

Missing links prediction: Comparing machine learning with physics-rooted approaches

Francesca Santucci^{1,*}, Giulio Cimini^{2,3} and Tiziano Squartini^{1,4,5}¹*IMT School for Advanced Studies, Piazza San Francesco 19, 55100 Lucca, Italy*²*'Enrico Fermi' Research Center (CREF), Via Panisperna 89A, 00184 Rome, Italy*³*Department of Physics and INFN, 'Tor Vergata' University of Rome, 00133 Rome, Italy*⁴*Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy*⁵*INdAM-GNAMPA Istituto Nazionale di Alta Matematica 'Francesco Severi', P.le Aldo Moro 5, 00185 Rome, Italy*

(Received 5 February 2026; accepted 4 August 2026; published 7 October 2026)

An active research line within the broader field of network science is the one concerning link prediction. Close in scope to network reconstruction, link prediction targets specific connections with the aim of uncovering the missing ones, as well as predicting those most likely to emerge in the future, from the available information. In this paper, we consider two families of methods, i.e., those rooted in statistical physics and those based upon machine learning: the members of the first family identify missing links as the most probable nonobserved ones, the probability coefficients being determined by solving maximum-entropy benchmarks over the accessible network structure; the members of the second family, instead, associate the presence of single edges with explanatory node-specific variables. Running likelihood-based models such as the configuration model, or one of its many fitness-based variants, in parallel with the gradient boosting decision tree algorithm reveals that the accuracy of the former is comparable to the accuracy of the latter. Such a result confirms that white-box algorithms are viable competitors to the currently available black-box ones, being more interpretable and computationally faster.

DOI: [10.1103/76kb-hyp4](https://doi.org/10.1103/76kb-hyp4)

I. INTRODUCTION

Link prediction is an active research line within the broader field of network science. Close in scope to network reconstruction [1], link prediction targets specific connections, aiming to uncover missing ones and predict those most likely to emerge in the future [2]. Generally speaking, the link prediction problem can be stated by asking the following question: Given a supposedly incomplete snapshot of a network, can the most-likely “yet-to-come” edges be predicted? Such an issue is relevant in many research areas, such as those concerning socioeconomic and financial networks [3–7]: Knowing the structure of the commercial partnerships between firms or of the financial exchanges between banks is, in fact, relevant for a number of reasons, such as quantifying the risk associated with the propagation of a shock [7–9].

Link prediction algorithms rank unconnected node pairs on the basis of a *score*: While some methods rely on purely structural information, others admit external information such as node- and edge-specific features.

The simplest framework to carry out link prediction includes the so-called *similarity-based* algorithms: Scores, here,

are induced by some measure of similarity between nodes; to this aim, *local* [3,10,11], *quasilocal* [11], or *global* [12–15] information—such as the degree, the degree of common neighbors, or the length of paths connecting any two nodes—has been employed [2,3,11,16].

A more refined framework includes the so-called *likelihood-based* algorithms, defined by a likelihood function whose maximization provides the probability that any two nodes are connected; this is usually achieved by assuming that a certain amount of information is accessible, hence treating it as a constraint to account for: Examples are provided by the methods relying on a graph modular structure [17,18] or those induced by entropy maximization [6,19].

More computer-science-oriented approaches are, instead, defined by models whose parameters represent node- and edge-specific features [2,7,20,21]: Specifying a model, thus, amounts to learning a function that determines the probability for nodes i and j to be connected while taking as inputs features that may be node specific (e.g., the sales, the number of employees, the degree of country i and country j), edge specific (e.g., the geographic distance between countries i and j , the volume of sector-specific exchanges), or both.

In all cases, algorithms are trained on a portion of the whole network, i.e., the *training set*, while the prediction is carried out on the links of the unobserved part, i.e., the *test set*.

The present work is devoted to comparing the performance of different link prediction methods in recovering the missing portion of the edge set of the (binary, undirected representation of the) World Trade Web (WTW) and of the electronic market for interbank deposits (eMID). The framework is the

*Contact author: santuccifrancesca26@gmail.com

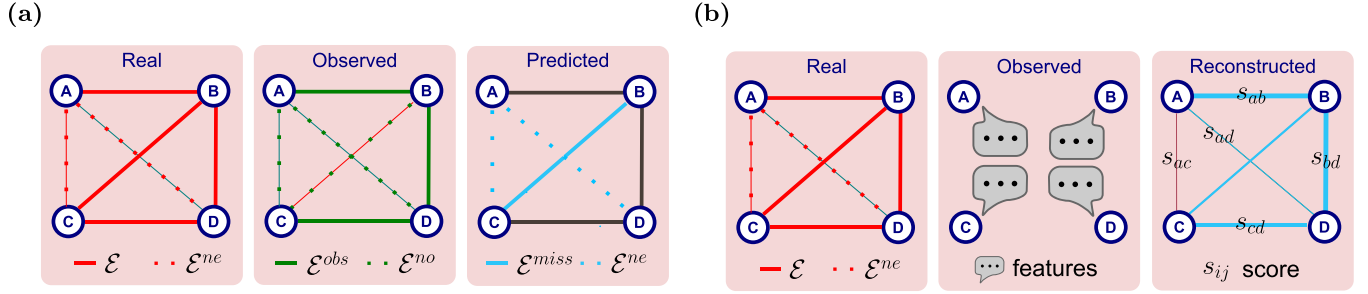


FIG. 1. Visual representation of the framework adopted to carry out link prediction in Ref. [6]. (a) The red, solid lines indicate the empirical edges, while the red, dotted lines indicate the nonexistent edges; the green, solid lines indicate the edges we have direct access to and the information about which can be employed to predict the ones indicated by the green, dotted lines; the aim of a link prediction exercise is understanding which of the nonobserved links is actually nonexistent (indicated by the light blue, dotted lines) and which is genuinely missing (indicated by the light blue, solid lines). (b) The three subpanels depict a slightly different framework, within which link prediction is carried out by employing node-specific features only—in a way that is reminiscent of network reconstruction exercises [9].

same as in Ref. [6], where the training set coincides with the whole network apart from a percentage of deleted connections and the proper prediction exercise solely concerns these missing edges (see also Fig. 1). Here, we consider algorithms belonging to each of the three aforementioned classes, such as the gravity model (GM) [taking as inputs “exogenous” features like each country’s gross domestic product (GDP) and their geographic distances], the configuration model (CM) (taking as inputs “endogenous” features like the degrees) and the gradient boosting decision tree (GBDT) (taking as inputs both “exogenous” and “endogenous” features [22]).

II. SETTING UP THE FRAMEWORK

Let us indicate with $\mathbf{A}$ the $N \times N$ adjacency matrix of the network we are considering, with $\mathcal{U}$ the set of all node pairs and with $\mathcal{E}$ the edge set: As a consequence, $\mathcal{E}^{\text{ne}} = \mathcal{U} \setminus \mathcal{E}$ indicates the set of nonexistent links [2,6]. The edge set is, then, partitioned into the set of observed, or accessible, links $\mathcal{E}^{\text{obs}}$ and the set of missing links $\mathcal{E}^{\text{miss}} = \mathcal{E} \setminus \mathcal{E}^{\text{obs}}$; finally, let us denote the union of the set of nonexistent links and the set of missing links—collecting the node pairs appearing as unconnected either because the link does not exist or because the information about its presence is missing—as $\mathcal{E}^{\text{no}} = \mathcal{E}^{\text{ne}} \cup \mathcal{E}^{\text{miss}} = \mathcal{U} \setminus \mathcal{E}^{\text{obs}}$ [2,6] [see also the first panel of Fig. 1(a)]. Link prediction algorithms output $|\mathcal{E}^{\text{no}}| = |\mathcal{E}^{\text{ne}}| + |\mathcal{E}^{\text{miss}}|$ scores, hereby indicated as $\{s_{ij}\}_{ij \in \mathcal{E}^{\text{no}}}$, to be assigned to the nonobserved links: While the $|\mathcal{E}^{\text{miss}}|$ largest scores identify the predicted missing links, the remaining $|\mathcal{E}^{\text{ne}}|$ scores identify the predicted nonexistent ones; let us indicate the two sets as $\overline{\mathcal{E}^{\text{miss}}}$ and $\overline{\mathcal{E}^{\text{ne}}}$.

A bit more compactly, upon considering that $\mathcal{U} = \mathcal{E}^{\text{obs}} \cup \mathcal{E}^{\text{miss}} \cup \mathcal{E}^{\text{ne}}$, one can define an $N \times N$ adjacency matrix $\mathbf{A}^{\text{obs}} = \{a_{ij}^{\text{obs}}\}_{ij \in \mathcal{U}}$, where

$$a_{ij}^{\text{obs}} = \begin{cases} 1, & \text{if } ij \in \mathcal{E}^{\text{obs}}, \\ 0, & \text{if } ij \in \mathcal{E}^{\text{miss}}, \\ 0, & \text{if } ij \in \mathcal{E}^{\text{ne}}, \end{cases} \quad (1)$$

in practice, the original matrix $\mathbf{A}$ where $|\mathcal{E}^{\text{miss}}|$ connections have been deleted; the *training set*, then, becomes $\mathbf{D}^{\text{tr}} = \{\mathbf{X}, \mathbf{A}^{\text{obs}}\}$, with $\mathbf{X} = \{\mathbf{x}_{ij}\}_{ij \in \mathcal{U}}$ and $\mathbf{x}_{ij}$ being a vector of fea-

tures, which may refer to nodes i and j , to the pair ij or both [see also the second panel of Fig. 1(a)]. Let us explicitly notice that (1) “endogenous” features are evaluated on the accessible portion of the dataset $\mathbf{A}^{\text{obs}}$; and (2) the structural portion of the training set consists of both connected and unconnected pairs of nodes—As we will see, the main difference between the black-box and white-box algorithms considered here lies precisely in the way such information is processed to make predictions.

After having trained a given model on $\mathbf{D}^{\text{tr}}$, prediction is performed on the *test set*, individuated by the node pairs with no observed connection: In symbols, $\mathbf{D}^{\text{te}} = \{a_{ij}\}_{ij \in \mathcal{E}^{\text{no}}}$ [see also the third panel of Fig. 1(a)].

The goodness of a given prediction performance can be evaluated according to a number of different statistical indicators. Let us indicate the number of true positives (i.e., the correctly identified missing links) with

$$\text{TP} = |\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{miss}}|, \quad (2)$$

the number of true negatives with

$$\text{TN} = |\overline{\mathcal{E}^{\text{ne}}} \cap \mathcal{E}^{\text{ne}}|, \quad (3)$$

and the number of false positives with

$$\text{FP} = |\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{ne}}|; \quad (4)$$

upon doing so, we can define the *true positives rate* (TPR) as

$$\text{TPR} = \frac{|\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{miss}}|}{|\mathcal{E}^{\text{miss}}|}, \quad (5)$$

the false positives rate (FPR) as

$$\text{FPR} = \frac{|\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{ne}}|}{|\mathcal{E}^{\text{ne}}|}, \quad (6)$$

the Jaccard index (JI) as

$$\text{JI} = \frac{|\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{miss}}|}{|\overline{\mathcal{E}^{\text{miss}}} \cup \mathcal{E}^{\text{miss}}|}, \quad (7)$$

and the accuracy (ACC) as

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{|\mathcal{E}^{\text{no}}|}. \quad (8)$$

In words, the TPR accounts for the percentage of missing links that are correctly identified as such; the JI enriches the picture provided by the TPR by accounting for the number of false positives too and the ACC accounts for the percentage of correctly identified links, be they missing or nonexistent.

Finally, the area under the receiver operating curve (AUROC) is defined as the area under the curve obtained upon scattering the TPR versus the FPR as the list of links ranked in decreasing order of the chosen score is gone through and quantifies the extent to which a given link prediction algorithm performs better than a random one—that flips a coin to classify each nonobserved link either as nonexistent or missing. A complementary index is the area under the precision-recall curve (AUPRC) [23], defined as the area under the curve obtained upon scattering the positive predictive value (PPV), defined as

$$\text{PPV} = \frac{|\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{miss}}|}{|\mathcal{E}^{\text{miss}}|}, \quad (9)$$

versus the TPR as the list of links ranked in decreasing order of the chosen score is gone through and quantifying the likelihood that a predictive positive is, indeed, a true one.

III. ALTERNATIVE SAMPLING SCHEMES

As our link prediction exercise needs the explicit definition of the aforementioned sets and there are many ways of selecting $|\mathcal{E}^{\text{miss}}|$ links out of $|\mathcal{E}|$ possible ones, we have sampled uniformly at random 10%, 20%, and 30% of links a system-specific number of times (see below) to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$; each of the statistical indicators above has, then, been averaged over such samples.

Such a sampling scheme, however, “favors” well-connected nodes; in order to test alternative ones, we have also considered a scheme that samples links from $\mathcal{E}$ with a probability that is inversely proportional to the product of the degrees of the involved vertices and the one named Random Node-Edge Sampler [24], sampling nodes and edges uniformly at random in two subsequent steps.

IV. MISSING LINKS PREDICTION ALGORITHMS

The present section is devoted to illustrate the missing links prediction algorithms considered in the present paper, noticing that each algorithm is characterized by a specific functional form of the prediction scores $\{s_{ij}\}_{ij \in \mathcal{E}^{\text{no}}}$: Given a node pair ij , each algorithm outputs a score s_{ij} accounting for the likelihood that the involved nodes are connected by a link, with higher values of s_{ij} indicating a higher chance that $ij \in \mathcal{E}$.

A. Econometric approaches

A simple, yet successful, recipe to model the WTW is the one provided by the GM [25,26]. According to it, the prediction score reads

$$s_{ij}^{\text{GM-I}} = \text{GDP}_i \text{GDP}_j, \quad (10)$$

where GDP_i (GDP_j) is the gross domestic product of country i (country j). A more refined GM-based link prediction recipe

reads

$$\begin{aligned} s_{ij}^{\text{GM-II}} &= \ln \left[\epsilon \frac{(\text{GDP}_i \text{GDP}_j)^\beta}{d_{ij}^\gamma} \right] \\ &= \alpha + \beta \ln(\text{GDP}_i \text{GDP}_j) - \gamma \ln d_{ij}, \end{aligned} \quad (11)$$

where d_{ij} is the geographic distance between countries i and j and $\alpha = \ln \epsilon$. To ease numerical manipulations, the GM features have been normalized, thus reading $\omega_i = \text{GDP}_i / \overline{\text{GDP}}$ and $\delta_{ij} = d_{ij} / \overline{d}$, where $\overline{\text{GDP}} = \sum_{i=1}^N \text{GDP}_i / N$ and $\overline{d} = 2 \sum_{i=1}^N \sum_{j(>i)} d_{ij} / (N(N-1))$.

Such a model can be fully determined upon maximizing the log-likelihood function

$$\begin{aligned} \mathcal{L}(\mathbf{A}^{\text{obs}} | \alpha, \beta, \gamma) \\ = \sum_{i=1}^N \sum_{j(>i)} [a_{ij}^{\text{obs}} \ln p_{ij} + (1 - a_{ij}^{\text{obs}}) \ln(1 - p_{ij})], \end{aligned} \quad (12)$$

where

$$p_{ij} = \frac{1}{1 + e^{-s_{ij}}}, \quad (13)$$

with $s_{ij} = s_{ij}^{\text{GM-II}}$. In order to evaluate the amount of information required by the recipe above, let us explicitly compute the gradient of the corresponding log-likelihood function. Since

$$\begin{aligned} \mathcal{L}(\mathbf{A}^{\text{obs}} | \alpha, \beta, \gamma) &= \sum_{i=1}^N \sum_{j(>i)} [a_{ij}^{\text{obs}} \ln(e^{s_{ij}}) - \ln(1 + e^{s_{ij}})] \\ &= \sum_{i=1}^N \sum_{j(>i)} [a_{ij}^{\text{obs}} s_{ij} - \ln(1 + e^{s_{ij}})], \end{aligned} \quad (14)$$

one finds

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{i=1}^N \sum_{j(>i)} \left[a_{ij}^{\text{obs}} - \frac{e^{s_{ij}}}{1 + e^{s_{ij}}} \right] \frac{\partial s_{ij}}{\partial \theta} \\ &= \sum_{i=1}^N \sum_{j(>i)} [a_{ij}^{\text{obs}} - p_{ij}] \frac{\partial s_{ij}}{\partial \theta}, \end{aligned} \quad (15)$$

where $\theta = \{\alpha, \beta, \gamma\}$, an expression showing that the required amount of information coincides with the entire set of observed connections.

The inferred coefficients, obtained by running the SCIKIT-LEARN Python package [27], are shown in Table I.

B. Machine learning approaches

The model we consider here is named GBDT and belongs to the family of *ensemble methods* that combine several base learners, trained in sequence, to achieve a better performance than the one that would be obtained by employing each of them singularly [7,28,29]: For binary classification tasks, the GBBDT outputs a raw score that is then mapped to a probability coefficient. More in detail, the raw score at the K th step reads

$$s_{ij}^{(K)} = \alpha + \sum_{k=1}^K \rho_{(k)} \phi(\mathbf{x}_{ij} | \theta_{(k)}), \quad (16)$$

TABLE I. Coefficients of the GM-based link prediction recipe defined by Eq. (11), trained on the years listed in the first column. The WTW has been considered in its binary, undirected representation.

