination of parameters, 100 networks have been generated. For each network, 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations for the two local methods CH2-L2 and RA-L3.

reported in Figure 4.6, where we can notice that RA-L3 is higher than or equal to CH2-L2 for higher clustering (β equal to 0.001 and 0.01), while CH2-L2 is higher than or equal to RA-L3 for lower clustering (β equal to 0.1). Since the difference is very low and tends to vanish for larger networks, only CH2-L2 is reported in Figure 4.5.

Also in this framework SPM and CH2-L2 obtain overall a better and more robust performance with respect to SBM, which has a particular drop in precision for $N = 500 - 1000$, and FBM, whose discrepancy with respect to the other methods is huge for $N = 100$. To notice that the SBM-FBM trend is similar to the one obtained for the link prediction on the nPSO model, where SBM is better than FBM for $N = 100 - 500$ and their ranking is inverted for $N = 1000$. Table

4.10 certifies what we have just asserted providing the average ranking and the permutation tests for the mean between the pairwise ranking of each pair of methods, whose difference is always significant.

Table 4.10: Permutation test and mean ranking of link prediction on WS networks. For each pair of methods, a permutation test for the mean has been applied to the two vectors of pairwise link prediction rankings on the WS networks, using 10000 iterations. The table reports the pairwise p-values, adjusted for multiple hypothesis comparison by the Benjamini–Hochberg correction. The p-values lower than the significance level of 0.05 are highlighted in bold. The methods are ordered by overall average ranking, which is shown as last column.

	SPM	**RA-L3**	**CH2-L2**	**FBM**	**SBM**	**ranking**
SPM		**0.0001**	**0.0001**	**0.0001**	**0.0001**	1.0
RA-L3			**0.0288**	**0.0001**	**0.0001**	2.6
CH2-L2				**0.0001**	**0.0001**	2.9
FBM					**0.0288**	4.2
SBM						4.3

4.2.4 Framework of LFR Networks

The Lancichinetti-Fortunato-Radicchi (LFR) model [49] is a synthetic networks generator, which addresses the heterogeneity of the node degrees and of the community sizes typically displayed in real world networks modelling them as two different (truncated) power-law distributions. The LFR benchmark is a special version of the degree-corrected stochastic block model [31], with the degree and the block size distributed according to truncated power laws [50]. In this evaluation framework it is natural to expect that the SBM family methods should perform better than the others since the inference is made on networks generated according to the same theory of SBM itself.

We performed the simulation using the LFR model considering parameters N = [100, 500, 1000] (network size), k = [20, 24, 28] (average degree), μ = [0.3, 0.5, 0.7] (mixing parameter) and $minc$ = [N/10, N/20] (minimum of the community size), fixing $maxk$ = 3*k (maximum degree of a node) and $maxc$ = 4*$minc$ (maximum of the community size) and trying to satisfy a desired clustering coefficient C = [0.7, 0.5, 0.3]. The values chosen for the parameters are the same used for the synthetic networks generated using the previous models, when available. The values chosen for $minc$, $maxc$ and $maxk$ are reasonable and intended to produce networks with characteristics similar to real networks.

Figure 4.7 represents the link prediction precision on the LFR networks with larger community

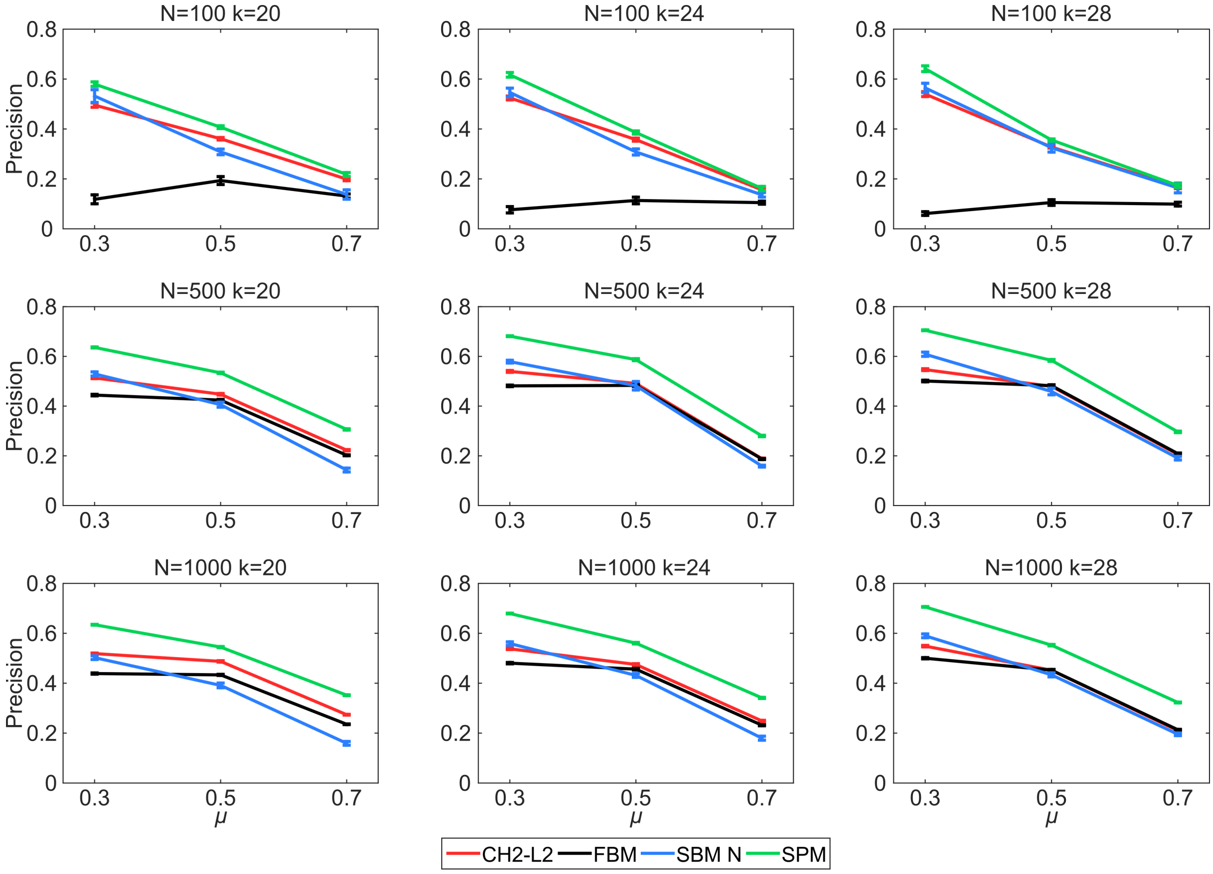

Figure 4.7: Precision evaluation of link prediction on LFR networks with minimum community size equal to $N/10$.
For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Note that only 10 networks have been considered in case of SBM, Nested SBM (SBM N), Degree Corrected and Nested SBM (SBM DC N) and Degree Corrected SBM (SBM DC) due to the high computational time. Only the best method among the SBM family, i.e. SBM N, is reported here for clarity reasons.

sizes ($minc = N/10$). Figure 4.7 includes the results for the following methods: the best global method, i.e. SPM; the best method among the SBM family, i.e. SBM N; the best local method, i.e. CH2-L2, and FBM. The comparison of the SBM family methods is shown in Figure 4.8 and outlines that SBM N is almost constant by increasing the network size, while the simple SBM experiences a critical drop for larger networks of 500 and 1000 nodes. Moreover, from the comparison of the two local methods reported in Figure 4.9 it can be easily verified that