Year	α	β	γ
1990	2.102 ± 0.020	0.436 ± 0.004	0.526 ± 0.009
1991	1.643 ± 0.159	0.409 ± 0.004	0.492 ± 0.007
1992	1.766 ± 0.191	0.425 ± 0.004	0.482 ± 0.009
1993	1.868 ± 0.012	0.443 ± 0.004	0.538 ± 0.013
1994	1.955 ± 0.195	0.442 ± 0.004	0.583 ± 0.028
1995	2.00 ± 0.027	0.442 ± 0.004	0.579 ± 0.015
1996	2.162 ± 0.017	0.449 ± 0.005	0.580 ± 0.016
1997	2.346 ± 0.035	0.472 ± 0.007	0.584 ± 0.016
1998	2.346 ± 0.014	0.474 ± 0.004	0.595 ± 0.019
1999	3.361 ± 0.017	0.477 ± 0.004	0.575 ± 0.015
2000	2.410 ± 0.030	0.486 ± 0.005	0.578 ± 0.017

i.e., is a summation over the aforementioned base learners, taking as input the vector of features $\mathbf{x}_{ij}$ and parametrized by θ ; the parameters ρ and θ are then learnt sequentially. More explicitly, at the zeroth step, one has

$$s_{ij}^{(0)} = \alpha, \quad (17)$$

where the value of α is chosen in order to minimize the loss function (equivalently, maximize the corresponding log-likelihood function)

$$\ell(\mathbf{A}^{\text{obs}}|\mathbf{s}^{(0)}) = - \sum_{i=1}^N \sum_{j(>i)} [a_{ij}^{\text{obs}} \ln p_{ij} + (1 - a_{ij}^{\text{obs}}) \ln(1 - p_{ij})], \quad (18)$$

where $p_{ij} = 1/(1 + e^{-s_{ij}^{(0)}})$. At the first step, one has

$$s_{ij}^{(1)} = \alpha + \rho_{(1)} \phi(\mathbf{x}_{ij}|\theta_{(1)}), \quad (19)$$

where $\theta_{(1)}$ is determined as

$$\theta_{(1)} = \arg \min_{\theta, \eta} \left\{ \sum_{i=1}^N \sum_{j(>i)} \left[- \frac{\partial \ell(a_{ij}^{\text{obs}}|s_{ij})}{\partial s_{ij}} \right]_{s_{ij}=s_{ij}^{(0)}} - \eta \phi(\mathbf{x}_{ij}|\theta) \right\}^2, \quad (20)$$

i.e., θ is chosen to minimize the distance from the gradient of the loss function, evaluated in $s_{ij} = s_{ij}^{(0)}$ [7,28]; subsequently, $\rho_{(1)}$ is chosen to minimize the “updated” loss function

$$\rho_{(1)} = \arg \min_{\rho} \left\{ \sum_{i=1}^N \sum_{j(>i)} \ell(a_{ij}^{\text{obs}}|\alpha + \rho \phi(\mathbf{x}_{ij}|\theta_{(1)})) \right\}. \quad (21)$$

By iterating the steps above K times, it is possible to build the predictor $s^{(K)}$ [28].

The base learners are usually chosen to be *decision trees* [28], hence returning a prediction by performing a sequence of data splits according to the decision rules embodied by the nodes of the trees, with the root corresponding to the initial split and the leaves corresponding to the prediction itself.

The GBDT, trained on a number of different combinations of features—i.e., each country’s GDP, each country’s GDP and their geographic distances, each country’s degree, and each country’s degree and their geographic distances—has been implemented by running the LIGHTGBM Python package [29]; all “endogenous” features—both those concerning the individual nodes i and j and those concerning the pair ij —are computed on $\mathbf{A}^{\text{obs}}$.

C. Entropy-based approaches

Maximum-entropy models have been widely employed, either to detect the empirical patterns of a networked configuration that cannot be interpreted as a mere by-product of the enforced constraints or to individuate the most likely networked configuration(s) that are compatible with the available information: also named exponential random graphs (ERGs) [30–36], the models of this class embody a physics-rooted approach to the study of network science [30,37,38], ensuring that the imposed constraints are preserved as ensemble averages.

More formally, entropy-based approaches define a probability distribution over a properly defined ensemble of graphs that is as random as possible, given a set of M constraints $\mathbf{C}(\mathbf{G}) = \{C_i(\mathbf{G})\}_{i=1}^M$ on the expected value of a number of observables: In symbols,

$$P(\mathbf{G}|\theta) = \frac{e^{-H(\mathbf{G},\theta)}}{Z(\theta)} = \frac{e^{-\mathbf{C}(\mathbf{G})\theta}}{Z(\theta)}, \quad (22)$$

where the parameters θ are the Lagrange multipliers associated with the constraints themselves. They can be numerically determined by maximizing the log-likelihood $\ln P(\mathbf{G}^*|\theta)$ of observing the empirical configuration $\mathbf{G}^*$. It is easy to prove that such a recipe leads to solve the system of nonlinear, coupled equations $\langle \mathbf{C} \rangle = \mathbf{C}(\mathbf{G}^*)$, i.e., $\langle C_i \rangle = C_i(\mathbf{G}^*)$, $i = 1, \dots, M$ [39–41].

In what follows, we will consider instances of the *separable* Hamiltonian [42]

$$H(\mathbf{A}, \theta) = \sum_{i=1}^N \sum_{j(>i)} H_{ij}(\mathbf{A}, \theta_{ij}) = \sum_{i=1}^N \sum_{j(>i)} a_{ij} \cdot f(\mathbf{A}, \theta_{ij}) \quad (23)$$

that induces a probability distribution reading

$$P(\mathbf{A}|\theta) = \prod_{i=1}^N \prod_{j(>i)} p_{ij}^{a_{ij}} (1 - p_{ij})^{1-a_{ij}}, \quad (24)$$

with

$$p_{ij} = \frac{e^{-f(\mathbf{A}, \theta_{ij})}}{1 + e^{-f(\mathbf{A}, \theta_{ij})}}; \quad (25)$$

notice that *linear* Hamiltonians can be recovered as special cases upon posing $f(\mathbf{A}, \theta_{ij}) = f(\theta_{ij})$.

Within such a framework, the $|\mathcal{E}^{\text{miss}}|$ largest coefficients of the set $\{p_{ij}\}_{(i,j) \in \mathcal{E}^{\text{no}}}$ identify a subgraph Σ^* whose probability reads

$$P(\Sigma^*|\theta) = \prod_i \prod_{j(>i)} p_{ij}^{\sigma_{ij}^*} (1 - p_{ij})^{1-\sigma_{ij}^*}, \quad (26)$$

TABLE II. Basic statistics concerning the WTW dataset [48], i.e., the total number of countries N , the total number of links L , the network density c , and the average degree $\bar{k}$.

Year	N	L	c	$\bar{k}$
1990	159	7639	0.608	96.08
1991	173	8249	0.554	95.36
1992	175	8552	0.562	97.74
1993	177	8837	0.567	99.85
1994	177	9052	0.591	103.94
1995	177	9199	0.591	103.94
1996	177	9623	0.618	108.73
1997	177	9864	0.633	111.46
1998	177	9866	0.633	111.48
1999	177	9864	0.633	111.46
2000	177	9865	0.633	111.47

the product running over the pairs of nodes in $\mathcal{E}^{\text{no}}$ and σ_{ij}^* being 1 only for the (i, j) s corresponding to the $|\mathcal{E}^{\text{miss}}|$ largest probability coefficients—the remaining entries reading zero: Among all subgraphs induced by precisely $|\mathcal{E}^{\text{miss}}|$ links, Σ^* is the one attaining the highest probability. In words, entropy-based methods lead us to interpret the nonobserved links that have been assigned the largest probability coefficients as the ones that are most likely to appear given the chosen constraints [6].

1. Link prediction via linear ERGs

Configuration model. Let us now consider the first specification of the entropy-based framework: It boils down to posing $\mathbf{C} = \mathbf{k}$, with $k_i = \sum_{j(\neq i)} a_{ij}$, an identification that leads to $H_{\text{CM}}(\mathbf{A}) = \sum_{i=1}^N \sum_{j(>i)} a_{ij} \cdot (\theta_i + \theta_j)$, further inducing

$$p_{ij}^{\text{CM}} = \frac{x_i x_j}{1 + x_i x_j}, \quad (27)$$

where $x_i = e^{-\theta_i}$; the parameters $\{x_i\}_{i=1}^N$ can be determined by solving the system of nonlinear, coupled equations

$$k_i(\mathbf{A}^{\text{obs}}) = \sum_{j(\neq i)} p_{ij}^{\text{CM}} = \sum_{j(\neq i)} \frac{x_i x_j}{1 + x_i x_j}, \quad \forall i, \quad (28)$$

a task that can be accomplished by implementing the iterative recipe [40]

$$x_i^{t+1} = \frac{k_i(\mathbf{A}^{\text{obs}})}{\sum_{j(\neq i)} \left(\frac{x_j^t}{1 + x_i^t x_j^t} \right)}. \quad (29)$$

The prescription above, characterizing the so-called CM [30,32–34,36], matches the one considered in Ref. [6], where the accessible portion of a graph was employed to estimate its degree sequence and, on this basis, obtain the probability for any two nodes to be connected—to be later employed as a score for the existence of a link between any two unconnected vertices. Such a procedure rests upon the assumption that if the accessible portion of a graph is reproduced with a certain accuracy by a certain topological information, the inaccessible portion will be reproduced with the same degree of accuracy by the same kind of information [43].

A simpler prescription ensuring that the expected degree sequence matches the observed one is provided by the Chung-Lu (CL) model:

$$p_{ij}^{\text{CL}} = \frac{k_i k_j}{2L}, \quad (30)$$

where k_i and k_j must be interpreted as the degrees of the nodes evaluated on the accessible portion of the dataset $\mathbf{A}^{\text{obs}}$, i.e., $k_i(\mathbf{A}^{\text{obs}})$ and $k_j(\mathbf{A}^{\text{obs}})$; as it can be recovered from the CM by posing $p_{ij}^{\text{CM}} \simeq x_i x_j$, the validity of this model is limited to the case in which the network is sparse and the degree distribution is not too broad [40].

Configuration model with distances. Let us now enrich the CM by adding information about the geographic distances between countries: This choice boils down to posing $\mathbf{C} = \{\mathbf{k}, d\}$, with $k_i = \sum_{j(\neq i)} a_{ij}$ and $d = \sum_{i=1}^N \sum_{j(>i)} a_{ij} d_{ij}$ [44,45]. Such an identification leads to $H_{\text{CMD}}(\mathbf{A}) = \sum_{i=1}^N \sum_{j(>i)} a_{ij} \cdot (\theta_i + \theta_j + \gamma d_{ij})$, an expression inducing

$$p_{ij}^{\text{CMD}} = \frac{x_i x_j w^{d_{ij}}}{1 + x_i x_j w^{d_{ij}}}, \quad (31)$$

where $x_i = e^{-\theta_i}$ and $w = e^{-\gamma}$. The parameters $\{x_i\}_{i=1}^N$ and w can be determined by solving the system of nonlinear, coupled equations

$$\begin{aligned} k_i(\mathbf{A}^{\text{obs}}) &= \sum_{j(\neq i)} p_{ij}^{\text{CMD}}, \quad \forall i, \\ d(\mathbf{A}^{\text{obs}}) &= \sum_{i=1}^N \sum_{j(>i)} p_{ij}^{\text{CMD}} d_{ij}, \end{aligned} \quad (32)$$

a task that can be accomplished by implementing the iterative recipe

$$\begin{aligned} x_i^{t+1} &= \frac{k_i(\mathbf{A}^{\text{obs}})}{\sum_{j(\neq i)} \left(\frac{x_j^t w^{d_{ij}}}{1 + x_i^t x_j^t w^{d_{ij}}} \right)}, \quad \forall i, \\ w_{t+1} &= \frac{d(\mathbf{A}^{\text{obs}})}{\sum_{i=1}^N \sum_{j(>i)} \left(\frac{x_i^t x_j^t w^{d_{ij}-1}}{1 + x_i^t x_j^t w^{d_{ij}}} \right) d_{ij}}; \end{aligned} \quad (33)$$

notice how the configuration model with distances (CMD) relies upon both “endogenous” (i.e., topological) and “exogenous” (i.e., nontopological) features. To ease numerical manipulations, the distances have been normalized, thus reading $\delta_{ij} = d_{ij}/\bar{d}$, where $\bar{d} = 2 \sum_{i=1}^N \sum_{j(>i)} d_{ij}/(N(N-1))$.

Fitness model. First introduced in Ref. [46] as a generative mechanism alternative to the “rich-gets-richer” one, the fitness model (FM) prescribes to associate each node with a hidden variable: Following Ref. [47], here we pose $x_i = \sqrt{z} \omega_i$, with $\omega_i = \text{GDP}_i / \overline{\text{GDP}}$ and $\overline{\text{GDP}} = \sum_{i=1}^N \text{GDP}_i / N$, thus obtaining

$$p_{ij}^{\text{FM}} = \frac{z \omega_i \omega_j}{1 + z \omega_i \omega_j}. \quad (34)$$

The parameter z can be inferred by imposing that the expected number of links matches the observed one

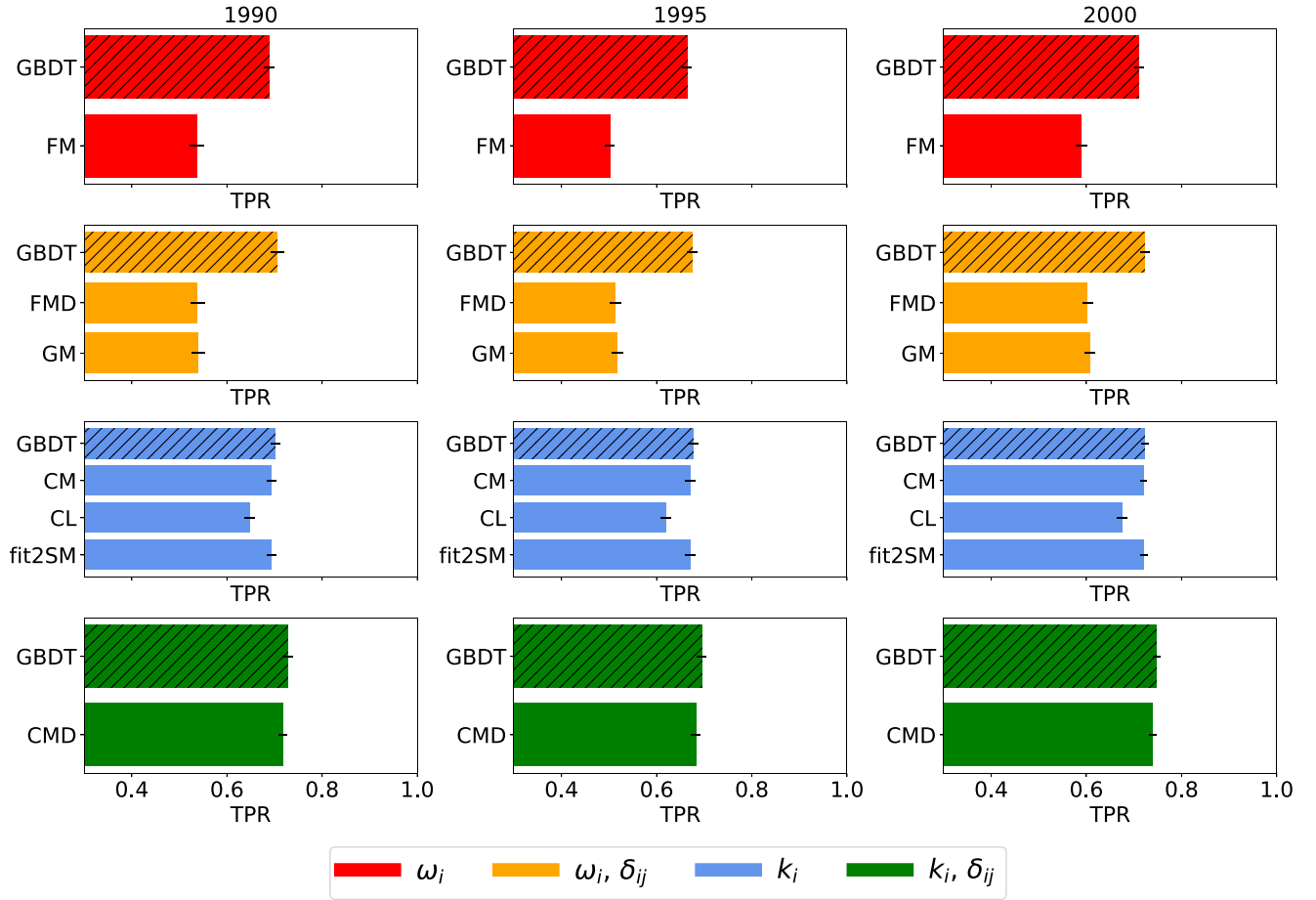


FIG. 2. Performance of the models, measured in terms of TPR, for the years 1990, 1995, and 2000 of the WTW. Each panel collects methods that rely on the same set of features (see the color legend). The upper bar in each panel, marked with a diagonal pattern, corresponds to the black-box GBDT. We have randomly selected 10% of links 40 times to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$; each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a horizontal black bar. When considering “exogenous” features, each instance of the GBDT outperforms its white-box counterpart. The two classes of models, instead, perform in a comparable way when coming to consider “endogenous” features such as the degrees.

$L(\mathbf{A}^{\text{obs}}) = |\mathcal{E}^{\text{obs}}|$, i.e., that

$$L(\mathbf{A}^{\text{obs}}) = \sum_{i=1}^N \sum_{j(>i)} p_{ij}^{\text{FM}}, \quad (35)$$

a task that can be accomplished by implementing the iterative recipe

$$z_{t+1} = \frac{L(\mathbf{A}^{\text{obs}})}{\sum_{i=1}^N \sum_{j(>i)} \left(\frac{\omega_i \omega_j}{1 + z_t \omega_i \omega_j} \right)}. \quad (36)$$

A different position reads $x_i = \sqrt{z} s_i$, where now $s_i = \sum_{j(\neq i)} w_{ij}$ indicates the strength of node i . This second choice is a popular one when analyzing financial networks, the available data about which usually concern lending and borrowing relationships. This second specification is also known with the name of density-corrected gravity model [49,50].

Fitness model with distances. As for the CM, also the CMD can be turned into a fitness model with distances (FMD). By operating the same replacement as above, one gets

$$p_{ij}^{\text{FMD}} = \frac{z \omega_i \omega_j w^{d_{ij}}}{1 + z \omega_i \omega_j w^{d_{ij}}}, \quad (37)$$

the system to be solved, now, reading

$$L(\mathbf{A}^{\text{obs}}) = \sum_{i=1}^N \sum_{j(>i)} p_{ij}^{\text{FMD}}, \quad (38)$$

$$d(\mathbf{A}^{\text{obs}}) = \sum_{i=1}^N \sum_{j(>i)} p_{ij}^{\text{FMD}} d_{ij},$$

and the iterative recipe becoming

$$z_{t+1} = \frac{L(\mathbf{A}^{\text{obs}})}{\sum_{i=1}^N \sum_{j(>i)} \left(\frac{\omega_i \omega_j w^{d_{ij}}}{1 + z_t \omega_i \omega_j w^{d_{ij}}} \right)}, \quad (39)$$

$$w_{t+1} = \frac{d(\mathbf{A}^{\text{obs}})}{\sum_{i=1}^N \sum_{j(>i)} \left(\frac{z_t \omega_i \omega_j w^{d_{ij}-1}}{1 + z_t \omega_i \omega_j w^{d_{ij}}} \right) d_{ij}}.$$

To ease numerical manipulations, the distances have been normalized, thus reading $\delta_{ij} = d_{ij}/\bar{d}$, where $\bar{d} = 2 \sum_{i=1}^N \sum_{j(>i)} d_{ij} / N(N-1)$.

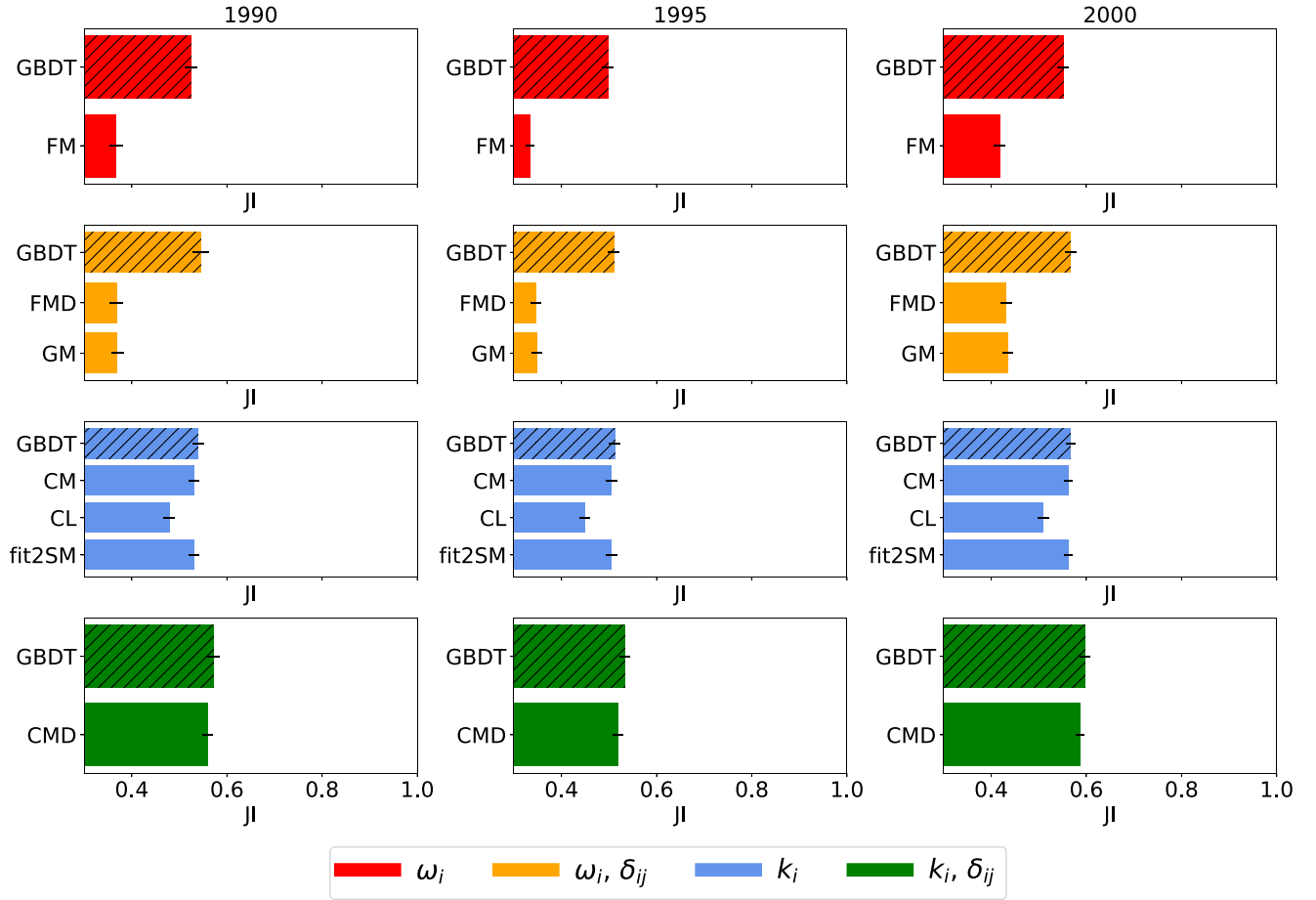


FIG. 3. Performance of the models, measured in terms of JI, for the years 1990, 1995, and 2000 of the WTW. Each panel collects methods that rely on the same set of features (see the color legend). The upper bar in each panel, marked with a diagonal pattern, corresponds to the black-box GBDT. We have randomly selected 10% of links 40 times to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$; each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a horizontal black bar. When considering “exogenous” features, each instance of the GBDT outperforms its white-box counterpart. The two classes of models, instead, perform in a comparable way when coming to consider “endogenous” features such as the degrees.

2. Link prediction via nonlinear ERGs

So far, we have focused on linear ERGs; let us now move to considering nonlinear ERGs.

Fitness-induced two-star model. The simplest model of this kind is the so-called two-star model (2SM) [30,51,52], induced by posing $\mathbf{C} = \{L, S\}$, with $L = \sum_{i=1}^N \sum_{j(>i)} a_{ij} = \sum_{i=1}^N k_i/2$ and $S = \sum_{i=1}^N \sum_{j(>i)} \sum_{m(\neq i, j)} a_{im} a_{mj} = \sum_{i=1}^N k_i^2/2 - L$. Such an identification leads to $H_{2\text{SM}}(\mathbf{A}) = \sum_{i=1}^N \sum_{j(>i)} a_{ij} \times [\alpha + \beta(k_i + k_j)]$, an expression inducing

$$p_{ij}^{2\text{SM}} = \frac{xy^{\langle k_i \rangle + \langle k_j \rangle}}{1 + xy^{\langle k_i \rangle + \langle k_j \rangle}}, \quad (40)$$

where $x = e^{-\alpha}$ and $y = e^{-\beta}$. One would be tempted to identify the expected value of the degrees with the observed one, but consistency forbids such a match: As the only available information concerns two global quantities, degrees are not directly accessible. In order to make the model fully consistent, a more refined formulation is thus needed.

First, the model we are looking for should be heterogeneous, thus ensuring that node-specific differences are

preserved. More quantitatively, one would be looking for an expression like

$$p_{ij}^{\text{dc2SM}} = \frac{x_i x_j y^{k_i + k_j}}{1 + x_i x_j y^{k_i + k_j}}, \quad (41)$$

where k_i and k_j must be interpreted as the degrees of the nodes evaluated on $\mathbf{A}^{\text{obs}}$, i.e., $k_i(\mathbf{A}^{\text{obs}})$ and $k_j(\mathbf{A}^{\text{obs}})$; as proven in Ref. [42], however, the model above, named degree-corrected two-star model (dc2SM), suffers from limitations inherited by the mean-field approach inducing it. One is, thus, forced to consider a fitness-based variant of it, reading either

$$p_{ij}^{\text{fit2SM-I}} = \frac{z\omega_i \omega_j y^{\langle k_i \rangle + \langle k_j \rangle}}{1 + z\omega_i \omega_j y^{\langle k_i \rangle + \langle k_j \rangle}}, \quad (42)$$

where $\langle k_i \rangle = \sum_{j(\neq i)} p_{ij}^{\text{fit2SM-I}}$, or

$$p_{ij}^{\text{fit2SM-II}} = \frac{zs_i s_j y^{\langle k_i \rangle + \langle k_j \rangle}}{1 + zs_i s_j y^{\langle k_i \rangle + \langle k_j \rangle}}, \quad (43)$$

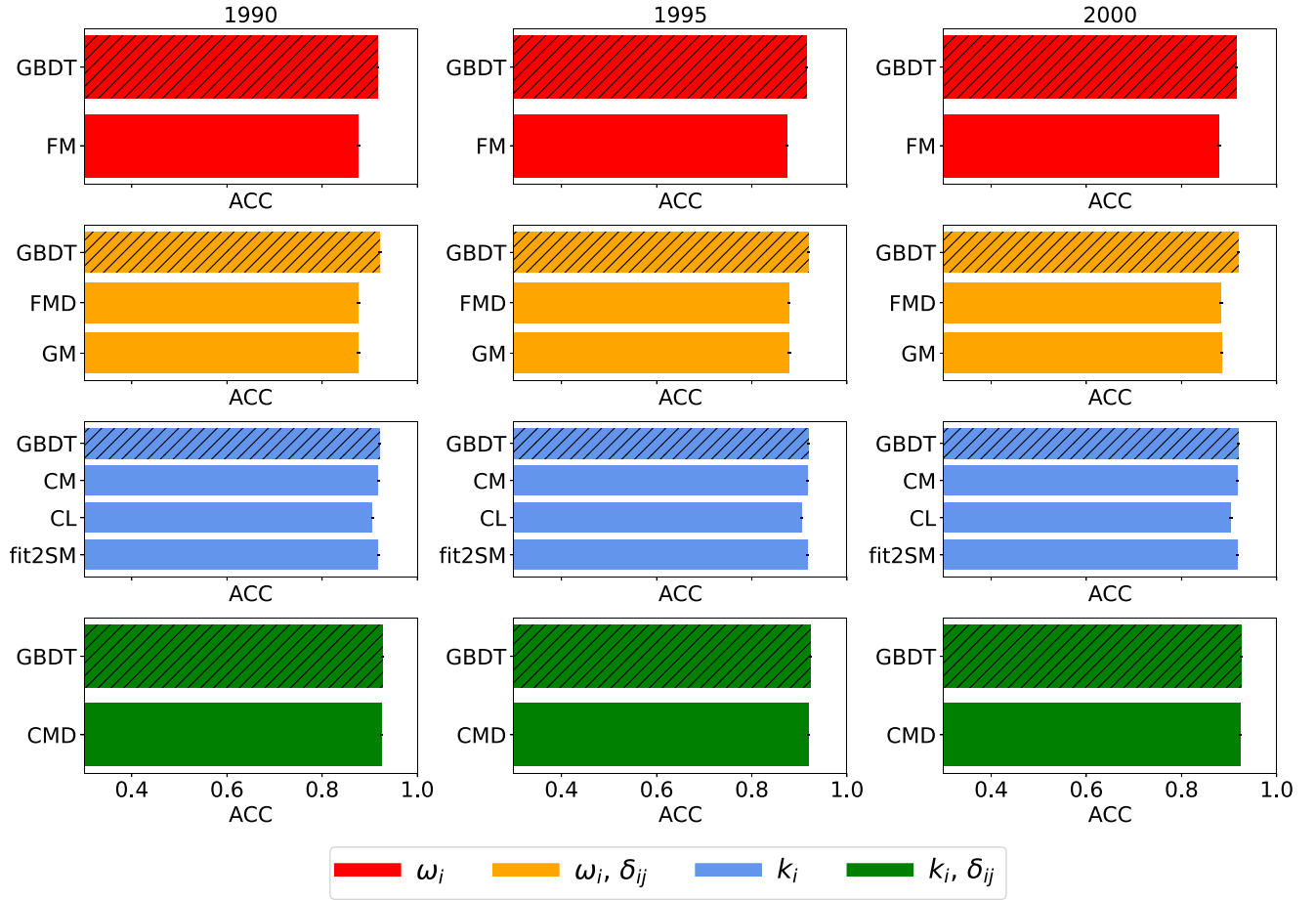


FIG. 4. Performance of the models, measured in terms of ACC, for the years 1990, 1995, and 2000 of the WTW. Each panel collects methods that rely on the same set of features (see the color legend). The upper bar in each panel, marked with a diagonal pattern, corresponds to the black-box GBDT. We have randomly selected 10% of links 40 times to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$; each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a horizontal black bar. When considering “exogenous” features, each instance of the GBDT outperforms its white-box counterpart. The two classes of models, instead, perform in a comparable way when coming to consider “endogenous” features such as the degrees.

where $\langle k_i \rangle = \sum_{j(\neq i)} p_{ij}^{\text{fit2SM-II}}$. A third variant can be defined as

$$p_{ij}^{\text{fit2SM-III}} = \frac{z \tilde{x}_i \tilde{x}_j y^{(k_i) + \langle k_j \rangle}}{1 + z \tilde{x}_i \tilde{x}_j y^{(k_i) + \langle k_j \rangle}}, \quad (44)$$

where $\langle k_i \rangle = \sum_{j(\neq i)} \tilde{x}_i \tilde{x}_j / (1 + \tilde{x}_i \tilde{x}_j) = k_i$; the major difference with respect to the first recipe is the disentangled resolution of the formula above: In other words, one first determines the parameters $\{\tilde{x}_i\}_{i=1}^N$ by solving the CM and then inserts them above to determine z and y .

The difference with respect to the 2SM is evident: The three variants above are heterogeneous, hence likely reproducing the observed degrees better than the 2SM. The parameters z and y can be determined by solving the system of nonlinear, coupled equations

$$L(\mathbf{A}^{\text{obs}}) = \sum_{i=1}^N \sum_{j(>i)} p_{ij}^{\text{fit2SM}},$$

$$S(\mathbf{A}^{\text{obs}}) = \sum_{i=1}^N \sum_{j(>i)} \sum_{m(\neq i, j)} p_{im}^{\text{fit2SM}} p_{mj}^{\text{fit2SM}}, \quad (45)$$

where we have dropped the reference to the model specification as the recipe is the same in all cases.

Let us now face the problem of determining the expected degrees; to this aim, let us focus on the third recipe: As the parameters $\{\tilde{x}_i\}_{i=1}^N$ have been determined by solving the CM on $\mathbf{A}^{\text{obs}}$, we can insert the observed degrees directly into the recipe above. As a consequence, the two parameters z and y can now be obtained by implementing the iterative recipe

$$z_{t+1} = \frac{L(\mathbf{A}^{\text{obs}})}{\sum_{i=1}^N \sum_{j(>i)} \frac{\tilde{x}_i \tilde{x}_j y_t^{k_i + k_j}}{1 + z_t \tilde{x}_i \tilde{x}_j y_t^{k_i + k_j}}},$$

$$y_{t+1} = \frac{S(\mathbf{A}^{\text{obs}})}{\sum_{i=1}^N \sum_{j(>i)} \sum_{m(\neq i, j)} \frac{z_t^2 \tilde{x}_i \tilde{x}_j \tilde{x}_m^2 y_t^{k_i + 2k_m + k_j - 1}}{(1 + z_t \tilde{x}_i \tilde{x}_m y_t^{k_i + k_m})(1 + z_t \tilde{x}_m \tilde{x}_j y_t^{k_m + k_j})}}, \quad (46)$$

where k_i , k_m , and k_j must be interpreted as the degrees of the nodes evaluated on the accessible portion of the dataset $\mathbf{A}^{\text{obs}}$, i.e., $k_i(\mathbf{A}^{\text{obs}})$, $k_m(\mathbf{A}^{\text{obs}})$, and $k_j(\mathbf{A}^{\text{obs}})$.