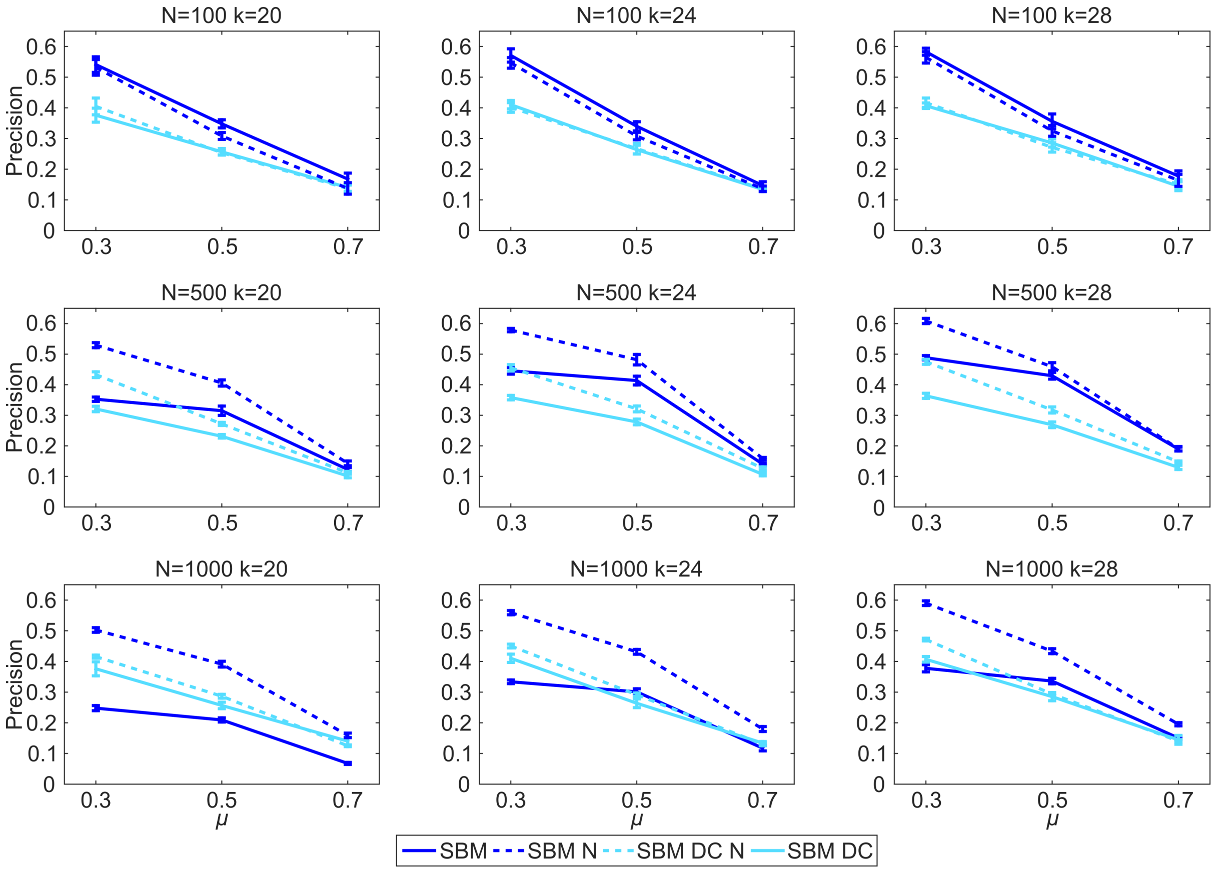

Figure 4.8: Precision evaluation of link prediction of SBM variations on LFR networks with minimum community size equal to $N/10$.
For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Note that only 10 networks have been considered for all the methods due to their high computational time.

CH2-L2 is the best local method.

From Figure 4.7 it appears clearly that SPM is the best performing link predictor in every scenario. As already pointed out on the other synthetic models, the FBM algorithm has very low performance for small-size networks and it increases its precision on larger networks. Similarly as before, all the methods exhibit a decreasing behavior going from low to high values of the mixing parameter μ, which is somehow related to the temperature shown in previous figures (inversely related to the clustering); for increasing temperature (i.e. for increasing values of μ) the networks tend to assume a more random and degenerate topology, which makes the link

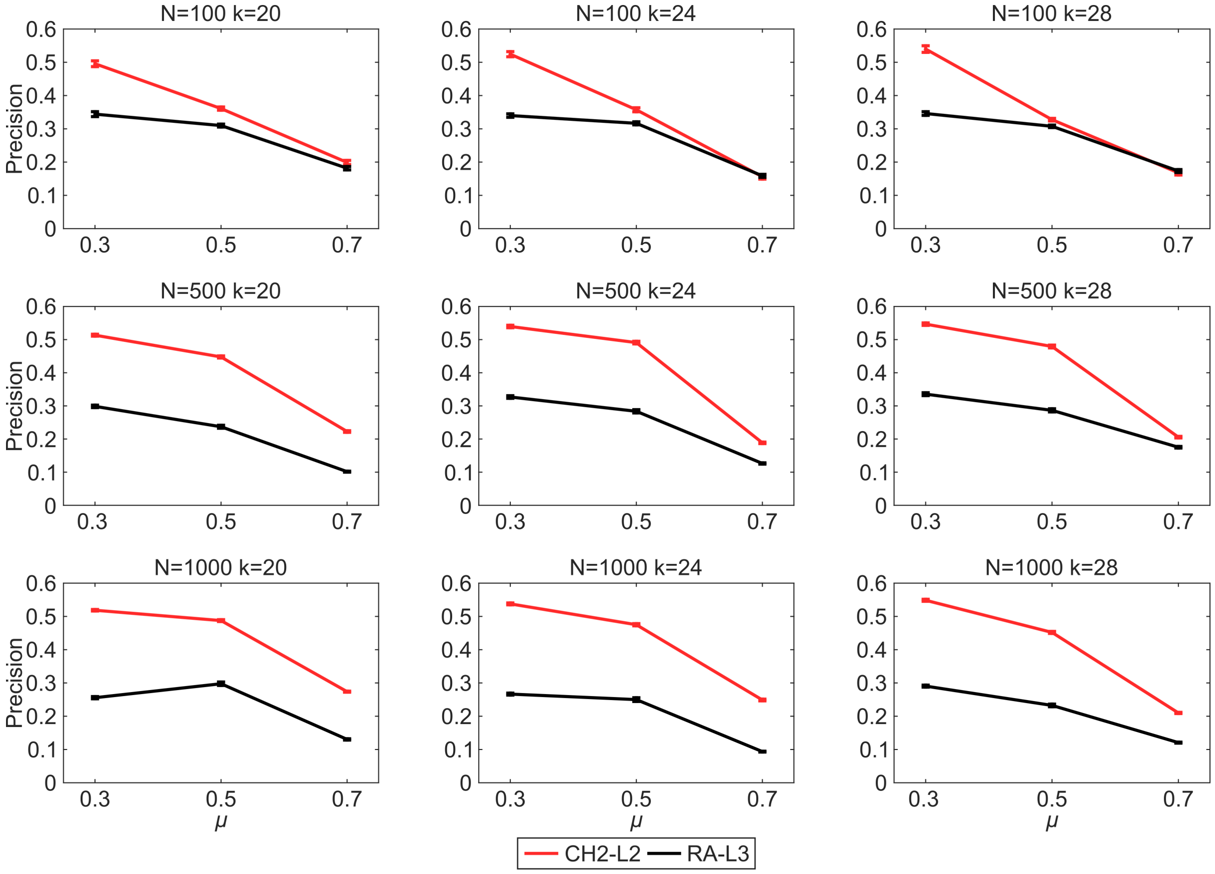

Figure 4.9: Precision evaluation of link prediction of CH2-L2 and RA-L3 on LFR networks with minimum community size equal to $N/10$.
For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations.

predictability harder.
Again, in Figure 4.7 can be verified that for low values of the mixing parameter μ the second best method is SBM N, which is slightly better than CH2-L2, while for increasing values of μ its performance is lower and the second best method tends to be CH2-L2.

In order to further corroborate the results and the analysis we conducted the same evaluation on LFR networks changing the minimum of the community sizes: this could be source of misleading results because the connection probability between two nodes assigned by SBM algorithms solely depends on the partitioning groups to which the nodes belong.
Similar trends are maintained for LFR networks with smaller community sizes (for example

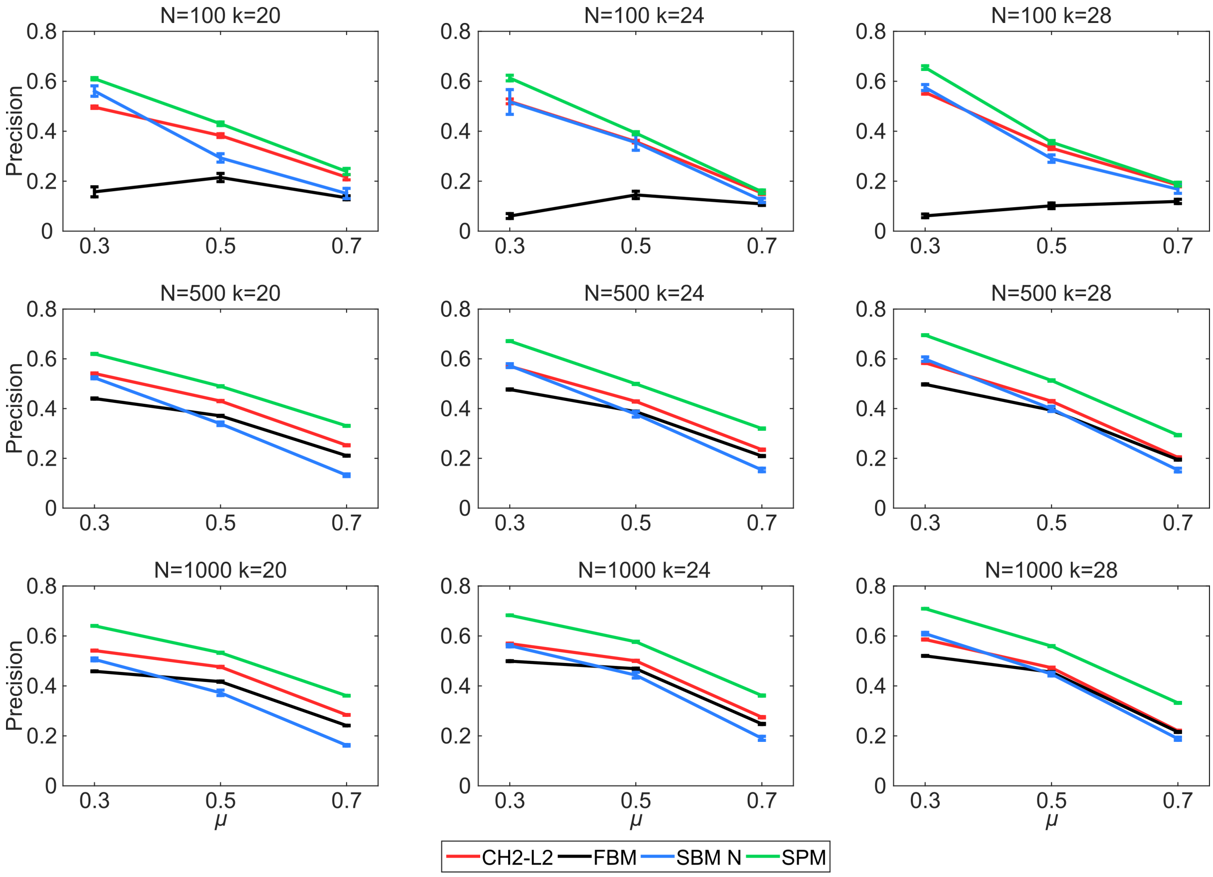

Figure 4.10: Precision evaluation of link prediction on LFR networks with minimum community size equal to $N/20$.