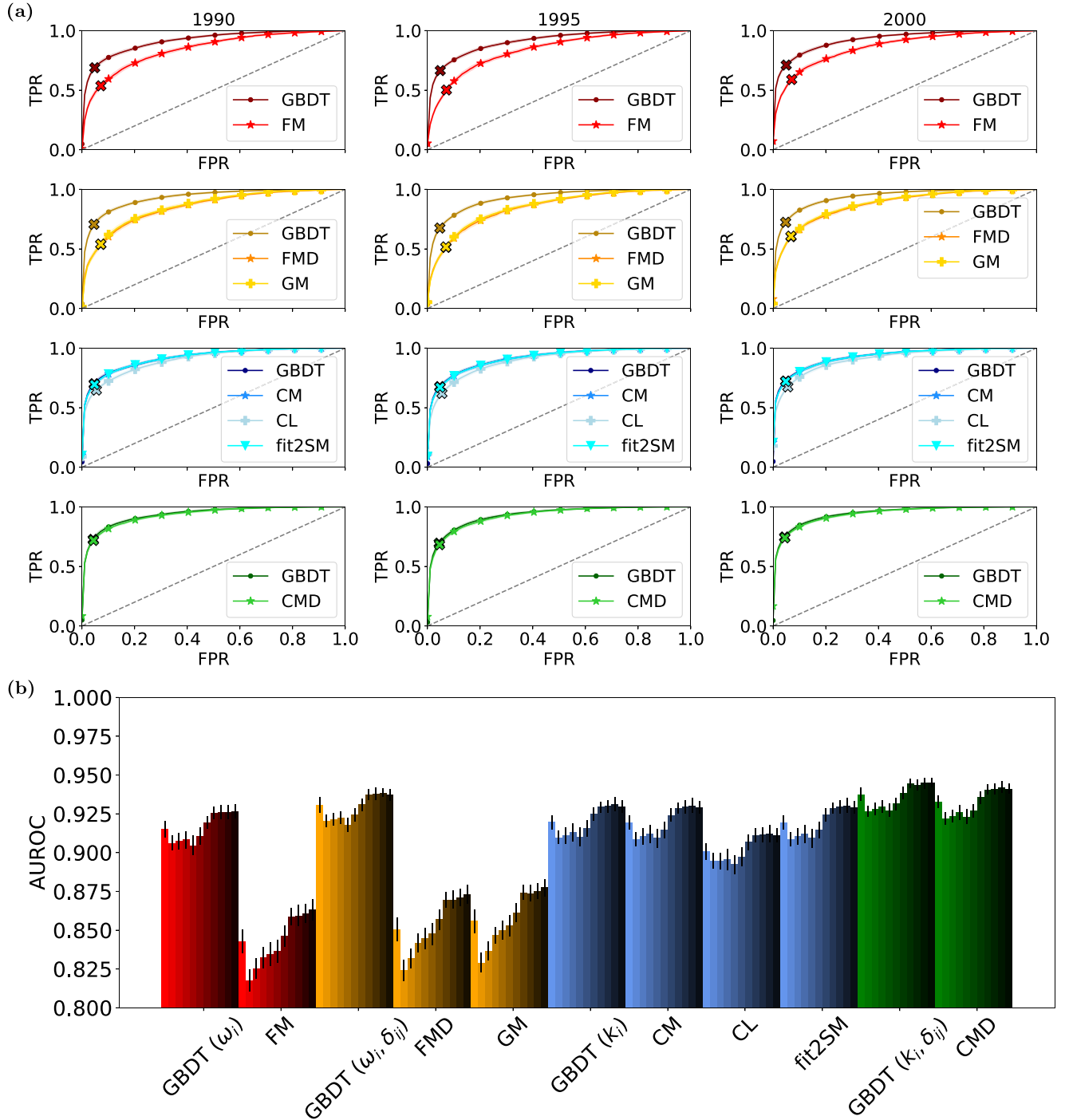


FIG. 5. (a) Performance of the models, measured in terms of ROC curves, for the years 1990, 1995, and 2000 of the WTW. Each panel collects methods that rely on the same set of features. We have randomly selected 10% of links 40 times to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$ (the curves relative to each indicator are, in fact, 40 partially overlapping curves corresponding to each realization). The “X” markers indicate the TPR and FPR values obtained by selecting a number of missing links equal to $|\mathcal{E}^{\text{miss}}|$. (b) Each statistical indicator has been averaged over the 40 realizations, the standard deviation being represented by a vertical black bar. The AUROC, depicted for all years (darker shades correspond to more recent years), enlarges when moving from less structured models, like the GM, to more structured ones, like the CMD.

V. DATA DESCRIPTION

The aforementioned link prediction algorithms have been run and compared on a number of instances of economic and financial datasets.

As a first dataset, we have considered the WTW curated by Gleditsch [48] and including yearly, bilateral, aggregated data on trade [the generic entry $\exp_{ij}(y)$ is the sum of the single, commodity-specific exports from i to j , during the year

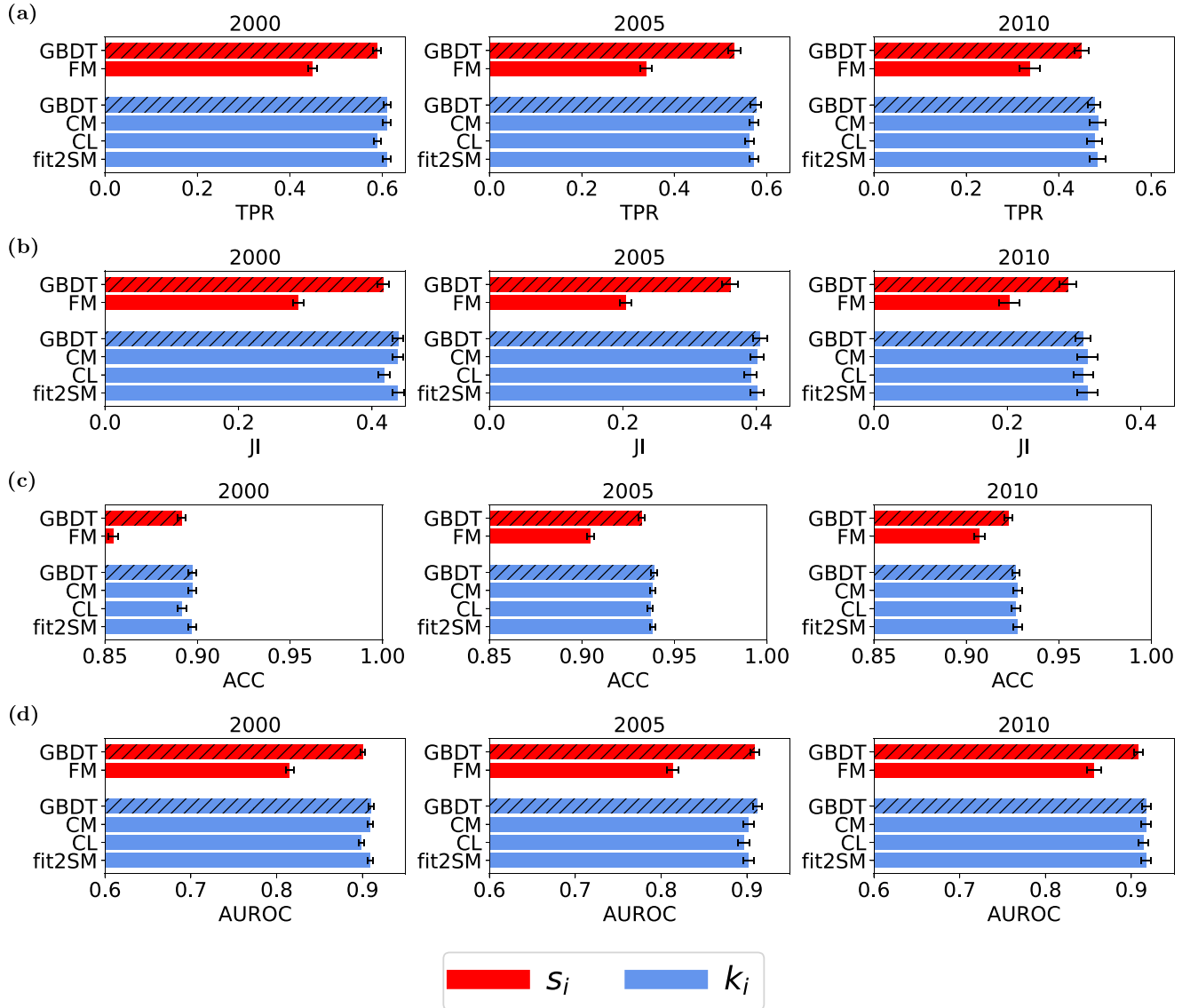


FIG. 6. Performance of the models, measured in terms of TPR [row (a)], JI [row (b)], ACC [row (c)], and AUROC [row (d)], for the years 2000, 2005, and 2010 of eMID. The upper bar in each panel, marked with a diagonal pattern, corresponds to the black-box GBDT. We have randomly selected 20% of links 30 times to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$; each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a horizontal black bar. The FM has been implemented by replacing ω_i with s_i in Eq. (34) (see the color legend). As for the WTW, each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterparts—especially for what concerns the CM and the fit2SM; overall, however, the higher sparsity of the system under consideration leads to (much) lower TPR and JI scores than those obtained when analyzing the WTW.

y , reported in millions of U.S. dollars], yearly GDP values (reported in millions of U.S. dollars), and the (time-independent) matrix of geographic distances between the capital cities of the countries in the data. We have then symmetrized and binarized the trade matrix by posing $a_{ij} = \Theta[w_{ij}]$, where Θ is the Heaviside step function and $w_{ij} = (\exp_{ij} + \exp_{ji})/2$ —in words, w_{ij} is the bilateral trade volume defined as the arithmetic mean of the export volume from country i to country j and the export volume from country j to country i —for each of the 11 years going from 1990 to 2000.

As a second dataset, we have considered the eMID [42], represented as a weighted, directed network whose nodes are

anonymized banks and weights represent exposures in millions of euros. Reported data cover the period January 1999 to December 2014 on a daily frequency: The presence of a weight w_{ij} at time t indicates the existence of a total exposure $w_{ij} \geq 50.000$ euros, registered at the end of the particular period t and directed from bank i to bank j . Considering that $\simeq 98\%$ of banks are Italian and that the volume of their transactions covers $\simeq 85\%$ of the total volume (as of 2011), our analysis solely focuses on the subgraph induced by such a subset of nodes. We have, then, symmetrized and binarized the exposure matrix by posing $a_{ij} = \Theta[w_{ij} + w_{ji}]$ —in words, $w_{ij} + w_{ji}$ is the sum of the exposures of the involved

TABLE III. Basic statistics concerning the eMID dataset [42], aggregated at the yearly level. The table shows the total number of banks N , the total number of links L , the network density c , and the average degree $\bar{k}$.

Year	N	L	c	$\bar{k}$
1999	215	9770	0.424	90.88
2000	195	8172	0.432	83.81
2001	182	6669	0.401	73.29
2002	176	5528	0.359	62.82
2003	176	4892	0.318	55.59
2004	175	4327	0.284	49.45
2005	172	4119	0.280	47.90
2006	172	3910	0.266	45.47
2007	170	3886	0.270	45.72
2008	157	3161	0.258	40.27
2009	137	2381	0.256	34.76
2010	127	2192	0.274	34.52
2011	112	2036	0.328	36.36
2012	97	1421	0.306	29.30

banks—for each aggregation period ranging from daily to yearly.

VI. RESULTS

As in a real-case scenario the choice between different methods would be determined by the available features, it is useful to compare models informed by the same quantities. To this aim, let us recall that (1) the GM, defined by Eq. (11), takes as inputs the GDPs and the geographic distances; (2)

TABLE IV. Basic statistics concerning the eMID dataset [42], aggregated at the quarterly, monthly, weekly, and daily levels. The table shows the total number of banks N , the total number of links L , the network density c , and the average degree $\bar{k}$ for each aggregation scale.

Quarter	N	L	c	$\bar{k}$
2000–Q01	181	5483	0.337	60.59
2000–Q03	178	5205	0.330	58.48
2010–Q01	113	1433	0.226	25.36
2010–Q03	107	1476	0.260	27.59
Month	N	L	c	$\bar{k}$
2000–M01	177	3634	0.230	41.06
2000–M08	175	3403	0.224	38.89
2010–M01	102	1014	0.197	19.88
2010–M08	97	920	0.198	18.97
Week	N	L	c	$\bar{k}$
2000–W01	170	1428	0.099	16.80
2000–W30	167	1493	0.108	17.88
2010–W01	82	346	0.104	8.44
2010–W30	90	423	0.106	9.40
Day	N	L	c	$\bar{k}$
2000-01-03	149	523	0.047	7.02
2005-04-15	108	320	0.055	5.93
2010-08-30	72	166	0.065	4.61

the first formulation of the FM, defined by Eq. (34), takes as inputs the GDPs; (3) the second formulation of the FM, defined by Eq. (37), takes as inputs the GDPs and the geographic distances; (4) the CM and the CL, respectively, defined by Eqs. (27) and (30), take as inputs the degrees; (5) the CMD, defined by Eq. (31), takes as inputs the degrees and the geographic distances; and (6) the fit2SM, defined by Eq. (44), takes as inputs the degrees (as the total number of links and the total number of two stars can be deduced from them). Naturally, we have compared the performance of each white-box method with the instance of the GBDT taking as input the same amount of information.

Let us start commenting on the WTW. The basic statistics are reported in Table II: As it can be appreciated, the number of nodes N ranges between 159 and 177 and the total number of links L ranges between 7639 and 9866; still, the link density $c = 2L/N(N-1)$ does not fluctuate much, ranging between 0.55 and 0.63, as well as the average degree $\bar{k} = 2L/N$, ranging between $\gtrsim 95$ and $\gtrsim 111$. Figure 2 depicts the TPR for each of the aforementioned models in the years 1990, 1995, and 2000. Let us recall that it is defined as $\text{TPR} = |\mathcal{E}^{\text{miss}} \cap \mathcal{E}^{\text{miss}}|/|\mathcal{E}^{\text{miss}}|$, hence accounting for the percentage of missing links that are correctly identified as such. First, let us notice that when considering “exogenous” features, each instance of the GBDT outperforms its white-box counterpart. The two classes of models, instead, perform in a comparable way when coming to consider “endogenous” features such as the degrees and the number of two stars: Quite interestingly, the CL is the method performing worst, while the CM and the fit2SM perform in a very similar way—a result indicating that the present application does not necessarily require to consider nonlinear ERGs.

Figure 3 depicts $\text{JI} = |\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{miss}}|/|\overline{\mathcal{E}^{\text{miss}}} \cup \mathcal{E}^{\text{miss}}|$, hence enriching the picture provided by the TPR by accounting for the number of false positives too; Fig. 4 depicts $\text{ACC} = (\text{TP} + \text{TN})/|\mathcal{E}^{\text{no}}|$ for each of the aforementioned models, hence accounting for the percentage of correctly identified links, be they missing or nonexistent: Both return a picture that is consistent with the one returned by the TPR.

A complementary picture is provided by Fig. 5, showing the ROC curves, obtained upon scattering the $\text{TPR} = |\mathcal{E}^{\text{miss}} \cap \mathcal{E}^{\text{miss}}|/|\mathcal{E}^{\text{miss}}|$ versus the $\text{FPR} = |\overline{\mathcal{E}^{\text{miss}}} \cap \mathcal{E}^{\text{ne}}|/|\mathcal{E}^{\text{ne}}|$ as the list of links ranked in decreasing order of the chosen score is gone through, for each of the aforementioned models, in the years 1990, 1995, and 2000. As evident upon looking at the picture, the AUROC of white-box models enlarges when moving from less structured models like the GM to more structured ones like the CMD.

All the aforementioned exercises have been carried out by randomly selecting 10% of links 40 times to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$ —each statistical indicator has, then, been averaged over such samples: The results we have obtained for a different percentage of randomly selected links, as well as for a different indicator, namely, AUPRC, are shown in Appendix A. As it can be appreciated, the performance of the CMD remains high up to a large percentage of removed connections—equivalently, of missing links; still, the performance of the GBDT deteriorates to a lesser extent than that of white-box models.

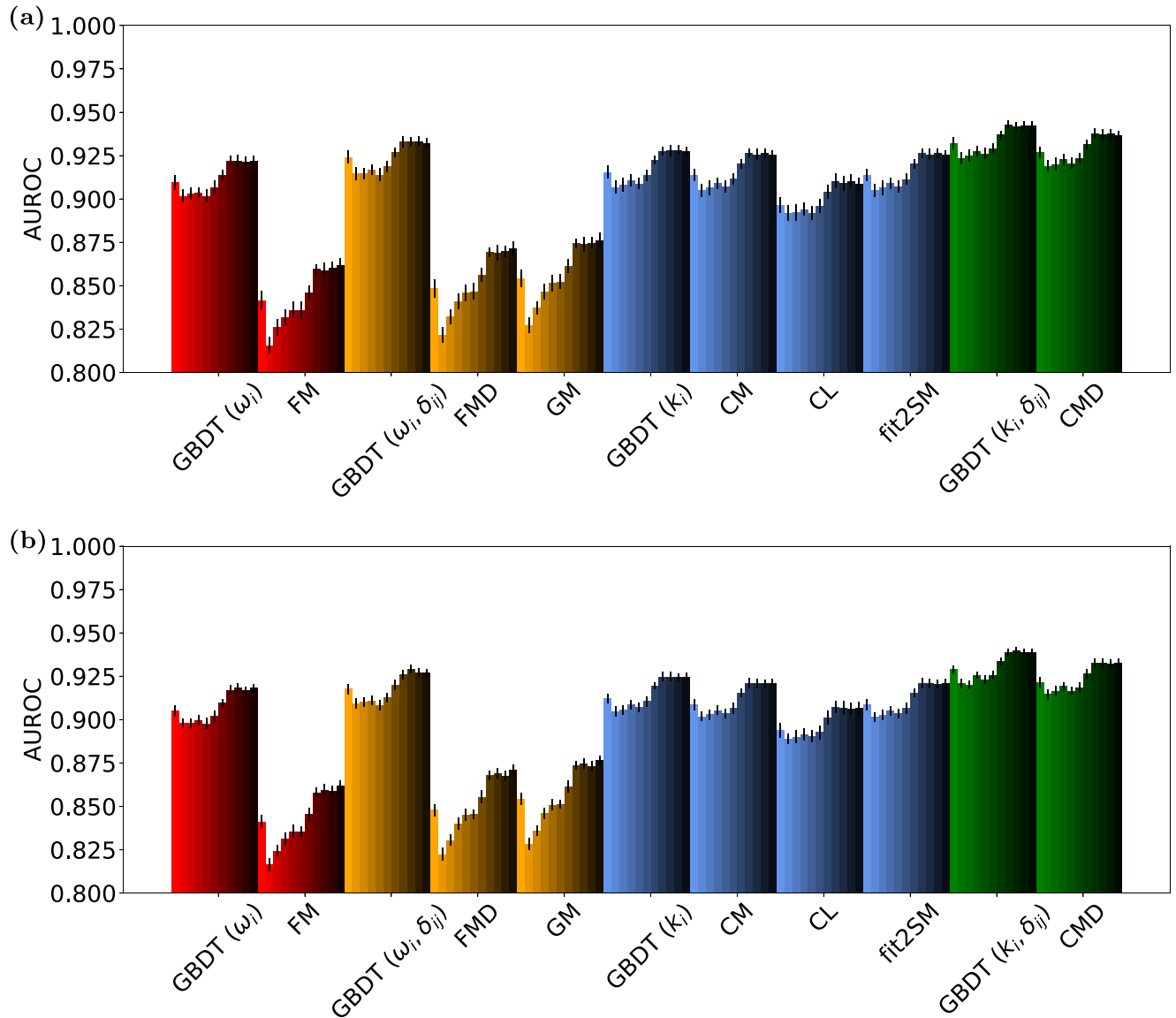


FIG. 7. Performance of the models, measured in terms of AUROC, for all years of the WTW (darker shades correspond to more recent years). We have randomly selected 20% [panel (a)] and 30% [panel (b)] of links 40 times to populate the test set, on which the prediction is performed after having trained the model on the remaining pairs of nodes. Each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a vertical black bar. The results are in agreement with those reported in the main text, as each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterparts—especially for what concerns the CM and the fit2SM.

Let us now move to commenting on eMID, aggregated at the yearly level as well. The basic statistics are reported in Table III: As it can be appreciated, the number of nodes N ranges between 97 and 215, and the total number of links L ranges between 1421 and 9770; the link density $c = 2L/N(N-1)$ fluctuates appreciably between 0.26 and 0.43, as well as the average degree $k = 2L/N$, ranging between $\gtrsim 29$ and $\lesssim 91$. As, in this case, no external feature is available *a priori*, we have excluded comparisons of the kind.

All results are shown in Fig. 6. Overall, the larger sparsity of the system under consideration leads to TPR and JI scores that are lower than the ones obtained when analyzing the WTW; in the case of eMID, in fact, reproducing the 1s is (much) more difficult than reproducing the 0s: As a conse-

quence, a (much) worse performance on the number of true positives than on the number of true negatives is expected. As for the WTW, however, each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterpart, especially for what concerns the CM and the fit2SM—although the sparsity of eMID lets the CL perform as well as the CM and the fit2SM.

All the aforementioned exercises on eMID have been carried out by randomly selecting 20% of links 30 times to populate $\mathcal{E}^{\text{miss}}$ and generate $\mathcal{E}^{\text{obs}} = \mathcal{E} \setminus \mathcal{E}^{\text{miss}}$ —each statistical indicator has, then, been averaged over such samples: The results we have obtained for a different aggregation level, as well as for a different indicator, namely, AUPRC, are shown

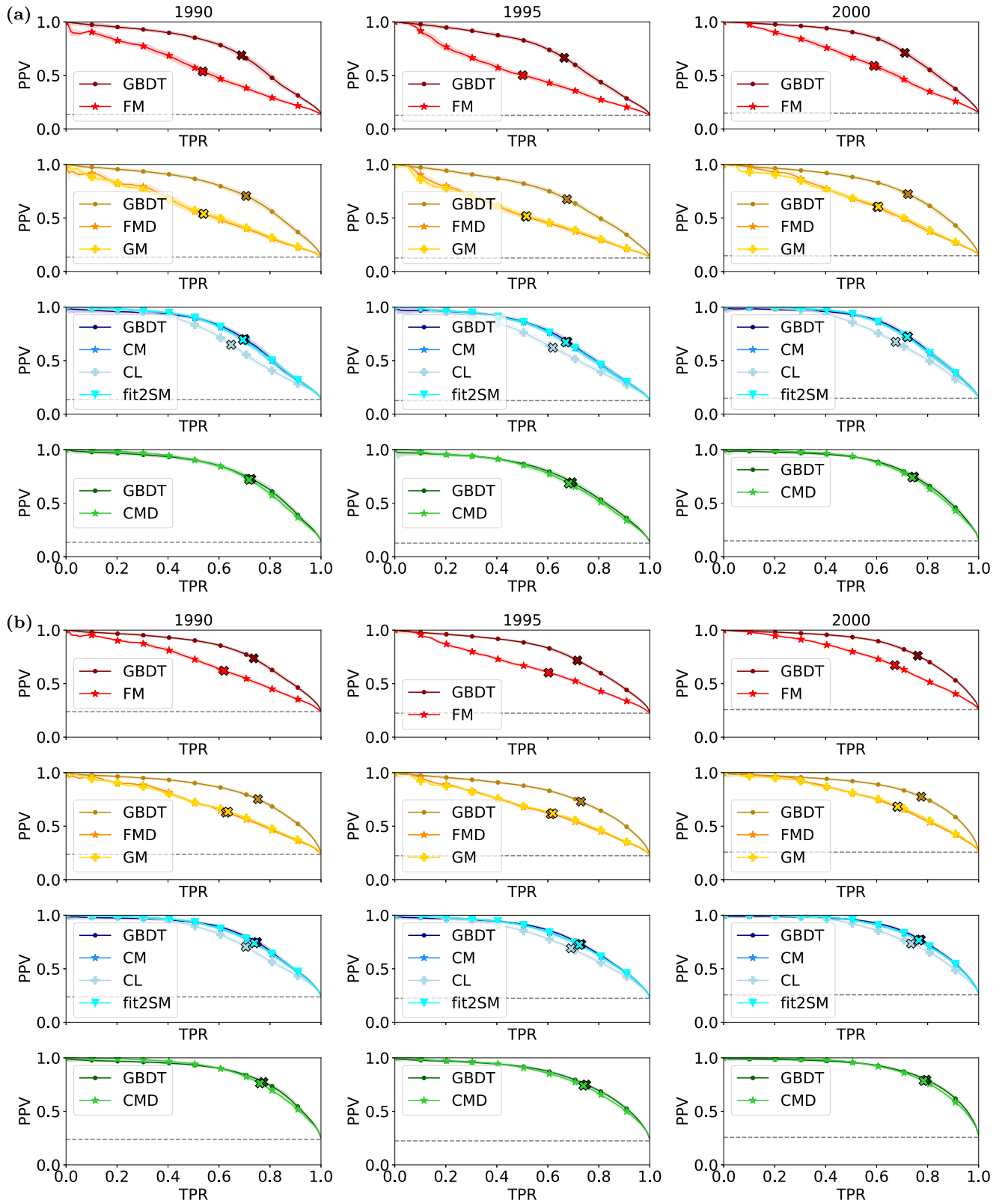


FIG. 8. Performance of the models, measured in terms of PRC curves, for the years 1990, 1995, and 2000 of the WTW. Each panel collects methods that rely on the same set of features. We have randomly selected 10% [panel (a)] and 20% [panel (b)] of links 40 times to populate the test set (the curves relative to each indicator are, in fact, 40 partially overlapping curves corresponding to each realization), on which the prediction is performed after having trained the model on the remaining pairs of nodes. The “X” markers indicate the PPV and TPR values obtained by selecting a number of missing links equal to $|\mathcal{E}^{\text{miss}}|$. The AUPRC enlarges when moving from less structured models, like the GM, to more structured ones, like the CMD.

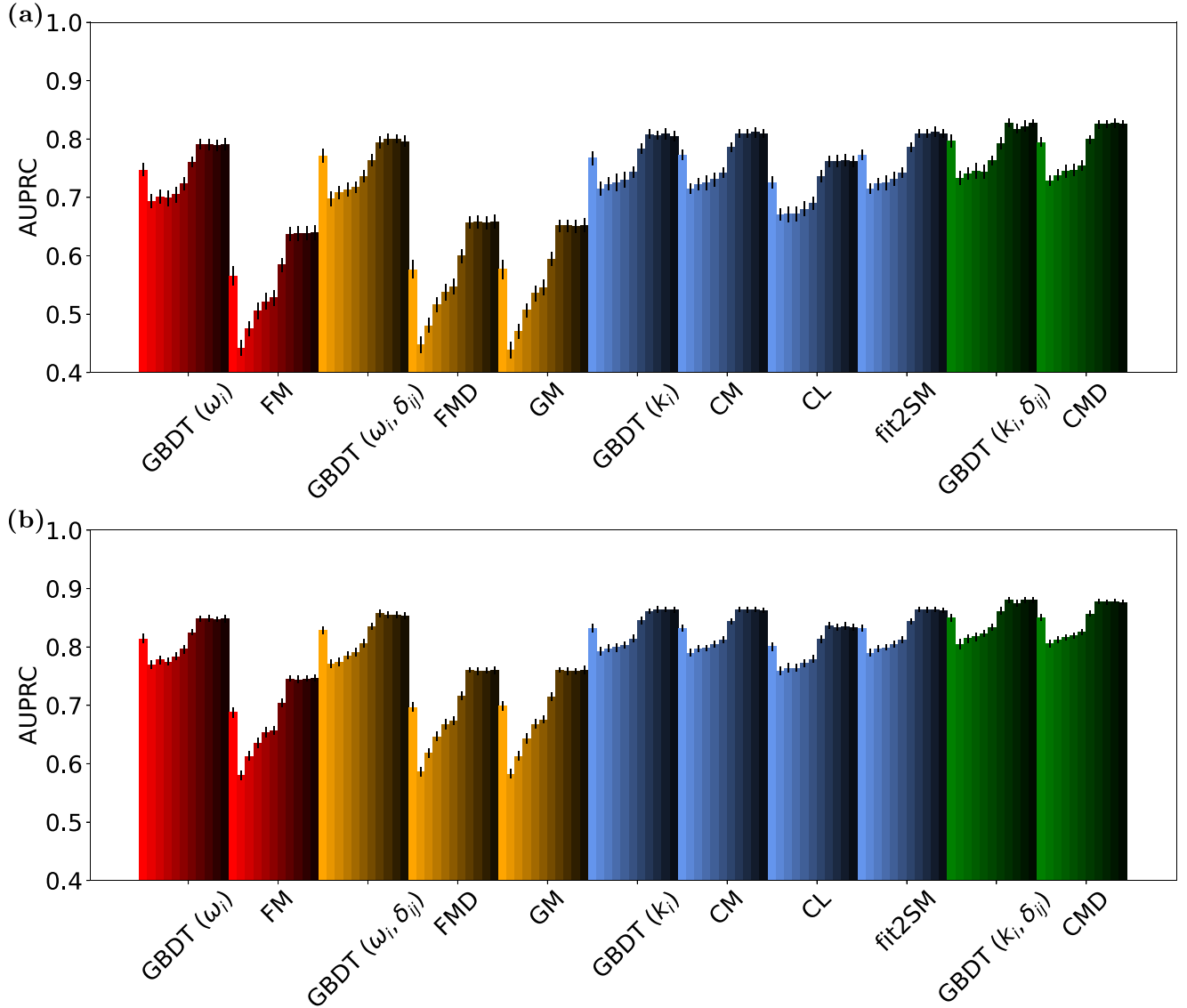


FIG. 9. Performance of the models, measured in terms of AUPRC, for all years of the WTW (darker shades correspond to more recent years). We have randomly selected 10% [panel (a)] and 20% [panel (b)] of links 40 times to populate the test set, on which the prediction is performed after having trained the model on the remaining pairs of nodes. Each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a vertical black bar. The results are in agreement with those reported in the main text, as each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterparts—especially for what concerns the CM and the fit2SM.

in Appendix B. As can be appreciated, very similar results to those characterizing the yearly aggregation level are obtained, with the only exception of the AUPRC, which appears to be sensitive to the configuration density.

VII. DISCUSSION

The aim of the present contribution is that of comparing the performance of several algorithms in carrying out a link prediction exercise, one of the selection criteria concerning data requirements: The algorithms implemented here have been instantiated with the same set of either “endogenous” or “exogenous” features—in other words, have been compared on the same ground.

In order to do so, we have posed ourselves in the same setting of Refs. [2,6], where the variables of interest are computed on the observed portion of the graph (i.e., $\mathbf{A}^{\text{obs}}$) and the prediction is carried out on the pairs of nodes in $\mathcal{E}^{\text{no}}$, which correspond to the entries $a_{ij}^{\text{obs}} = 0, \forall ij \in \mathcal{E}^{\text{no}}$, hence appearing as not connected. Within such a framework, the best performance is obtained when the purely structural information is combined with the available, external information: for what concerns the WTW, the latter consists of two sets of properties, i.e., the GDPs of countries and the geographic distances between them; while, however, the FMD does not perform better than the FM and the GM, enriching the CM with the information concerning the geographic distances leads to a better performance than the one characterizing the GBDT taking as inputs the same sets of quantities.

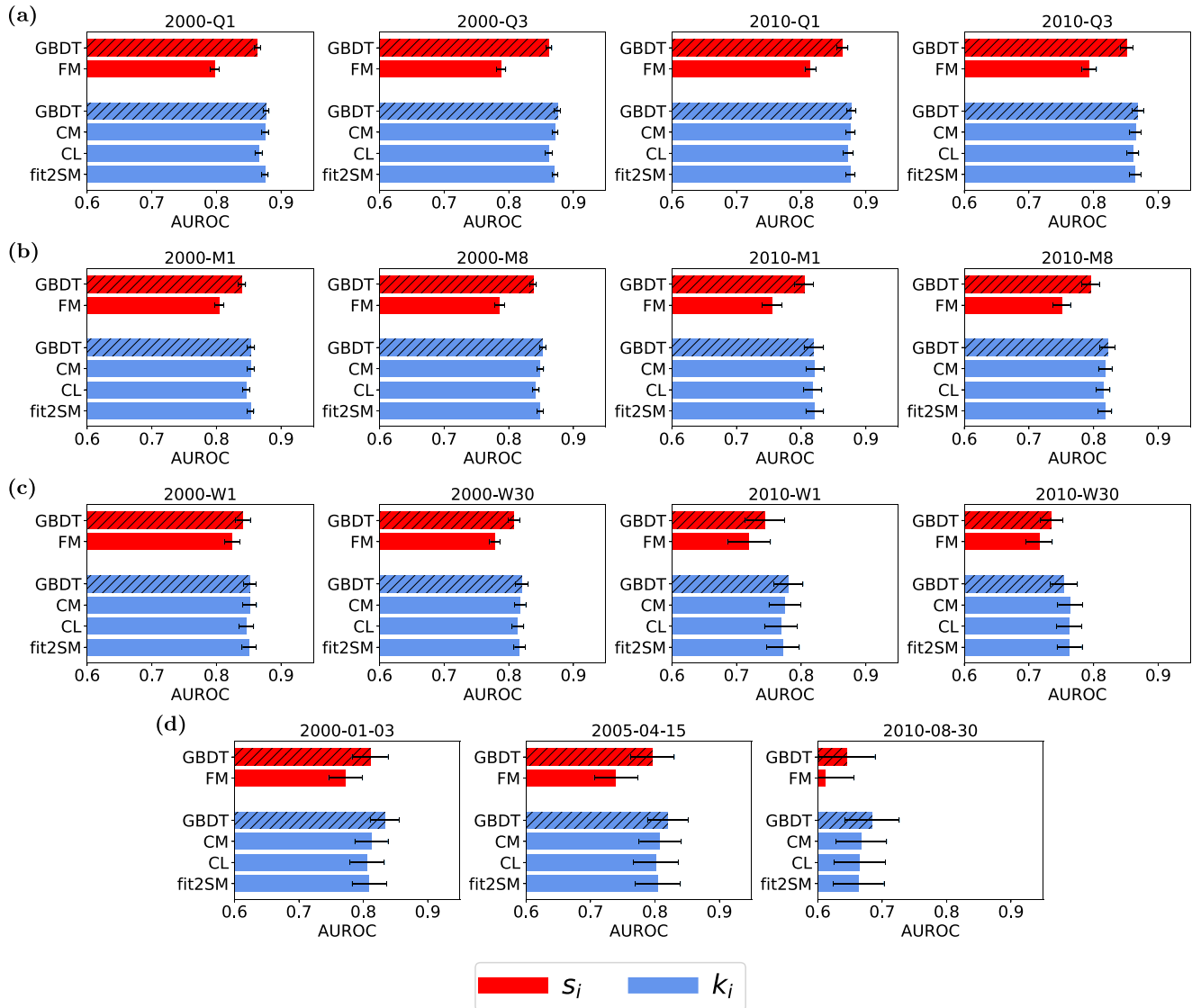


FIG. 10. Performance of the models, measured in terms of AUROC, for snapshots of eMID at different aggregation levels: quarterly [panel (a)], monthly [panel (b)], weekly [panel (c)], and daily [panel (d)]. The upper bar in each panel, marked with a diagonal pattern, corresponds to the black-box GBDT. We have randomly selected 20% of links 30 times to populate the test set. Each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a horizontal black bar. The FM has been implemented by replacing ω_i with s_i in Eq. (34) (see the color legend). As for the WTW, each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterpart; the results are overall consistent across aggregation levels.