For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Note that only 10 networks have been considered in case of SBM, Nested SBM (SBM N), Degree Corrected and Nested SBM (SBM DC N) and Degree Corrected SBM (SBM DC) due to the high computational time. Only the best method among the SBM family, i.e. SBM N, is reported here for clarity reasons.

considering $minc = N/20$) as shown in Figure 4.10. The best method among the SBM family remains SBM N and the best local method is CH2-L2, as can be confirmed by looking respectively to Figure 4.11 and Figure 4.12. Generally, SBM N achieves slightly lower precision scores than in the previous framework with $minc = N/10$: this leads CH2-L2 to be more clearly the second best method (after SPM) in almost every configuration. The comparison of these figures with the previous set of figures (compare respectively Figures 4.7 and 4.10, 4.8 and 4.11, 4.9 and

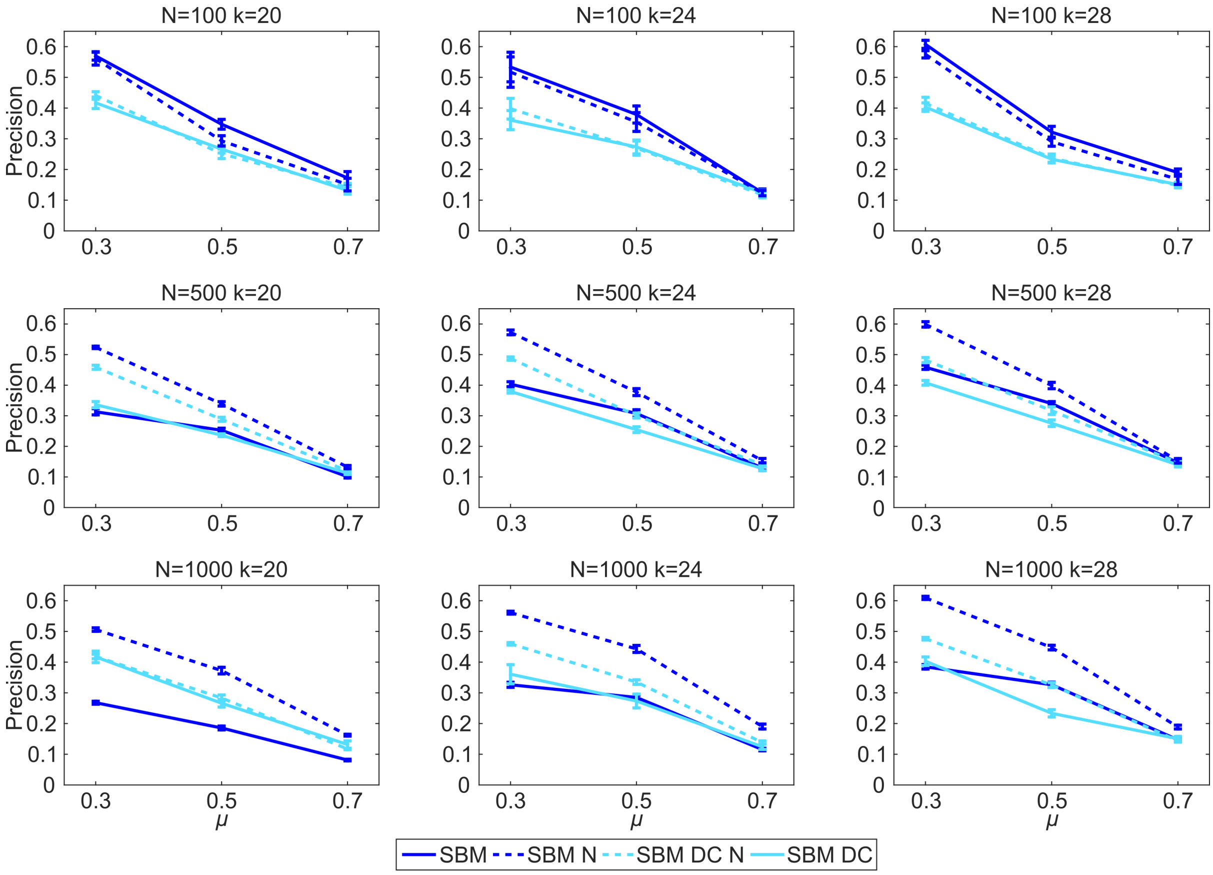

Figure 4.11: Precision evaluation of link prediction of SBM variations on LFR networks with minimum community size equal to $N/20$.
For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations. Note that only 10 networks have been considered for all the methods due to their high computational time.

4.12) suggests that the size of the communities does not heavily influence the precision of the methods belonging to the SBM family. It was important to clarify this last aspect, because the results could have been biased by the network community size, since

Finally, the overall mean ranking averaged over all the possible parameter combinations (also over the two values assigned to *minc*) is reported in Table 4.11, together with a permutation test for the mean applied to the pairwise link prediction rankings for each pair of methods. Clearly, SPM is by far the best method with 1.1 of ranking followed by CH2-L2 with 2.6. SBM N is only third with 3.7 of ranking and its difference from the first two methods is statistically significant.

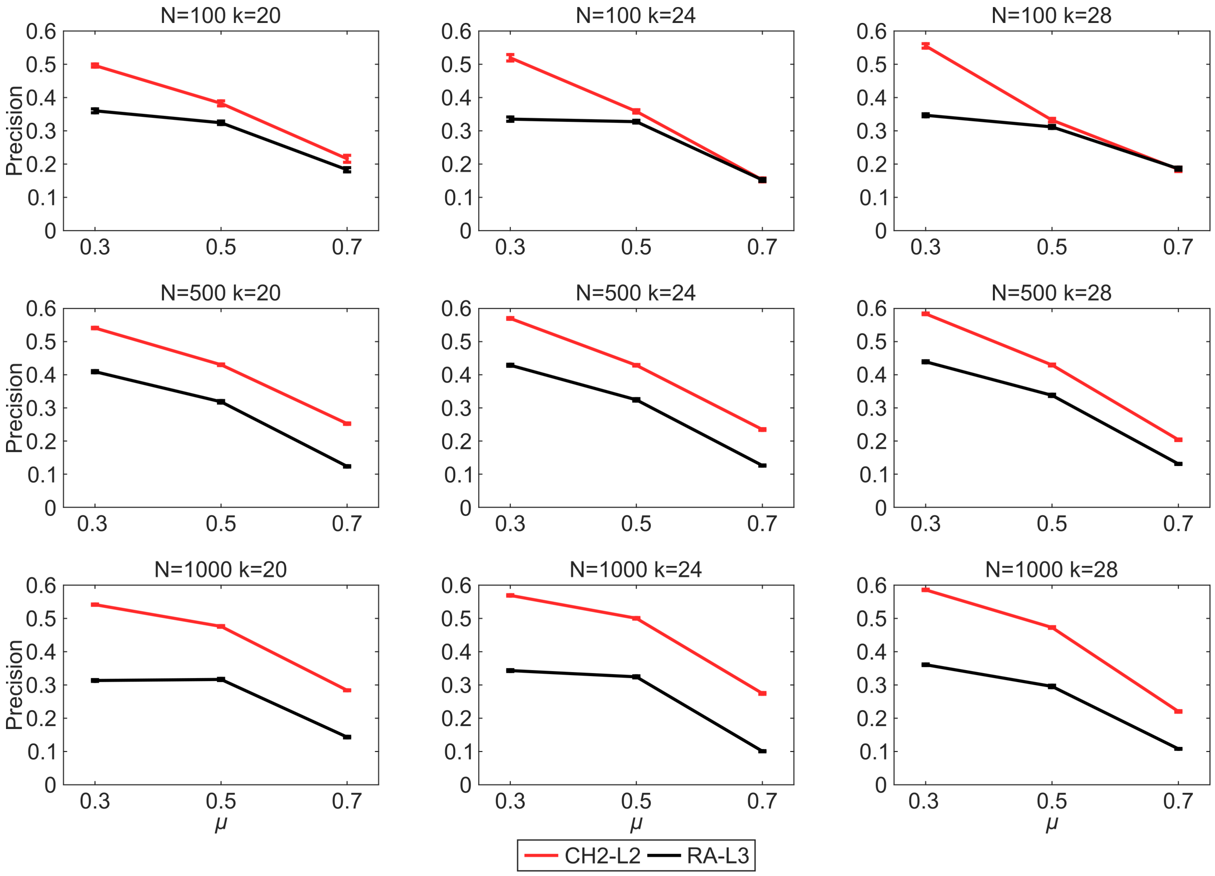

Figure 4.12: Precision evaluation of link prediction of CH2-L2 and RA-L3 on LFR networks with minimum community size equal to $N/20$.
For each combination of parameters, 100 networks have been generated. For each network 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The plots report, for each parameter combination, the mean precision and standard error over the random iterations.

Then FBM comes, although it highly suffers for small-size networks of 100 nodes, and the simple SBM is only the fifth method with 5.1 of ranking.

This is a noteworthy result which triggered our attention and led us to the interesting conclusion that SBM exhibits a clear problem of inference even on a generative model based on the same exact theory of the link predictor itself. Indeed, it has been shown that the LFR benchmark for synthetic networks is a special version of the degree-corrected stochastic block model with the degree and the block size distributed according to truncated power laws [50], [31], [30]. This was actually the original motivation to conduct this test, in order to verify the inference power of the SBM algorithm.

Table 4.11: Permutation test and mean ranking of link prediction on LFR networks. For each pair of methods, a permutation test for the mean has been applied to the two vectors of pairwise link prediction rankings on the LFR networks (aggregated results on LFR networks with minimum of the community size $minc{=}N/10$ and $N/20$), using 10000 iterations. The table reports the pairwise p-values, adjusted for multiple hypothesis comparison by the Benjamini–Hochberg correction. The p-values lower than the significance level of 0.05 are highlighted in bold. The methods are ordered by overall average ranking, which is shown as last column.

	SPM	CH2-L2	SBM N	FBM	SBM	SBM DC N	RA-L3	SBM DC	ranking
SPM		**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	1.1
CH2-L2			**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	2.6
SBM N				0.1822	**0.0010**	**0.0002**	**0.0002**	**0.0002**	3.7
FBM					**0.0010**	**0.0113**	**0.0010**	**0.0003**	4.9
SBM						0.0899	**0.0006**	**0.0003**	5.1
SBM DC N							0.0899	**0.0002**	5.9
RA-L3								0.3443	6.1
SBM DC									6.7

These considerations will be better developed throughout the next Chapter where all these results will be discussed in more depth.

5 | Discussion

Link prediction studies are not always convincing in the selection of the approaches to adopt as a reference for comparison. Global methods are believed to be the best performing and SBM, in spite of its remarkably high computational time, is often considered among the best state-of-the-art methods to use as a baseline. However, we did not find in the scientific literature well-grounded proofs of significant outperformance with respect to other state-of-the-art methods. Consequently, we decided to conduct an accurate study that compares SBM with Structural Perturbation Method (SPM) and CH2-L2, the two methods that in many studies were recently pointed out as the best respectively for global and local link prediction. In addition, for completeness, some variants of SBM have been considered: namely FBM, representing a faster variant of SBM, SBM DC, SBM N and SBM DC N, since they represent some modifications to the underlying SBM theory trying to compensate for the variations in node degree and for the grouping of nodes in small but well-clustered communities.

In contrast to the malpractice of testing the methods in a reduced benchmark of small-size networks, this study is characterized by an extensive analysis evaluating the methods on many different frameworks considering both real and artificial networks.

From the wide investigation several key messages emerged, which will be now summarized. First, SPM proved itself to be the best global method, significantly outperforming SBM in both real and artificial networks. Second, SBM, commonly adopted as a state-of-the-art baseline for comparison, displayed a huge computational time with respect to the other approaches, without an overall gain in prediction performance even versus the best local method CH2-L2. Third, mean precision resulted to be an inappropriate metric of overall performance. In fact, the mean is a central measure affected by the presence of peculiar networks that strongly favor the prediction only for some methods, whereas the precision-ranking provides a more robust and unbiased overview. Fourth, the evaluation on multiple frameworks highlighted that the adoption of a single benchmark with only small-size networks, although the number of networks tested is large, can easily bring to misleading conclusions showing only part of the truth. Last but not least, CH2-L2, the best local approach, has been found to be comparable to the best global approach, SPM, in some evaluation frameworks.

Actually, surprised by the impressive results obtained by a simple yet powerful local, parameter-free and model-based deterministic rule such as CH2-L2, we decided to investigate further its capability of link prediction on real networks. It has been shown that CH2-L2 can outperform the global models especially for networks of increasing size, thus we were encouraged to look further using also large-size networks and a different evaluation framework. Therefore, we conducted an investigation that considers the link-growth evolution of a real network over time. The networks represent six Autonomous Systems (AS) Internet topologies extracted from the data collected by the Archipelago active measurement infrastructure (ARK) developed by CAIDA [76], from September 2009 to December 2010 at time steps of 3 months. Several statistics of the AS snapshots are shown in Table 3.1, they are large-size networks with a number of nodes going from 24000 to almost 30000. It has already been shown in Tables 4.6, 4.1, 4.2 and 4.3 the computational demand of SBM and its variations which clearly makes them unusable for bigger networks. Anyway, it has been appreciated that the best methods are also the fastest ones and can be applied to bigger networks; for this reason only SPM and CH2-L2 have been employed for the link prediction evaluation on this dataset, but it is clear that this comparison is enough for the prefigured purpose.

Since in this case the information about the links that will appear is available, the evaluation framework differs from the one previously presented. For every snapshot at times $i = [1, 5]$ the algorithms have been executed in order to assign likelihood scores to the non-observed links and the link prediction performance has been evaluated with respect to every future time point $j = [i{+}1, 6]$. Considering a pair of time points (i, j), the non-observed links at time i are ranked by decreasing likelihood scores and the precision is computed as the percentage of links that appear at time j among the top-r in the ranking, where r is the total number of non-observed links at time i that appear at time j. Non-observed links at time i involving nodes that disappear at time j are not considered in the ranking. Table 5.1 reports for each method a 5-dimensional upper triangular matrix, containing as element (i, j) the precision of the link prediction from time i to time $j{+}1$.

As seen on nPSO synthetic networks, CH2-L2 outperformed SPM and ranked first in the prediction for all the pairwise time points, with a mean precision of 0.13 versus 0.09. It can be noticed that the precision improves as the two time points become further, going from 0.11 to 0.14 for CH2-L2 and from 0.08 to 0.11 for SPM. In addition to this, in line to what reported for small-size networks, CH2-L2 is faster than SPM and here the execution time is much smaller in favor of the local method, with a difference of around 5 hours, suggesting that the computational requirements of the global method considerably increase with the network size. This is actually in agreement with the computational complexity of the two methods, since SPM executes in

Table 5.1: Precision evaluation of link prediction in time on AS Internet networks. The table reports for each method a 5-dimensional upper triangular matrix, containing as element (i,j) the precision of the link prediction from time i to time j+1. On the right side, the methods are ranked by the mean precision computed over all the time combinations. The last column shows the time required for executing the methods on the biggest network (last snapshot, December 2010), after the removal of 10% of the links, as average over 10 iterations. For each comparison the best method is highlighted in bold.

CH2-L2					**SPM**						**mean precision**	**mean ranking**	**mean time**
0.12	**0.13**	**0.13**	**0.14**	**0.14**	0.08	0.09	0.09	0.10	0.11	**CH2-L2**	**0.13**	**1**	**1.2 h**
	0.11	**0.12**	**0.14**	**0.14**		0.07	0.08	0.09	0.10	SPM	0.09	2	6.8 h
		0.12	**0.13**	**0.14**			0.08	0.09	0.10				
			0.12	**0.13**				0.08	0.09				
				0.13					0.09				

$O(N^3)$ whereas CH2-L2, on sparse networks as the ones considered (last 12 rows of Table 3.1 shows the low density), requires only $O(N^2)$ (for further details please refer to Section 3.2 and to the Appendix A). Differently from the removal and re-prediction framework in which the set of missing links is artificially generated by a random procedure, here the set of links that will appear between two consecutive time points is given by ground-truth information, which makes the result even more significant and truthful, confirming the effectiveness of CH2-L2.

Since the above considered Internet networks were characterized by a high number of nodes and the local CH2-L2 model outperformed the best global model SPM (and SBM could not reach these predictions), one can advance the hypothesis that network size could play an important role. Consequently, with a substantial computational effort, we created the first study ever conducted to also perform and include removal and re-prediction evaluation for the best global and local methods on 12 large-size networks (from 3000 up to 40000 nodes), several statistics are shown as last entries of Table 3.1. Considering the same evaluation framework described for the small-size real networks, Table 5.2 shows the precision for each network, the mean precision and mean ranking for a further comparison of the two overall performance scores in discussion; the respective p-value of the permutation test (10000 iterations) for the mean ranking is 0.06. Although the difference is not statistically significant, we can appreciate how CH2-L2 surpasses SPM in 8 out of 12 networks (and one tie). This result is confirmed by the mean ranking, 1.29 for CH2-L2 against 1.71 for SPM. Table 5.3 reports the computational time required by the methods in order to perform the link prediction. The table is a further confirmation of the lower computational complexity of CH2-L2, as already discussed. Noteworthy is the increase of time from *thesaurus* (around 24000 nodes) to *facebook* (around 44000 nodes), where CH2-L2 goes from 1.3 to 2.1 hours, whereas SPM passes from 2.5 to 15.3 hours, pointing out in a tangible way the stronger computational time dependency on the network size.