Such a result leads us to the conclusion that the structural information plays a fundamental role in shaping a network topology: It is, in fact, so relevant to make the CM alone capable of competing with the instance of the GBDT taking as inputs the degrees; in other words, although machine learning algorithms appear as among those performing best, white-box models taking as inputs purely “endogenous” features offer a comparable performance. This is even more relevant when considering that the GBDT needs a very fine-grained kind of information, i.e., the exact position of the connections populating the training set, while the CM takes as input a much coarser one, i.e., (just) the number of connections that are incident to each node (its degree); see also below.

While the first consideration is in agreement with the results of Ref. [7], the second is not. Such a discrepancy may be

due to several reasons: (1) There, the considered systems are supply chains, represented as binary, directed networks, while, here the binary, undirected representation of the world trade and of a financial system has been considered. While these are not big differences *per se*, the chosen representation affects the way models are defined, as well as the number and type of “exogenous” features (e.g., industry, location, size) that can be employed—the FM implemented in Ref. [7] is undirected in nature, hence particularly unsuited to model directed configurations characterized by a small value of reciprocity such as production networks. (2) There, local, topological information such as the one encoded into the degree sequence(s) has been apparently ignored. Here, instead, it represents the backbone of many of the models taking as inputs “endogenous” features. (3) There, algorithms are trained on subsets of the original

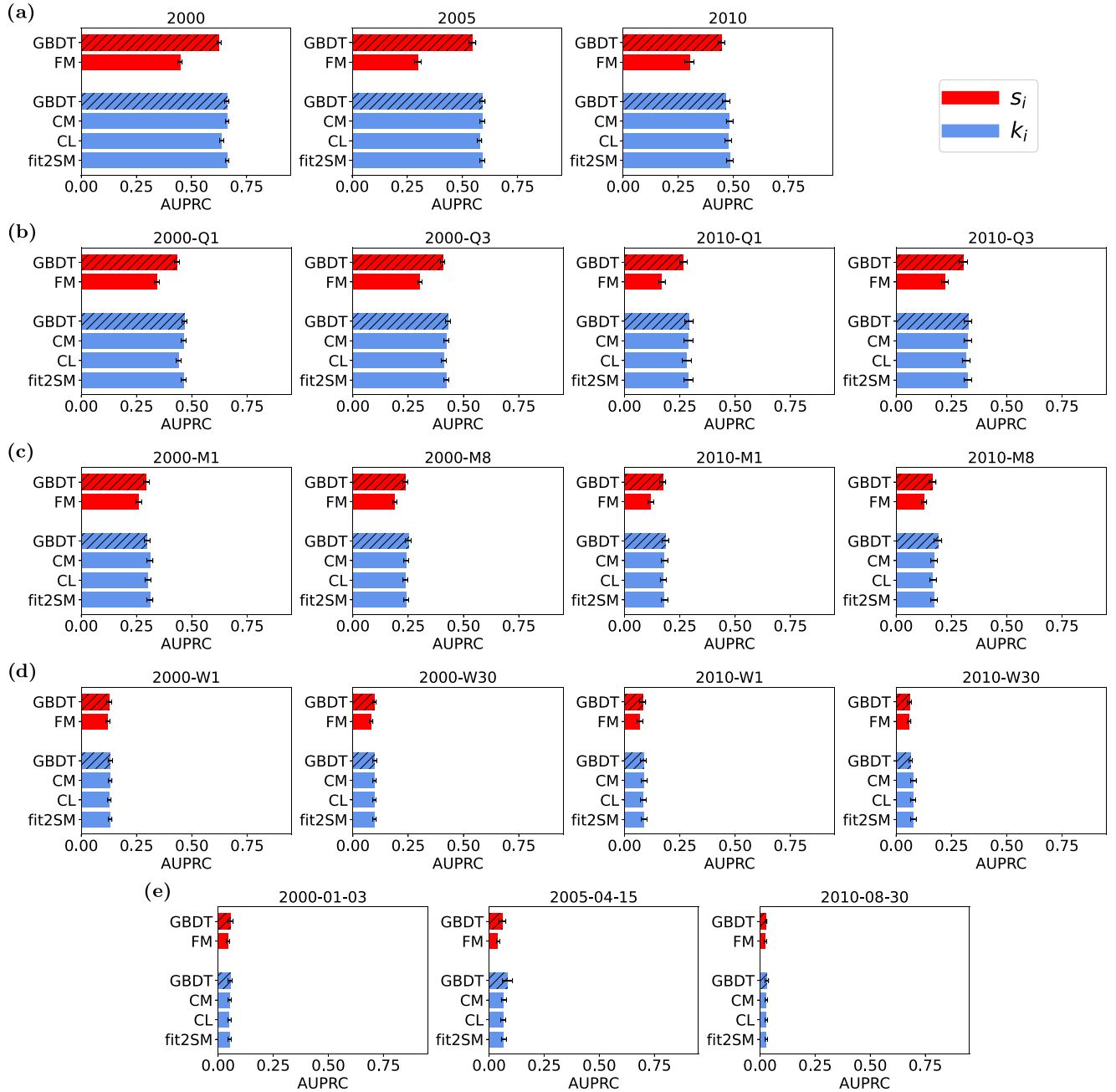


FIG. 11. Performance of the models, measured in terms of AUPRC, for snapshots of eMID at different aggregation levels: yearly [panel (a)], quarterly [panel (b)], monthly [panel (c)], weekly [panel (d)], and daily [panel (e)]. The upper bar in each panel, marked with a diagonal pattern, corresponds to the black-box GBDT. We have randomly selected 20% of links 30 times to populate the test set. Each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a horizontal black bar. The FM has been implemented by replacing ω_i with s_i in Eq. (34) (see the color legend). As for the WTW, each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterpart; the results change as finer aggregation levels are considered.

dataset, identified to contain both connected and unconnected pairs of nodes—More specifically, $\mathcal{U}$ is split into training and test sets, i.e., $\mathcal{U} = \mathcal{U}^r \cup \mathcal{U}^e$, and, then, $\mathcal{U}^r$ is sampled, the collected information consisting of one observation for each pair of firms that has been drawn, i.e., whether the two firms are connected or not, plus a vector of features. Here, instead, all algorithms are trained on the same (incomplete) portion of the dataset of interest, to let all of them start on equal footing. Still, we have carried out an additional set of comparisons,

more in line with the approach pursued in Ref. [7]; the results are shown in Appendix C: As evident upon inspecting the figures there, very similar considerations to the ones above hold true.

As we have already noticed, the usual sampling scheme “favors” well-connected nodes. Appendix D has been, now, dedicated to compare it with a couple of alternative ones: the first one prescribes to remove links by selecting edges with a probability that is inversely proportional to the product

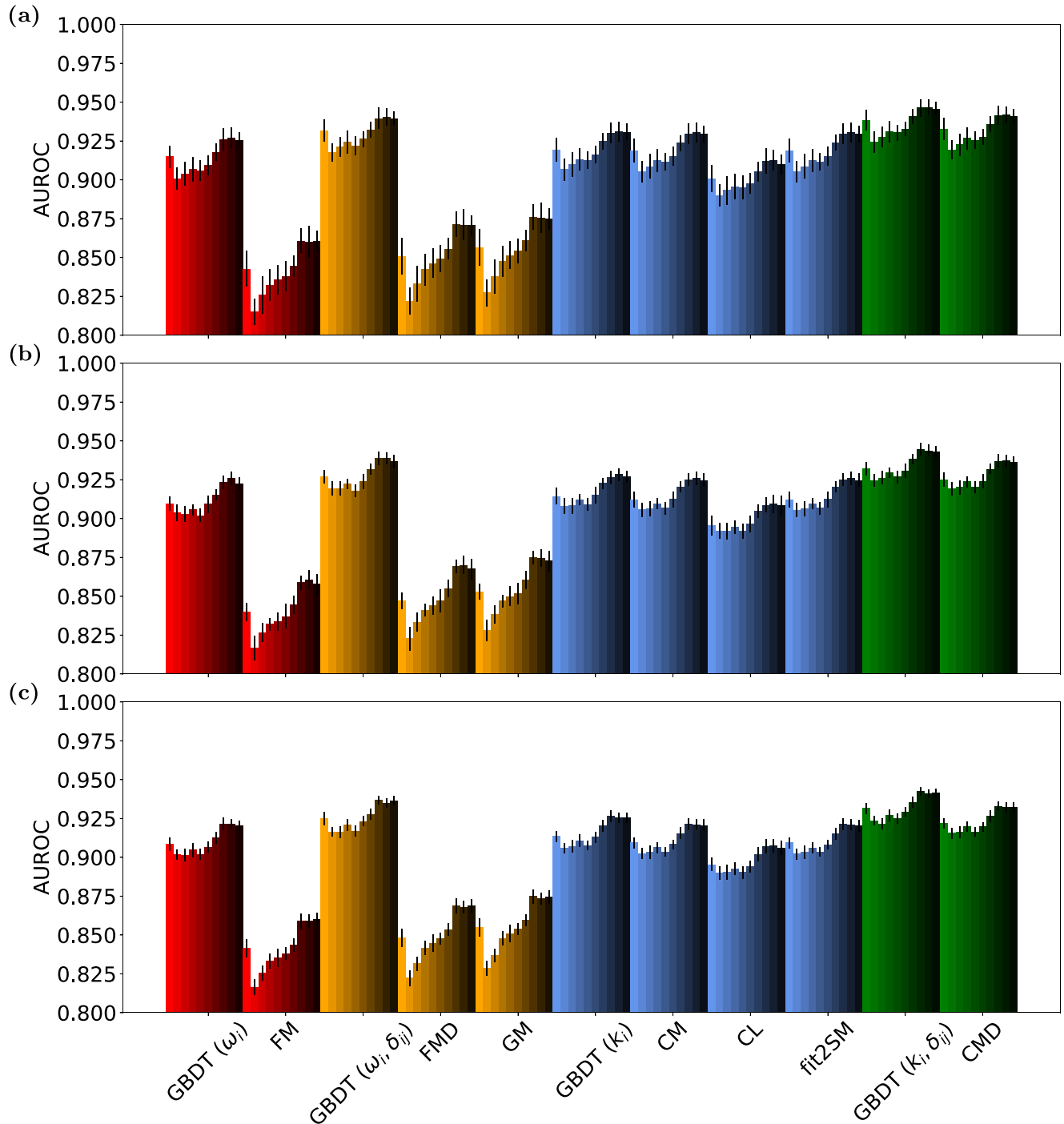


FIG. 12. Performance of the models, measured in terms of AUROC, for all years of the WTW (darker shades correspond to more recent years), when the alternative prediction setting described in Appendix C is adopted. We have randomly selected 10% [panel (a)], 20% [panel (b)], and 30% [panel (c)] of node pairs 40 times to populate the test set, on which the prediction is performed after having trained the model on the remaining pairs of nodes. Each statistical indicator has, then, been averaged over such samples, the standard deviation being represented by a vertical black bar. The results are in agreement with those reported in the main text, as each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterparts—especially for what concerns the CM and the fit2SM.

of the degrees of the involved vertices; the second one is the so-called Random Node-Edge Sampler [24], first sampling nodes uniformly at random and, then, sampling edges uniformly at random from the ones that are incident to the sampled nodes—Although both alternatives aim at contrasting the aforementioned bias, giving the less connected nodes a

higher chance to be selected, this is achieved only to a partial extent.

Overall, it should be noticed that the performance of all methods degrades as “targeted” link removal recipes are implemented, whence our conclusion that existing link prediction algorithms (be they black box or white box) perform

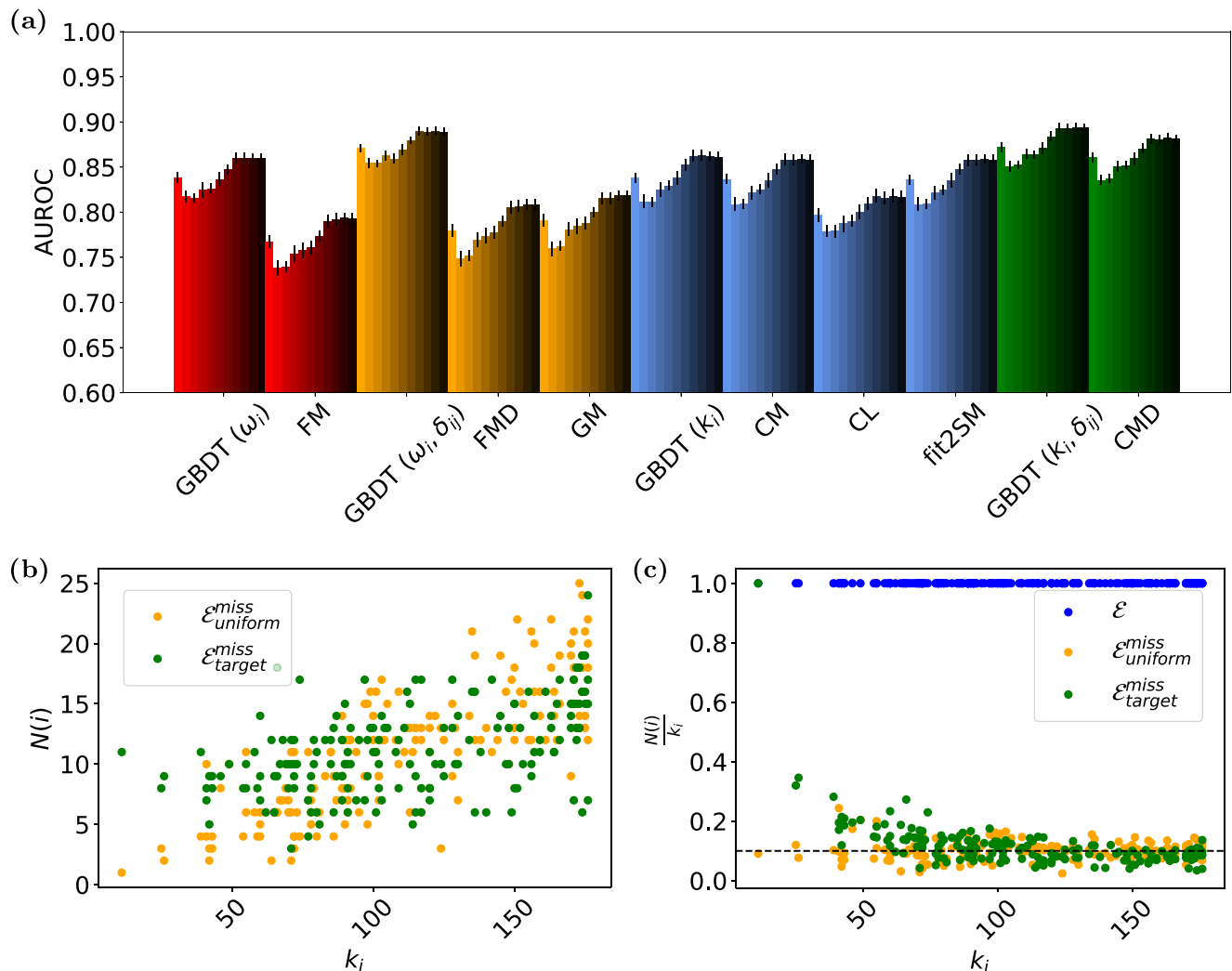


FIG. 13. (a) AUROC scores for “targeted” link removal, for all years of the WTW (darker shades correspond to more recent years). The statistical indicators have been averaged over all samples, the standard deviation being represented by a vertical black bar. We have selected 10% of links 20 times to populate the test set, on which the prediction is performed after having trained the model on the remaining pairs of nodes. (b) Number of occurrences of node i in $\mathcal{E}_{\text{uniform}}^{\text{miss}}$ (populated with randomly selected links from the year 2000 of the WTW) and $\mathcal{E}_{\text{target}}^{\text{miss}}$ (populated by selecting links with a probability that is inversely proportional to the product of the degrees of the involved vertices from the year 2000 of the WTW), scattered vs the degree k_i . (c) Number of occurrences of node i in $\mathcal{E}$, $\mathcal{E}_{\text{uniform}}^{\text{miss}}$, and $\mathcal{E}_{\text{target}}^{\text{miss}}$, rescaled by the degree, scattered vs degree k_i . The results indicate that the performance of all methods degrades as “targeted” link removal recipes are implemented.

better for better connected nodes. Although negative, such a result offers an additional indication about the identity of the “best” explanatory variables in a link-prediction exercise: As also noticed elsewhere [1,49,50] economic, financial (and social) systems are constituted by nodes the “activity” of which is proportional to (if not driven by) their connectedness; as a consequence, controlling for the degrees implies being capable of explaining (most of) each node’s “activity”—as well as most of each node’s structural properties.

Besides incurring into the additional cost required by tuning hyperparameters (see Appendix E), machine learning algorithms need to be fed with (much) more information than that required by white-box algorithms—what is required to make predictions on the “test” portion of the dataset is, in fact, the exact knowledge of the “training” portion of the dataset—in order to achieve a performance that is not necessarily better. While such a need may even represent an advantage when

considering a specific dataset, it turns out to constitute a limitation when coming to generalize, as machine learning approaches either perform quite poorly (e.g., whenever trained on a dataset to predict the missing links of another) or cannot perform at all (e.g., whenever in the presence of missing information). White-box models, on the contrary, perform robustly: Examples are provided by the network reconstruction tasks carried out in Refs. [9,50], where the considered systems have been fully reconstructed by (solely) employing the strengths of nodes and (an estimation of) the link density.

Finally, a statistical analysis about the differences characterizing the performances of our algorithms (see Appendix F) reveals that a clear winner cannot always be unambiguously identified, as the presence of a winner depends on both the score and the algorithm. When considering the AUPRC, for instance, (1) machine learning approaches perform better than the white-box models not taking structural information as

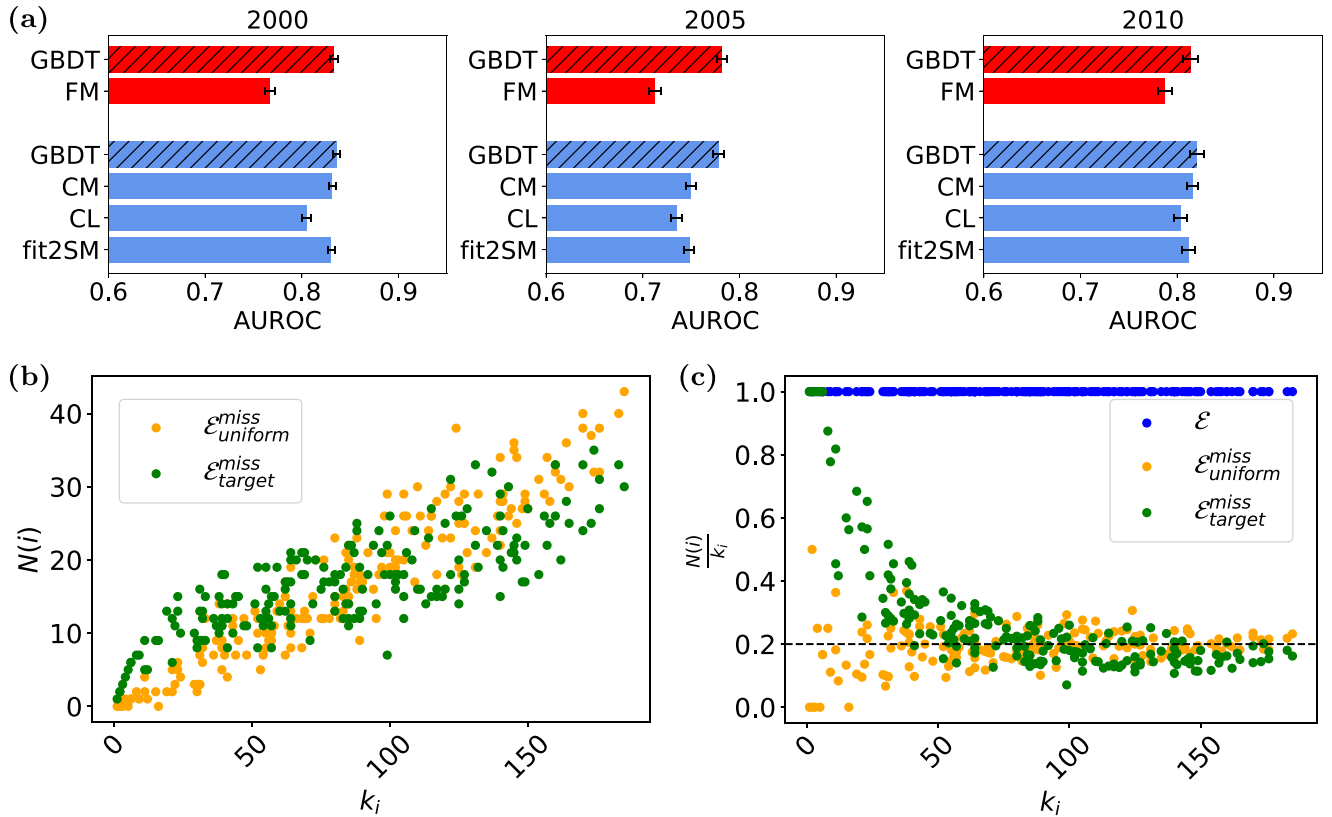


FIG. 14. (a) AUROC scores for targeted link removal, for the years 2000, 2005, and 2010 of eMID. The statistical indicators have been averaged over all samples, the standard deviation being represented by a vertical black bar. We have selected 20% of links 30 times to populate the test set, on which the prediction is performed after having trained the model on the remaining pairs of nodes. (b) Number of occurrences of node i in $\mathcal{E}_{\text{uniform}}^{\text{miss}}$ (populated with randomly selected links from the yearly aggregation of the eMID dataset, where year is 2000) and $\mathcal{E}_{\text{target}}^{\text{miss}}$ (populated by selecting links with a probability that is inversely proportional to the product of the degrees of the involved vertices from the same dataset snapshot), scattered vs degree k_i . (c) Number of occurrences of node i in $\mathcal{E}$, $\mathcal{E}_{\text{uniform}}^{\text{miss}}$, and $\mathcal{E}_{\text{target}}^{\text{miss}}$ rescaled by the degree, scattered vs degree k_i . The results indicate that the performance of all methods degrades as “targeted” link removal recipes are implemented.

input; (2) the former and the latter perform in a statistically undistinguishable way when “mixed” structural information is taken as input; and (3) white-box models (typically) perform better than machine learning approaches when either the plain degree sequence or nonlinear structural information is taken as input.

Let us conclude by stressing once again that our motivation, here, is not that of challenging the performance of machine learning approaches, rather checking under which conditions they can be said to outperform white-box models in link prediction tasks: What emerges from our analysis is that answering such a question crucially depends on the way these exercises are carried out. Still, the results shown in Appendixes E and F motivate us to indicate maximum-entropy models as strong competitors of machine learning approaches, often performing in a comparable way at a substantially lower computational cost.

ACKNOWLEDGMENTS

The authors thank Sadamori Kojaku and Mattia Marzi for insightful discussions. This work has been supported by the following projects: “RE-Net—Reconstructing eco-

nomic networks: from physics to machine learning and back”—2022MTBB22, funded by the European Union Next Generation EU, PNRR Mission 4 Component 2 Investment 1.1, CUP: D53D23002330006; “C2T—From crises to theory: towards a science of resilience and recovery for economic and financial systems”—P2022E93B8, funded by the European Union Next Generation EU, PNRR Mission 4 Component 2 Investment 1.1, CUP: D53D23019330001; “SoBigData RI PPP—SoBigData RI Preparatory Phase Project”, funded by the European Union under the scheme HORIZON-INFRA-2021-DEV-02-01, preparatory phase of new ESFRI research infrastructure projects, Grant Agreement No. 101079043; and “FAIR—Future Artificial Intelligence Research”—Spoke 1 “Human-centered AI”, funded by the European Commission under the Next Generation EU program, PNRR Mission 4 Component 2 Investment 1.3, Grant Agreement No. PE00000013.

Study conception and design: F.S., G.C., and T.S. Data collection: F.S. Analysis and interpretation of results: F.S., G.C., and T.S. Draft manuscript preparation: F.S., G.C., and T.S. Draft manuscript revision: F.S., G.C., and T.S.

The authors declare no competing interests.

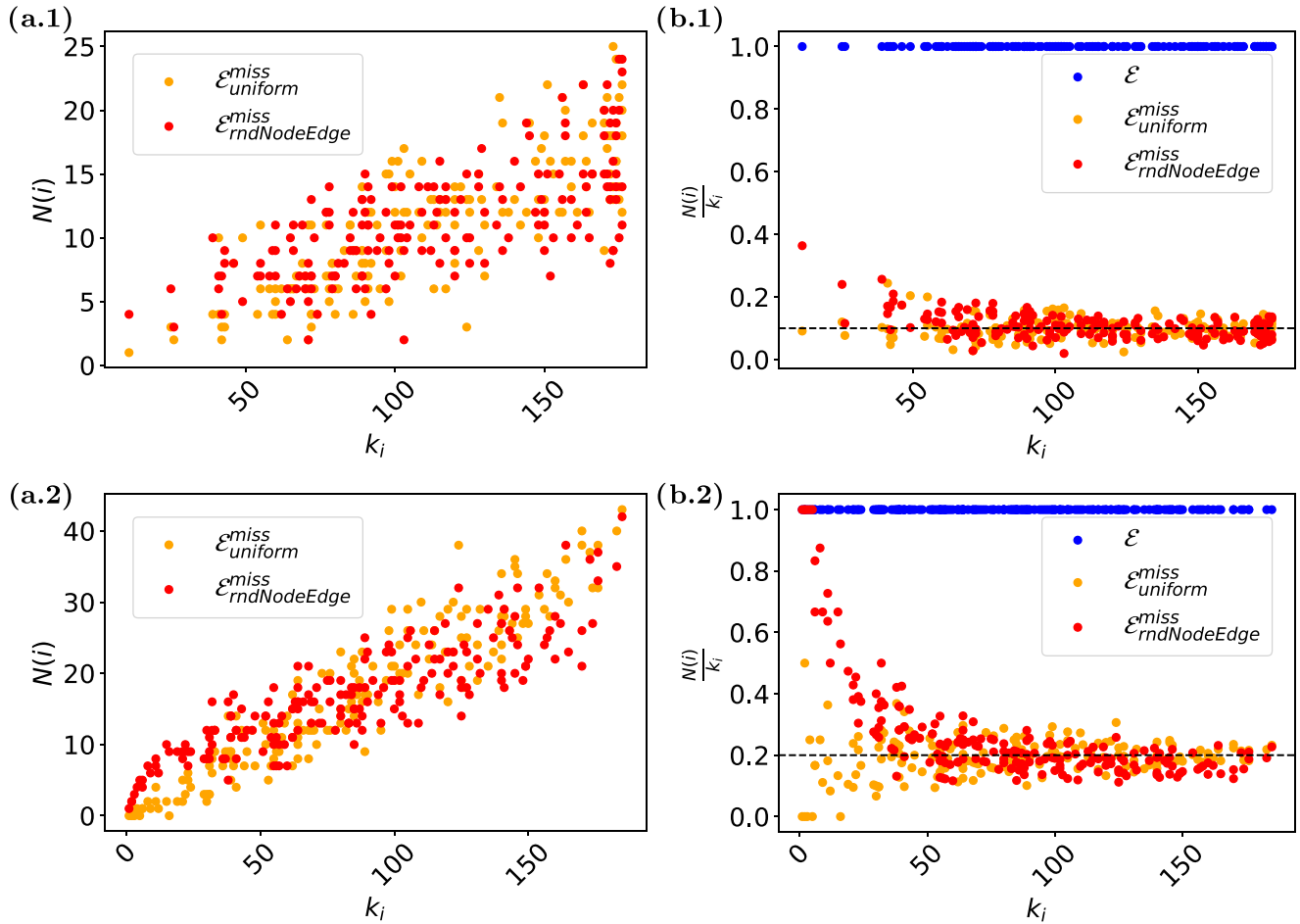


FIG. 15. Number of occurrences of node i in $\mathcal{E}_{\text{uniform}}^{\text{miss}}$, populated with randomly selected links from the year 2000 of the WTW [panel (a.1), 10% of links] and eMID [panel (a.2), 20% of links], and $\mathcal{E}_{\text{rndNodeEdge}}^{\text{miss}}$, populated by implementing the Random Node-Edge Sampler for the year 2000 of the WTW [panel (a.1), 10% of links] and eMID [panel (a.2), 20% of links], scattered vs the degree k_i . (b.1) and (b.2) Number of occurrences of node i in $\mathcal{E}$, $\mathcal{E}_{\text{uniform}}^{\text{miss}}$, and $\mathcal{E}_{\text{rndNodeEdge}}^{\text{miss}}$ rescaled by the degree, scattered vs degree k_i .

DATA AVAILABILITY

Some of the data that support the findings of this article are openly available (World Trade Web curated by Gleditsch [48]). All results corresponding to Figs. 6, 10, 11, 14, 15 (lower), 18, 19, and 20(b) and Tables III, IV, VI are based on the eMID whose nodes are anonymized banks and weights represent exposures in millions of euros are not publicly available because they are owned by a third party and the terms of use prevent public distribution. The data are available from the authors upon reasonable request.

APPENDIX A: MISSING LINKS PREDICTION ON THE WORLD TRADE WEB

Figure 7 depicts the performance of our models for different shares of removed links: As it can be appreciated, the performance of the CMD remains high up to a large percentage of missing links; to be noticed that the performance of the GBDT taking as inputs both “exogenous” and “endogenous” features deteriorates to a lesser extent than the one of the CMD, i.e., its white-box counterpart, as the size of the training set decreases.

The results concerning the AUPRC, shown in Figs. 8 and 9, are in agreement with those reported in the main text as well, as each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterpart—especially for what concerns the CM and the fit2SM.

APPENDIX B: MISSING LINKS PREDICTION ON THE ELECTRONIC MARKET FOR INTERBANK DEPOSITS

Table IV reports the basic statistics for each aggregation scale; Figs. 10 and 11 depict the performance of our models for different aggregation scales. As it can be appreciated, while (1) the results concerning the AUROC are consistent with the ones characterizing the yearly aggregation level, (2) the results concerning the AUPRC change as finer aggregation levels are considered: More specifically, the AUPRC shrinks (much) faster than the AUROC as the aggregation level becomes finer; such a result indicates that, as the list of potentially missing links is gone through, (1) the overall number of truly missing links is (practically, always) recovered quite accurately, although (2) the number of false positives rises—In other words, the performance of all algorithms in

TABLE V. Computational complexity of the models considered in this present article (expressed in milliseconds), for all years of the WTW. We have selected 10% of links 40 times to populate the test set, on which the prediction is performed after having trained the model on the remaining pairs of nodes. Values are reported as averages (across training set–test set realizations) plus/minus one standard deviation.

		WTW features							
		ω_i		ω_i, δ_{ij}		k_i		k_i, δ_{ij}	
Black-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	753 ± 154		813 ± 168		203 ± 97		550 ± 162	
	T^{tune}	$(1.23 \pm 0.19) \times 10^5$		$(1.30 \pm 0.15) \times 10^5$		$(1.41 \pm 0.21) \times 10^5$		$(1.30 \pm 0.18) \times 10^5$	
White-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	FM	53 ± 4	FMD	3023 ± 325	CM	34 ± 12	CMD	546 ± 918
				GM	62 ± 9	CL	4 ± 1		
						fit2SM	22640 ± 2827		

spotting single connections degrades quite rapidly as we move from yearly to daily snapshots (equivalently, toward sparser configurations).

Upon comparing these results with those in Appendix A, we are led to conclude that existing link prediction algorithms (be they black box or white box) perform better for denser configurations.

APPENDIX C: MISSING LINKS PREDICTION IN A DIFFERENT SETTING

Another common setting for prediction—similar to the one adopted in Ref. [7]—consists in randomly splitting all node pairs into a training and a test set, i.e., $\mathcal{U} = \mathcal{U}^{\text{tr}} \cup \mathcal{U}^{\text{te}}$. The

training set is then identified with $\mathbf{D}^{\text{tr}} = \{\mathbf{x}_{ij}, a_{ij}\}_{ij \in \mathcal{U}^{\text{tr}}}$ and the prediction is performed on $\mathbf{D}^{\text{te}} = \{a_{ij}\}_{ij \in \mathcal{U}^{\text{te}}}$. As the evaluation of “endogenous” features is based on the links populating $\mathcal{U}^{\text{tr}}$, this is equivalent to treating all pairs of nodes in $\mathcal{U}^{\text{te}}$ as unconnected.

The results, shown in Fig. 12, are in agreement with those reported in the main text, as each “endogenous” instance of the GBDT performs in a way that is comparable to that of its purely structural, white-box counterparts—especially for what concerns the CM and the fit2SM. To be noticed that the performance of the GBDT taking as inputs both “exogenous” and “endogenous” features deteriorates to a lesser extent than that of the CMD, i.e., its white-box counterpart, as the size of the training set decreases.