Table 5.2: Link prediction on large-size real networks.
For each network 10% of links have been randomly removed (10 repetitions) and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The table reports for each network the mean precision over the random repetitions. The last rows show the mean precision and the mean ranking over the entire dataset. For each network the best method is highlighted in bold. The networks are sorted by increasing number of nodes N. For each pair of methods, a permutation test for the mean has been applied to the two vectors of pairwise link prediction rankings on each network using 10000 iterations. The table reports in the last row the pairwise p-value, adjusted for multiple hypothesis comparison by the Benjamini–Hochberg correction.

	CH2-L2	**SPM**
odlis	**0.12**	0.08
advogato	**0.17**	0.15
arxiv astroph	0.60	**0.67**
thesaurus	0.06	**0.07**
arxiv hepth	0.22	**0.27**
ARK200909	**0.17**	0.10
ARK200912	**0.17**	0.09
ARK201003	**0.17**	0.10
ARK201006	**0.17**	0.10
ARK201009	**0.18**	0.10
ARK201012	**0.18**	0.11
facebook	**0.10**	**0.10**
mean precision	**0.19**	0.16
mean ranking	**1.29**	1.71
p-value	0.0576	

These considerations on large-size real networks led to the assertion that the conclusions we can infer from the results on the nPSO model are in general true to predict the behavior of the algorithms also on real networks. Furthermore, in this study it has been extensively proven that global models are not always better than local methods and actually often the opposite is true (at least for CH2-L2 versus SBM and SPM), as we have shown with regards to the mean precision, the mean ranking and the computational time. For a better understanding of the mechanics interplaying in the CH2-L2 model we refer to two recent articles [35], [36] and to Chapter 2.

A particular doubt emerges due to the fact that CH2-L2 is able to outperform SPM on nPSO artificial networks and large-size real networks, whereas, paradoxically, in small-size real networks and on other synthetic models the contrary is true. We thought that this could be due to the hyperbolicity of the networks. It is known that the nPSO model generates artificial net-

Table 5.3: Computational time evaluation of link prediction on large-size real networks.
For each network, 10% of links have been randomly removed and the algorithms have been executed in order to assign likelihood scores to the non-observed links in these reduced networks. The table reports for each large-size real network the mean precision and the mean computational time (*h* for hours, *min* for minutes) over the random iterations for the entire dataset. The networks are sorted by increasing number of nodes N. The methods have been run in the same workstation, see the Appendix B for further details. Note that only one Internet network is shown in the table, for the others similar computational timings apply.

	CH2-L2	**SPM**
odlis	0.7 min	0.3 min
advogato	2.2 min	1.6 min
arxiv_astroph	21.9 min	70.8 min
thesaurus	1.3 h	2.5 h
arxiv_hepth	1.1 h	3.9 h
ARK201012	1.2 h	6.8 h
facebook	2.1 h	15.3 h
mean	0.9 h	4.2 h

works with an underlying hyperbolic geometry [88], [44] and large-size real networks, in order to make efficient the navigation and the global information delivery [89], are often characterized by a marked hyperbolic geometry [44], [90]. A representative case is the one of the Internet AS topologies: many studies demonstrated that they have a distinct hyperbolic geometry, in fact the greedy routing efficiency, robustness and scalability is generally maximized when the space is hyperbolic, both in single-layer [90–92] and in multiplex networks [93]. However, this might not be necessarily true for the small-size real networks, where, due to the reduced size, the density is high, and therefore their topology cannot often respect a hyperbolic geometry. This is confirmed by the results in Table 3.1, where it is shown that all the small-size networks (first 25 entries) have a high density in comparison to large-size networks (last 12 entries). Since a peculiar and necessary feature of networks with underlying hyperbolic geometry is a scale-free degree distribution [44], [90], [94], we performed a comparison between the estimated power-law degree distribution exponents of small-size and large-size real networks. As highlighted in Figure 5.1, the large-size real networks have a significantly lower exponent (p-value < 0.01) and therefore are characterized by a significantly higher power-lawness than small networks. Furthermore, the large-size networks average value is 2.54, which perfectly falls in the typical range $2 < \gamma < 3$ [81]. On the contrary, small-size networks have a mean exponent of 4.30 (4.22 without

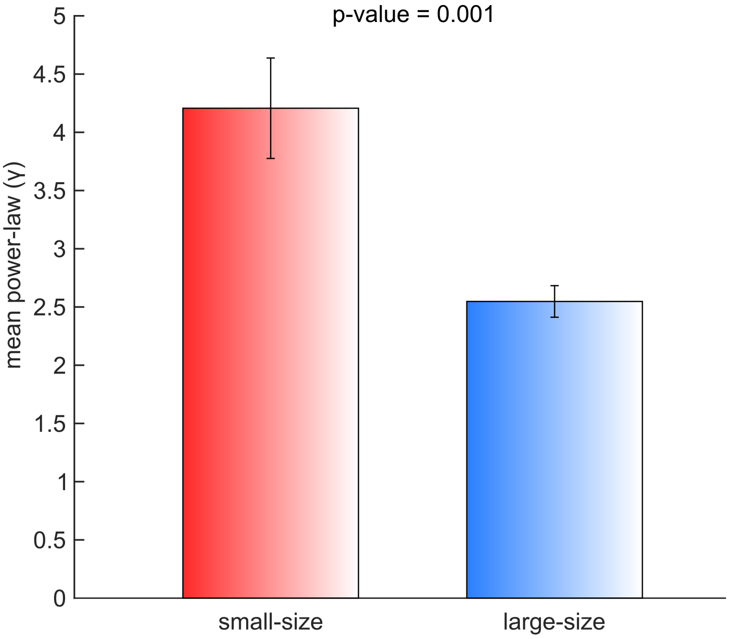

Figure 5.1: Comparison of power-law exponent between small-size and large-size real networks.
The barplot reports the mean and standard error of the power-law exponent γ estimated from the observed degree distribution of the small-size and large-size real networks. A permutation test for the mean (10000 iterations) has been applied to the two vectors of power-law exponents (rightmost column of Table 3.1) and the p-value is shown on top of the barplot.

considering the dataset of 486 brain networks), which represents an outlier with respect to that range. In brief, CH2-L2 is a physical rule that exhibits a stronger performance in comparison to a general learning-algorithm in networks characterized by an underlying hyperbolic geometry, hence emerges the speculation that the physical model behind CH2-L2 might be able to well capture the dynamics of organization of systems with this intrinsic characteristic. In fact, CH2-L2 might be one of the basic principles and generative mechanisms that contributes to give origin to the growth of hyperbolic networks by facilitating the transition from local-tunnels (i.e. the ensemble of all the local paths, which can be the smallest shortest-paths definable on a given network topology or the paths of a fixed arbitrary length, that connect two nonadjacent nodes, extremities of the tunnel) to local-rings (i.e. the closure of a local-tunnel obtained by adding to the topology the missing link for which the likelihood to appear is computed) and, in turn, generating local-community link-clustering in the network topology [35]. Previous studies demonstrated how bioinspired modelling can capture the basic dynamics of network adaptability through iteration of local rules, and produces in few hours of computing solutions with properties comparable to or better than those of real-world infrastructure networks, which would require many months of designing by teams of engineers [95]. Similarly, this thesis aims at promoting interest for both bioinspired computing and network automata, demonstrating that a simple unsupervised rule that emulates principles of network self-organization and adaptiveness arising during learning in living intelligent systems (like the brain), can equiperform, and sometimes outperform, advanced learning-machines (algorithms based on inference such as SPM, SBM and FBM) that exploit global network information. Furthermore, in support to the more accurate predictions of CH2-L2 on the time-evolving AS topologies, a recent study highlighted similar optimization principles between synaptic plasticity rules that regulate neural network activity and algorithms commonly used for controlling the flow of data in engineered networks such as Internet [96]. In particular, the Additive Increase and Multiplicative Decrease (AIMD) rule, which is the congestion control algorithm adopted in the Transmission Control Protocol (TCP) of the Internet [97], has also strong theoretical and experimental support for Long Term Potentiation (LTP) and Long Term Depression (LTD) in brain [96]. Moreover, the algorithm is very similar to an edge-weight update rule shown to produce stable Hebbian learning compared to many other rules [98], [99]. This similarity was at the moment proven only for changing weights of existing connectivity, hence it represents a *geometrical learning*. The results presented here are promising because they pave the way to extend the similarity between neural networks and Internet networks architectures also from the mere topological point of view, where, according to the LCP theory and the related epitopological learning, the process of *structural learning* is given by addition or deletion of connectivity.

The results in Figure 4.2 for the nPSO model and in Tables 5.1 and 5.2 for large-size real networks, as already discussed, show that in most of the cases CH2-L2 is able to predict links in the hyperbolic networks with a precision even higher than all the global methods. This is an evidence in support of the hypothesis that CH2-L2 might be one of the basic principles and generative mechanisms that contributes to give origin to the growth of hyperbolic networks by facilitating the transition from local-tunnels to local-rings and, in turn, growing local-community link-clustering in the network topology. Therefore, to prove this intuition many evaluation frameworks have been performed in order to give a concrete proof based on simulations that CH2-L2 is a generative rule particularly valid for hyperbolic geometry.

To this aim, it should be noticed that scale-freeness seems a necessary condition for hyperbolicity [90], [94]. This means that non-scale-free networks are non-hyperbolic, therefore theoretically if it is true that CH2-L2 is a generative rule particularly valid for hyperbolic geometry, then on non-hyperbolic networks the link prediction performance of CH2-L2 should be reduced and inferior to SPM, like we noticed in real small-size networks that having a high power-law exponent are weakly hyperbolic. Actually, to be more precise, since small-size real networks seem weakly hyperbolic, CH2-L2 performance was in general lower than SPM but not 'significantly' lower

- from a statistical point of view - because it is very rare to detect real networks that are not scale-free at all and therefore not hyperbolic. However, using an artificial random model of non-scale-free networks, like the Watts-Strogatz model [48], whose input parameters give the possibility to tune the clustering coefficient, we were able to prove that the performance of CH2-L2 is significantly lower than SPM for high levels of clustering. In fact, according to a recent study of Krioukov [100], non-scale-free networks with strong clustering have a latent network geometry that is Euclidean. On the other hand, using the same Watts-Strogatz model with low level of clustering, the random networks lose any latent geometry, and therefore both link predictors should dramatically lose their prediction power in general. These conclusions are solidly and perfectly reflected in Figure 4.5 and they ultimately demonstrate that latent geometry is at the basis of link prediction and that SPM performs better for Euclidean latent geometry given by non-scale-free and non-hyperbolic networks, whereas CH2-L2 performs better for hyperbolic latent geometry given by scale-free hyperbolic networks. Since many real world networks tend to exhibit hyperbolicity, and therefore scale-freeness, this simulation on the Watts-Strogatz model is the demonstration that the finding that CH2-L2 seems to perform better than SPM and than the other global methods on real networks is true in general and has theoretical well-grounded basis in the latent geometry of the real networks. Additionally, it has been seen that also in Watts-Strogatz networks SPM and CH2-L2 obtain overall the best and more robust performance with respect to SBM, which has a particular drop in precision for N = 500-1000, and FBM, whose discrepancy with respect to the other methods is huge for N= 100. Again, in order to verify to a greater extent the behavior in non-hyperbolic networks, an evaluation on Euclidean RGG networks has been conducted, which has proven that SPM outperforms the other methods. In this scenario the second method is CH2-L2 and SBM is the last one.

These last tests on RGG and Watts-Strogatz networks further certifies the inference difficulties of the SBM algorithm, which we propose to definitely prove applying the link prediction algorithms based on the SBM theory to synthetic networks generated through the same SBM theory itself, i.e. the LFR networks (which are based on a slight variation of the degree corrected SBM). For networks of 100 nodes the results of the best algorithm of the SBM family, i.e. SBM N, are comparable with the ones obtained using CH2-L2, but has to be noticed that they are lower than the best global method SPM. As the network size increases the difference in performance between SPM and the other methods becomes more evident and CH2-L2 becomes clearly the second best method outperforming SBM N. These considerations are maintained when varying the number of communities (compare Figure 4.7 and 4.10) and this testifies the robustness of the claim. This evaluation framework offered an ultimate proof of the inference problems of the SBM family in general and it has been confirmed that, regardless of the generative model used for creating the artificial networks, SBM does not perform at the same level as the other methods, and therefore as a model-learning machine it is not able to generalize enough, at least not as much as SPM and CH2-L2 for link prediction. This suggests that SBM overfits the structure of the network used for learning. In fact, a recent study [101] highlighted that if only the single partition with the highest posterior probability is used for predicting the links, the performance decreases with respect to the case in which an ensemble of likely partitions are considered (as in the algorithms here adopted), because the single partition significantly overfits the network. What all these simulations additionally spot out is that, although the ensemble procedure should mitigate the overfitting, it seems that it still remains a major drawback of the SBM-based algorithms.

6 | Conclusions and Future Works

In this study, an extensive evaluation of state-of-the-art topological link prediction methods for complex networks has been conducted.
Based on the results obtained from the analysis, some guidelines emerged for forthcoming link prediction studies. First, the widespread consideration of SBM as a state-of-the-art baseline for a performance comparison with new proposed methods should be firmly rejected. From the wide evaluations presented appear clear the significant outperformance of SPM and CH2-L2, respectively as best global and local methods, therefore we strongly encourage their adoption as references for a fair comparison to the state-of-the-art. Second, in order to prevent erroneous or partial conclusions, it has been stressed the importance to follow a robust evaluation framework, based on multiple types of link prediction evaluations. Methods should be tested on different frameworks, for instance re-prediction of randomly removed links and prediction in time-evolving networks. Both real networks and artificial models should be taken into account, considering a rich benchmark dataset that ranges over different network sizes, diverse topological features and various nature of the networks. Among the ones tested, the nonuniform PSO model turned out to be the closest to generate artificial networks with realistic topologies, at least for large size real networks. Precision should be adopted as metric of evaluation on the single network and, while comparing the methods over several networks, the precision-ranking should be used to obtain an unbiased score of overall performance.
Throughout the analysis it has been proven that the best state-of-the-art local method (CH2-L2) can perform equally or even better than the best state-of-the-art global approach (SPM). This is a remarkable result which contradicts the misleading common belief that global methods generally achieve higher performance with respect to local methods. In particular SBM has displayed poor inference capabilities on all the evaluation frameworks proposed, also on synthetic networks built with generative rules based on the same SBM theory.

6.1 Future Work

Several interesting points for future work emerges. First, it could be valuable to further expand the dataset of real-world networks, in order to robustly confirm the assertions made throughout the study. Second, it could be interesting to evaluate the performance of the three "theories" behind the link prediction methods for the inference on larger networks using CH2-L2, SPM and FBM (i.e. the only methods that can scale to large networks up to almost 50000 nodes). Third, it is important to further investigate the predictive power of RA-L3 on more classes of networks. Fourth, it is urgent to develop a synthetic networks generator able to produce networks with the required characteristics and statistics, allowing for more control on both clustering and hyperbolicity (features typically displayed in real world networks).

Finally, the detailed analysis of the links correctly predicted by the methods suggested that on average only half of them overlap between two different approaches, whereas the remaining part is peculiar of a single method and distributed in a similar abundance among the two. This offers a margin of improvement that could be exploited by a proper combination of methods and paves the way for the investigation of hybrid approaches potentially able to reach higher performance in topological link prediction. But, in order to build these ‘intelligent hybrid methods’ for topological link prediction, there is the urge to find the basis of a latent geometry theory of link prediction in complex networks, and this study aims to be a first landmark towards this direction.

A | CH2-L2 Complexity

The CH2-L2 algorithm for topological link prediction consists of a main loop going over all the non-observed links and at every iteration it independently evaluates the likelihood of one link. Given the number of nodes N and the number of observed links E, the number of iterations is:

$$\frac{N\left(N-1\right)}{2}-E$$

Considering an iteration in which the non-observed link between two nodes i and j is evaluated, the dominant operation is the intersection between the two sets of neighbors for finding the common-neighbors between i and j.
Since the set intersection complexity is linear in the number of elements, the cost is:

$$O\left(k_i+k_j\right)$$

Where k_i and k_j are the degrees of the nodes i and j.
Although different iterations could have different costs, the average complexity will be:

$$O\left(2\cdot avg_k\right)=O\left(4\cdot\frac{E}{N}\right)=O\left(\frac{E}{N}\right)$$

Where avg_k is the average node degree.
Given that there are $\frac{N\left(N-1\right)}{2}-E$ iterations with average complexity $O\left(\frac{E}{N}\right)$, the overall complexity is:

$$O\left(\left(\frac{N\left(N-1\right)}{2}-E\right)\frac{E}{N}\right)=O\left(\frac{E\left(N-1\right)}{2}-\frac{E^2}{N}\right)$$

Gathering the factor $\frac{E\left(N-1\right)}{2}$ we obtain:

$$O\left(\frac{E\left(N-1\right)}{2}\left(1-\frac{2E}{N\left(N-1\right)}\right)\right)=O\left(\frac{E\left(N-1\right)}{2}\left(1-D\right)\right)$$

Where D is the network density $D = \dfrac{2E}{N\left(N-1\right)}$.
Removing the multiplicative factor $\dfrac{1}{2}$ and the constant -1 on which N is dominant, we can rewrite in a more compact form:

$$O\left(EN\left(1-D\right)\right)$$

Let's analyze the complexity in three particular cases:

1. Minimum number of links for a connected network (tree): $E = N-1$

$$O\left(\frac{E\left(N-1\right)}{2}\left(1-\frac{2E}{N\left(N-1\right)}\right)\right) = O\left(\frac{\left(N-1\right)^2}{2}\left(1-\frac{2\left(N-1\right)}{N\left(N-1\right)}\right)\right) =$$

$$= O\left(\frac{\left(N-1\right)^2}{2} - \frac{\left(N-1\right)^2}{2} - \frac{\left(N-1\right)^2}{N}\right) = O\left(N^2 - \frac{N^2}{N}\right) = O\left(N^2\right)$$

2. Half of the number of possible links: $E = \dfrac{N\left(N-1\right)}{4}$

$$O\left(\frac{E\left(N-1\right)}{2}\left(1-\frac{2E}{N\left(N-1\right)}\right)\right) = O\left(\frac{N\left(N-1\right)^2}{8}\left(1-\frac{1}{2}\right)\right) =$$

$$= O\left(\frac{N\left(N-1\right)^2}{16}\right) = O\left(N^3\right)$$

3. Fully connected network (no non-observed links to evaluate): $E = \dfrac{N\left(N-1\right)}{2}$

$$O\left(\frac{E\left(N-1\right)}{2}\left(1-\frac{2E}{N\left(N-1\right)}\right)\right) = O\left(\frac{N\left(N-1\right)^2}{8}\left(1-1\right)\right) = 0$$

The analysis of the complexity function highlights that the complexity is $O\left(N^2\right)$ for sparse networks, it increases as the number of links increases reaching $O\left(N^3\right)$ for middle density, and then decreases arriving at a null computational cost at the maximum density, since there are not non-observed links to evaluate.
Due to the fact that reasonable values of density for real-networks are much lower than 0.5, as confirmed by Table 3.1, we may assert that within the domain of real and practical problems in which topological link prediction is applied, the complexity of CH2-L2 can be more simply expressed as $O\left(EN\right)$, and very often approximated by $O\left(N^2\right)$.

Note that the link likelihoods are computed independently from each other and therefore the implementation can be easily parallelized in order to speed up the running time.

B | Hardware and Software

Unless stated otherwise, Matlab code was used for all the simulations.

The simulations on small-size real networks have been carried out on a Dell workstation under Windows 7 professional 64-bit with 24 GB of RAM and one Intel(R) Xenon(R) X5660 processor with 2.80 GHz.

The simulations on large-size networks have been carried out on a workstation under Windows 8.1 Pro with 512 GB of RAM and two Intel(R) Xenon(R) CPU E5-2687W v3 processors with 3.10 GHz.

The simulations on the SBM variations have been performed on a workstation under Debian GNU Linux 9.4 64-bit with 264 GB of RAM and 32 Intel(R) Xeon(R) CPU E5-2650 v2 processors with 2.60 GHz.

All the other simulations have been run on nodes having 128 GB of RAM and two processors Intel(R) Xeon(R) CPU E5-2680 v3 (each with 12 cores) at 2.50 GHz.

Bibliography

[1] A. L. Barabási, "Linked: The new science of networks," 2003.

[2] T. G. Lewis, *Network science: Theory and applications.* John Wiley & Sons, 2011.

[3] N. R. Council *et al.*, *Network science.* National Academies Press, 2006.

[4] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[5] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[6] 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," *Scientific Reports*, vol. 3, no. 1613, pp. 1–13, 2013.

[7] J. O'Madadhain, J. Hutchins, and P. Smyth, "Prediction and ranking algorithms for event-based network data," *ACM SIGKDD explorations newsletter*, vol. 7, no. 2, pp. 23–30, 2005.

[8] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, p. 98, 2008.

[9] J. Kunegis, E. W. De Luca, and S. Albayrak, "The link prediction problem in bipartite networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6178 LNAI, 2010, pp. 380–389.

[10] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang, "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4511–4515, 2010.

[11] S. Daminelli, J. M. Thomas, C. Durán, and C. V. Cannistraci, "Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks," *New Journal of Physics*, vol. 17, no. 11, p. 113037, 2015.

[12] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, "Bipartite network projection and personal recommendation," *Physical Review E*, vol. 76, no. 4, p. 046115, 2007.

[13] W. Zeng, M.-S. Shang, Q.-M. Zhang, L. Lü, and T. Zhou, "Can dissimilar users contribute to accuracy and diversity of personalized recommendation?" *International Journal of Modern Physics C*, vol. 21, no. 10, pp. 1217–1227, 2010.

[14] Q.-M. Zhang, M.-S. Shang, W. Zeng, Y. Chen, and L. Lü, "Empirical comparison of local structural similarity indices for collaborative-filtering-based recommender systems," *Physics Procedia*, vol. 3, no. 5, pp. 1887–1896, 2010.

[15] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-commerce recommendation applications," *Data mining and knowledge discovery*, vol. 5, no. 1-2, pp. 115–153, 2001.

[16] P. Holme and M. Huss, "Role-similarity based functional prediction in networked systems: application to the yeast proteome," *Journal of the Royal Society Interface*, vol. 2, no. 4, pp. 327–333, 2005.

[17] Z. Huang and D. D. Zeng, "A link prediction approach to anomalous email detection," in *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, vol. 2. IEEE, 2006, pp. 1131–1136.

[18] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos, "Using ghost edges for classification in sparsely labeled networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2008, pp. 256–264.

[19] C. Durán, S. Daminelli, J. M. Thomas, V. J. Haupt, M. Schroeder, and C. V. Cannistraci, "Pioneering topological methods for network-based drug–target prediction by exploiting a brain-network self-organization theory," *Briefings in Bioinformatics*, vol. 8, no. W1, pp. 3–62, 2017.

[20] M. Jalili, Y. Orouskhani, M. Asgari, N. Alipourfard, and M. Perc, "Link prediction in multiplex online social networks," *Royal Society Open Science*, no. 4, p. 160863, 2017.

[21] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 52, pp. 22 073–22 078, 2009.

[22] 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.

[23] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, "Toward link predictability of complex networks," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, pp. 2325–2330, 2015.

[24] W. Peng, X. Baowen, W. Yurong, Z. Xiaoyu, W. Citation, X. Peng, W. Baowen, and Z. X. Yurong, "Link Prediction in Social Networks: the State-of-the-Art," *Sci China Inf Sci*, vol. 58, no. 58, pp. 11 101–38, 2015.

[25] M. Kim and J. Leskovec, "The Network Completion Problem: Inferring Missing Nodes and Edges in Networks," *SIAM International Conference on Data Mining*, pp. 47–58, 2011.

[26] X. Feng, J. Zhao, and K. Xu, "Link Prediction in Complex Networks: A Clustering Perspective," *The European Physical Journal B*, vol. 85, no. 3, p. 9, 2012.

[27] E. Dong, J. Li, and Z. Xie, "Link Prediction via Convex Nonnegative Matrix Factorization on Multiscale Blocks," *Journal of Applied Mathematics*, vol. 2014, 2014.

[28] L. Pan, T. Zhou, L. Lü, and C.-K. Hu, "Predicting missing links and identifying spurious links via likelihood analysis," *Scientific Reports*, vol. 6, pp. 1–10, 2016.

[29] J. Ding, L. Jiao, J. Wu, and F. Liu, "Prediction of missing links based on community relevance and ruler inference," *Knowledge-Based Systems*, vol. 98, pp. 200–215, 2016.

[30] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, 2014.

[31] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 83, no. 1, 2011.

[32] X. Zhang, X. Wang, C. Zhao, D. Yi, and Z. Xie, "Degree-corrected stochastic block models and reliability in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 393, pp. 553–559, 2014.

[33] L. Gulikers, M. Lelarge, and L. Massoulié, "A spectral method for community detection in moderately sparse degree-corrected stochastic block models," *Advances in Applied Probability*, vol. 49, no. 3, pp. 686–721, 2017.

[34] Y. Zhao, E. Levina, and J. Zhu, "Consistency of community detection in networks under degree-corrected stochastic block models," *The Annals of Statistics*, vol. 40, no. 4, pp. 2266–2292, 2012. [Online]. Available: http://projecteuclid.org/euclid.aos/1358951382

[35] 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," 2018.

[36] A. Muscoloni, U. Michieli, and C. V. Cannistraci, "Local-ring network automata and the impact of hyperbolic geometry in complex network link-prediction," 2018. [Online]. Available: http://arxiv.org/abs/1707.09496

[37] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 64, no. 2, p. 4, 2001.

[38] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[39] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks." *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 80, no. 4 Pt 2, p. 046122, 2009.

[40] I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao, M. A. Calderwood, M. Vidal, and A.-L. Barabási, "Network-based prediction of protein interactions," *bioRxiv*, p. 275529, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/03/02/275529.full.pdf+html

[41] T. P. Peixoto, "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 89, no. 1, 2014.

[42] ——, "The Graph-tool Python Library," *Figshare*, 2014.

[43] A. Muscoloni and C. V. Cannistraci, "A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities," *New Journal of Physics*, 2018.

[44] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguna, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, 2012.

[45] G. McLachlan and D. Peel, *Finite Mixture Models*, N. Hoboken, Ed. John Wiley & Sons, Inc., 2000.

[46] E. N. Gilbert, "Random Plane Networks," *Journal of the Society for Industrial and Applied Mathematics*, vol. 9, no. 4, pp. 533–543, 1961.

[47] J. Dall and M. Christensen, "Random geometric graphs," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 66, no. 1, pp. 1–9, 2002.

[48] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[49] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical review E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, vol. 78, no. 4, p. 46110, 2008.

[50] S. Fortunato and D. Hric, "Community detection in networks: A user guide," pp. 1–44, 2016.

[51] D. D. Bock, W.-C. A. Lee, A. M. Kerlin, M. L. Andermann, G. Hood, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim, and R. C. Reid, "Network anatomy and in vivo physiology of visual cortical neurons." *Nature*, vol. 471, no. 7337, pp. 177–182, 2011.

[52] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[53] D. Baird, J. Luczkovich, and R. R. Christian, "Assessment of spatial and temporal variability in ecosystem attributes of the St Marks national wildlife refuge, Apalachee Bay, Florida," *Estuarine, Coastal and Shelf Science*, vol. 47, no. 3, pp. 329–349, 1998.

[54] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait?" *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

[55] J. E. Cohen, D. N. Schittler, D. G. Raffaelli, and D. C. Reuman, "Food webs are more than the sum of their tritrophic parts," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22 335–22 340, 2009. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0910582106

[56] R. Kötter, "Online Retrieval, Processing, and Visualization of Primate Connectivity Data From the CoCoMac Database," *Neuroinformatics*, vol. 2, no. 2, pp. 127–144, 2004.

[57] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. V. den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 166–180, 2011.

[58] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *PNAS*, vol. 99, no. 12, pp. 7821–7826, 2002.

[59] J. Coleman, E. Katz, and H. Menzel, "The diffusion of an innovation among physicians," *Sociometry*, vol. 20, no. 4, pp. 253–270, 1957.

[60] M. Sageman, "Understanding terror networks." *International journal of emergency mental health*, vol. 7, no. 1, pp. 5–8, 2005.

[61] R. Michalski, S. Palus, and P. Kazienko, "Matching Organizational Structure and Social Network Extracted from Email Communication," *Business Information Systems*, vol. 87, pp. 197–206, 2011.

[62] P. M. Geiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, no. 4, pp. 565–573, 2003.

[63] L. C. Freeman, C. M. Webster, and D. M. Kirke, "Exploring social structure using dynamic three-dimensional color images," *Social Networks*, vol. 20, no. 2, pp. 109–118, 1998.

[64] L. Harriger, M. P. van den Heuvel, and O. Sporns, "Rich Club Organization of Macaque Cerebral Cortex and Its Role in Network Communication," *PLoS ONE*, vol. 7, no. 9, 2012.

[65] A. J. M. Martin, M. Vidotto, F. Boscariol, T. Di Domenico, I. Walsh, and S. C. E. Tosatto, "RING: networking interacting residues, evolutionary information and energetics in protein structures," *Bioinformatics*, vol. 27, no. 14, pp. 2003–2005, 2011.

[66] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," in *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, 2007, pp. 606–620.

[67] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 74, no. 3, 2006.

[68] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[69] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions." *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, vol. 68, no. 6, pp. 1–4, 2003.

[70] L. A. Adamic and N. Glance, "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog," *LinkKDD 2005*, pp. 36–43, 2005.

[71] J. M. Reitz, *Online Dictionary for Library and Information Science*, 2002.

[72] P. Massa, M. Salvetti, and D. Tomasoni, "Bowling alone and trust decline in social network sites," in *8th IEEE International Symposium on Dependable, Autonomic and Secure Computing, DASC 2009*, 2009, pp. 658–663.

[73] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and Shrinking Diameters," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 2007.

[74] G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper, "An associative thesaurus of English and its computer analysis," in *The computer and literary studies*, A. J. Aitkin, R. W. Bailey, and N. Hamilton-Smith, Eds. Edinburgh: University Press, 1973.

[75] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the Evolution of User Interaction in Facebook," *Proceedings of the 2nd ACM workshop on Online social networks - WOSN '09*, p. 37, 2009.

[76] K. Claffy, Y. Hyun, K. Keys, M. Fomenkov, and D. Krioukov, "Internet mapping: From art to science," in *Proceedings - Cybersecurity Applications and Technology Conference for Homeland Security, CATCH 2009*, 2009, pp. 205–211.

[77] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, "The minimal preprocessing pipelines for the Human Connectome Project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[78] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, and E. Yacoub, "The Human Connectome Project: A data acquisition perspective," pp. 2222–2231, 2012.

[79] M. P. van den Heuvel, L. H. Scholtens, L. Feldman Barrett, C. C. Hilgetag, and M. A. de Reus, "Bridging Cytoarchitectonics and Connectomics in Human Cerebral Cortex,"

Journal of Neuroscience, vol. 35, no. 41, pp. 13 943–13 948, 2015. [Online]. Available: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.2630-15.2015

[80] M. P. van den Heuvel, L. H. Scholtens, M. A. de Reus, and R. S. Kahn, "Associated Microscale Spine Density and Macroscale Connectivity Disruptions in Schizophrenia," *Biological Psychiatry*, vol. 80, no. 4, pp. 293–301, 2016.

[81] A. Clauset, C. Rohilla Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[82] W. Wang, F. Cai, P. Jiao, and L. Pan, "A perturbation-based framework for link prediction via non-negative matrix factorization," *Scientific Reports*, vol. 6, no. December, p. 38938, 2016.

[83] Y. Yang, R. N. Lichtenwalter, N. V. Chawla, Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, "Evaluating link prediction methods," *Knowledge and Information Systems*, vol. 45, pp. 751–782, 2015.

[84] R. Pech, D. Hao, L. Pan, H. Cheng, and T. Zhou, "Link Prediction via Matrix Completion," *CoRR*, vol. abs/1606.0, 2016.

[85] N. E. Ziv and E. Ahissar, "New tricks and old spines," *Nature*, vol. 462, no. December, pp. 859–861, 2009.

[86] V. Corti, Y. Sanchez-Ruiz, G. Piccoli, A. Bergamaschi, C. V. Cannistraci, L. Pattini, S. Cerutti, A. Bachi, M. Alessio, and A. Malgaroli, "Protein fingerprints of cultured CA3-CA1 hippocampal neurons: Comparative analysis of the distribution of synaptosomal and cytosolic proteins," *BMC Neuroscience*, vol. 9, 2008.

[87] A. Muscoloni and C. V. Cannistraci, "Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction," *New Journal of Physics*, 2018.

[88] ——, "A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities," *New Journal of Physics*, vol. 20, p. 052002, 2018.

[89] M. Boguñá, D. Krioukov, and K. C. Claffy, "Navigability of complex networks," *Nature Physics*, vol. 5, no. 1, pp. 74–80, 2008.

[90] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá, "Hyperbolic geometry of complex networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 82, no. 3, p. 036106, 2010.

[91] M. Boguna, F. Papadopoulos, and D. Krioukov, "Sustaining the Internet with Hyperbolic Mapping," *Nature Communications*, vol. 1, no. 6, pp. 1–8, 2010.

[92] F. Papadopoulos, D. Krioukov, M. Boguñá, and A. Vahdat, "Greedy Forwarding in Dynamic Scale-free Networks Embedded in Hyperbolic Metric Spaces," in *Proceedings of the 29th Conference on Information Communications*, ser. INFOCOM'10. Piscataway: IEEE Press, 2010, pp. 2973–2981.

[93] K.-K. Kleineberg, M. Boguñá, M. Ángeles Serrano, and F. Papadopoulos, "Hidden geometric correlations in real multiplex networks," *Nature Physics*, vol. 12, no. November, p. DOI: 10.1038/NPHYS3812, 2016.

[94] D. Krioukov, F. Papadopoulos, A. Vahdat, and M. Boguñá, "Curvature and temperature of complex networks," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 80, no. 3, 2009.

[95] A. Tero, S. Takagi, K. Ito, D. P. Bebber, M. D. Fricker, K. Yumiki, R. Kobayashi, and T. Nakagaki, "Rules for Biologically Inspired Adaptive Network Design," *Science*, vol. 327, no. January, pp. 439–442, 2010.

[96] J. Y. Suen and S. Navlakha, "Using inspiration from synaptic plasticity rules to optimize traffic flow in distributed engineered networks," *CoRR*, vol. abs/1611.0, 2016.

[97] M. Corless, C. King, R. Shorten, and F. Wirth, *AIMD Dynamics and Distributed Resource Allocation.* USA: SIAM-Society for Industrial and Applied Mathematics, 2016.

[98] M. C. van Rossum, G. Q. Bi, and G. G. Turrigiano, "Stable Hebbian learning from spike timing-dependent plasticity." *The Journal of Neuroscience*, vol. 20, no. 23, pp. 8812–8821, 2000.

[99] G. Billings and M. C. W. van Rossum, "Memory retention and spike-timing-dependent plasticity." *Journal of neurophysiology*, vol. 101, no. 6, pp. 2775–2788, 2009.

[100] D. Krioukov, "Clustering Implies Geometry in Networks," *Physical Review Letters*, vol. 116, no. 20, pp. 1–5, 2016.

[101] T. Vallès-Català, T. P. Peixoto, R. Guimerà, and M. Sales-Pardo, "On the consistency between model selection and link prediction in networks," *arXiv:1705.07967*, pp. 1–12, 2017.