TABLE VI. Computational complexity of the models considered in this present article (expressed in milliseconds), for snapshots of eMID at different aggregation levels (top: yearly; middle: monthly; bottom: daily). We have selected 20% of links 30 times to populate the test set, on which the prediction is performed after having trained the model on the remaining pairs of nodes. Values are reported as averages (across training set–test set realizations) plus/minus one standard deviation.

		eMID features (yearly)	
		s_i	k_i
Black-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	695 ± 848	140 ± 90
	T^{tune}	$(7.94 \pm 1.75) \times 10^4$	$(8.32 \pm 2.97) \times 10^4$
White-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	FM 15 ± 10	CM 35 ± 32
			CL 3 ± 2
			fit2SM 13348 ± 8345
		eMID features (monthly)	
		s_i	k_i
Black-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	283 ± 172	92 ± 40
	T^{tune}	$(6.26 \pm 0.86) \times 10^4$	$(5.94 \pm 0.93) \times 10^4$
White-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	FM 7 ± 2	CM 17 ± 5
			CL 2 ± 0
			fit2SM 6906 ± 3637
		eMID features (daily)	
		s_i	k_i
Black-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	FM 90 ± 45	47 ± 26
	T^{tune}	$(5.15 \pm 1.05) \times 10^4$	$(4.23 \pm 0.91) \times 10^4$
White-box algorithms	$T^{\text{tr}} + T^{\text{te}}$	FM 4 ± 1	CM 12 ± 1
			CL 2 ± 0
			fit2SM 2529 ± 1300

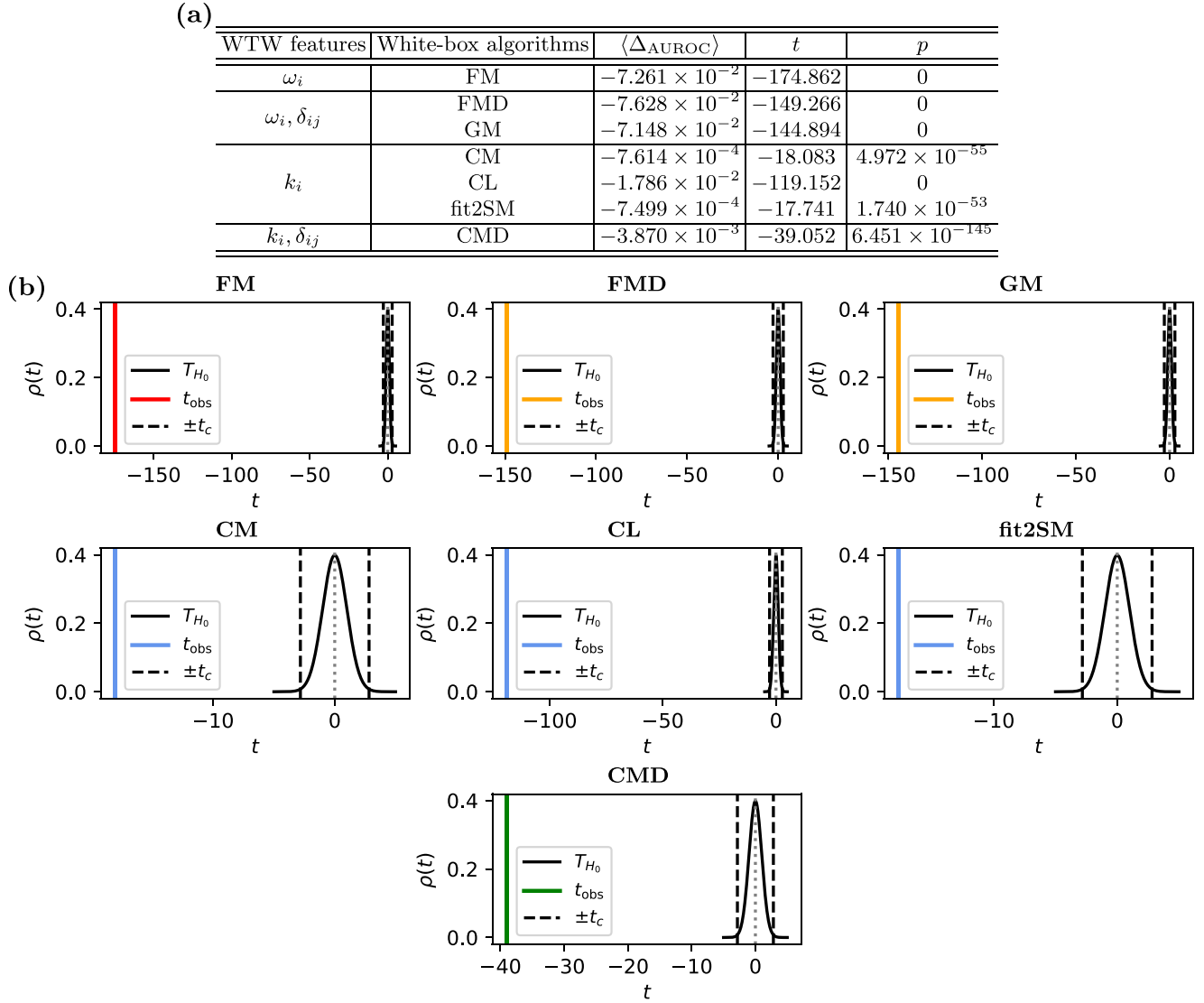


FIG. 16. (a) Statistical significance of the prediction performances of each white-box method vs its black-box counterpart (evaluated in terms of AUROC score), for the WTW when $\mathcal{E}^{\text{miss}}$ is populated with 10% of the existing links. For each pair of methods, the difference of performances is paired across all iterations (40) per year (11), corresponding to a sample size of 440 points. The table entries report the (sample) average difference $\bar{\Delta}$, the t value expressing the distance of the (sample) average difference from 0, in units of the (sample) variance, and the p value. (b) Distributions representing our paired samples t -test for the AUROC score and the WTW. Results do not change if the analysis is carried out in a yearly fashion.

APPENDIX D: LINK REMOVAL IN A DIFFERENT SETTING

The results that we have shown in the main text have been obtained by populating $\mathcal{E}^{\text{miss}}$ with links sampled from $\mathcal{E}$ with a uniform probability [2,6,7,11]. Since node i appears in $\mathcal{E}$ as a vertex of an undirected edge exactly k_i times, we expect the same dependence on the degree to characterize $\mathcal{E}^{\text{miss}}$ as well: In other words, uniformly sampling $\mathcal{E}$ leads to populate $\mathcal{E}^{\text{miss}}$ with (links connecting) high-degree nodes. It is, thus, natural to ask how a different sampling scheme would affect the results.

In order to answer such a question, we have replicated the analysis under a couple of “targeted” sampling prescriptions. In the first case, we have removed links by selecting edges with a probability that is inversely proportional to the product

of the degrees of the involved vertices: More specifically, at the generic iteration, the link connecting nodes i and j was sampled with a probability reading $p_{ij} \propto 1/k_i k_j$ (see Figs. 13 and 14). In the second case, we have implemented the so-called Random Node-Edge Sampler [24], first sampling a node uniformly at random and, then, sampling an edge uniformly at random from the ones that are incident to the sampled node—The procedure is repeated until the desired size is reached.

Comparing Figs. 13 and 14 with Figs. 5 and 6 reveals that (1) although removing links randomly leads to remove more links from the more connected nodes, the “targeted” sampling prescriptions considered as alternatives to the uniform one—basically performing the same, as evident upon looking at Fig. 15—may incur into the same problems, and (2) the performance of all methods degrades when “targeted”

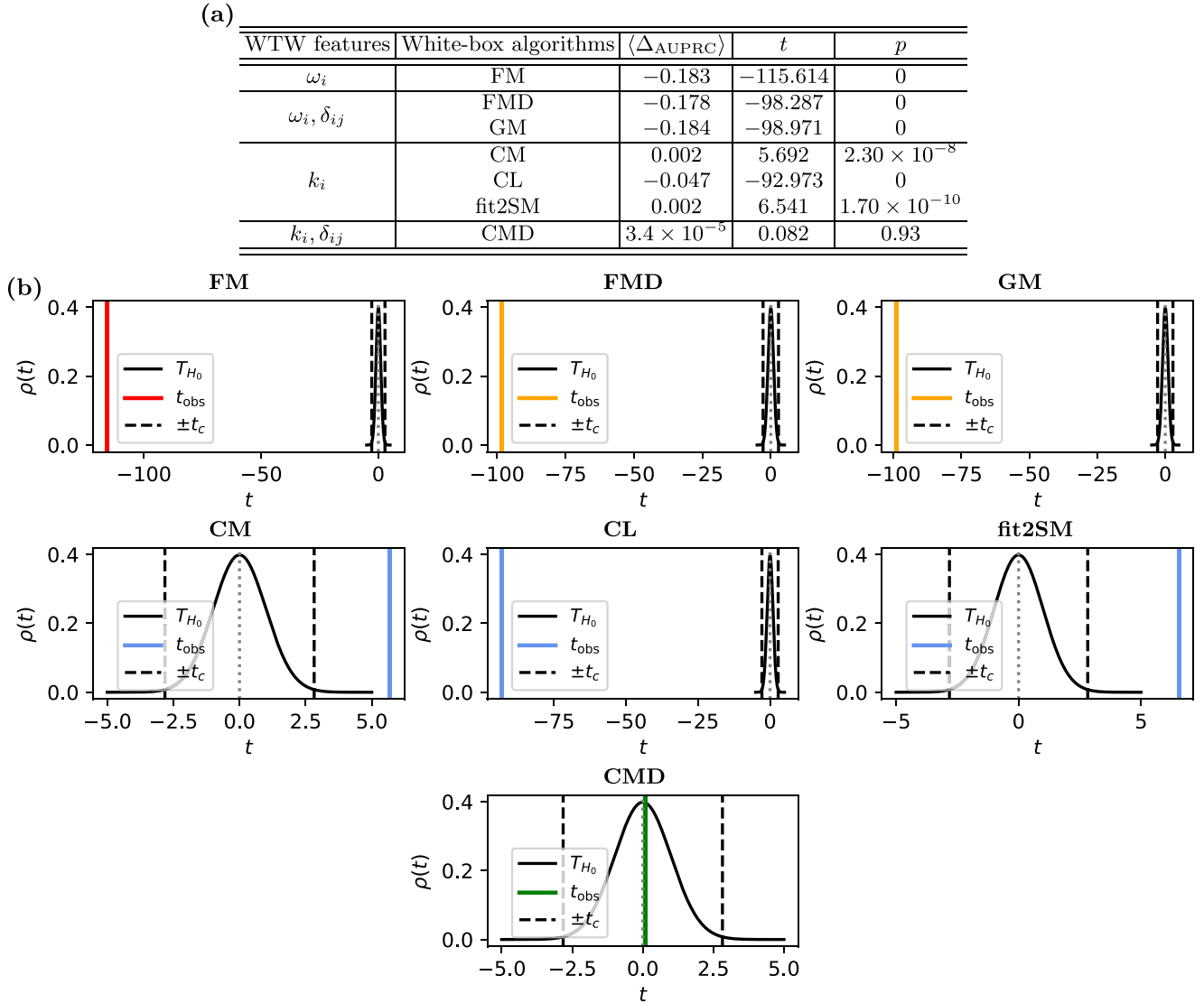


FIG. 17. (a) Statistical significance of the prediction performances of each white-box method vs its black-box counterpart (evaluated in terms of AUPRC score), for the WTW when $\mathcal{E}^{\text{miss}}$ is populated with 10% of the existing links. For each pair of methods, the difference of performances is paired across all iterations (40) per year (11), corresponding to a sample size of 440 points. The table entries report the (sample) average difference $\bar{\Delta}$, the t value expressing the distance of the (sample) average difference from 0, in units of the (sample) variance, and the p value. (b) Distributions representing our paired samples t -test for the AUPRC score and the WTW. Results do not change if the analysis is carried out in a yearly fashion.

sampling prescriptions are considered—the methods taking as inputs the degrees being more affected than others. We are, thus, led to conclude that existing link prediction algorithms (be they black box or white box) perform better for better connected nodes.

APPENDIX E: COMPUTATIONAL COMPLEXITY OF MISSING LINKS PREDICTION ALGORITHMS

Let us now study the computational complexity of the algorithms considered in the present work: the results are reported in Tables V and VI. In the case of white-box methods, the total computational time corresponds to the average time needed for training and predicting across training set–test set realizations (i.e., $T^{\text{tr}} + T^{\text{te}}$): Let us observe that four white-box algorithms (CL, CM, FM, GM) neatly beat their black-box

counterpart. In the case of black-box methods, instead, it is also necessary to take into account the time needed to tune the hyperparameters (i.e., T^{tune}), which, however, does not need to be repeated across training set–test set realizations; still, adding it leads the computational complexity of the black-box methods to significantly exceed that of white-box methods.

APPENDIX F: STATISTICAL SIGNIFICANCE OF PREDICTION PERFORMANCES

Let us now check whether the prediction performances of the algorithms considered in the present work are significantly different or not. To this aim, let us consider the score difference

$$\Delta_{\text{Score}} = \text{Score}_{\text{wb}} - \text{Score}_{\text{bb}} \quad (\text{F1})$$

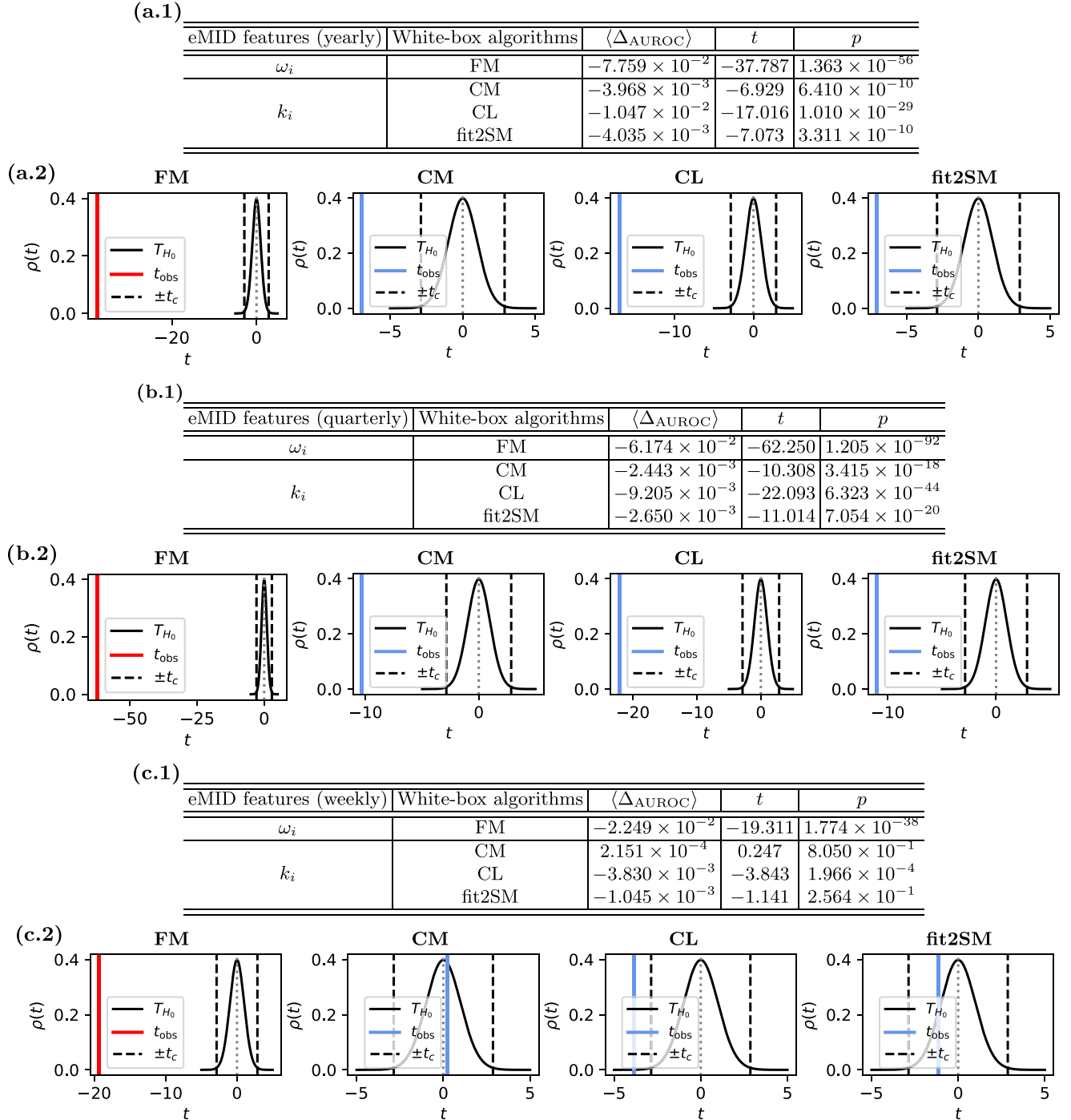


FIG. 18. (a.1)–(c.1) Statistical significance of the prediction performances of each white-box method vs its black-box counterpart (evaluated in terms of AUROC score), for eMID when $\mathcal{E}^{\text{miss}}$ is populated with 20% of the existing links. For each pair of methods, the difference of performances is paired across all iterations for all snapshots at the (a) yearly (3 snapshots with 30 iterations each), (b) quarterly (4 snapshots with 30 iterations each), and (c) weekly (4 snapshots with 30 iterations each) aggregation levels, corresponding to a sample size of 90 points (yearly) and 120 points (quarterly and weekly). The table entries report the (sample) average difference Δ , the t value expressing the distance of the (sample) average difference from 0, in units of the (sample) variance and the p value. (a.2)–(c.2) Distributions representing our paired samples t -test for the AUROC score and eMID, at the corresponding aggregation level.

between the white-box methods considered here and their black-box counterpart(s) and ask whether it is significantly different from zero. More formally, we carry out a number of paired samples t -tests: Under the null hypothesis $\Delta_{\text{Score}} = 0$,

the test statistics

$$t = \frac{\bar{\Delta} - 0}{s/\sqrt{n}}, \quad (\text{F2})$$

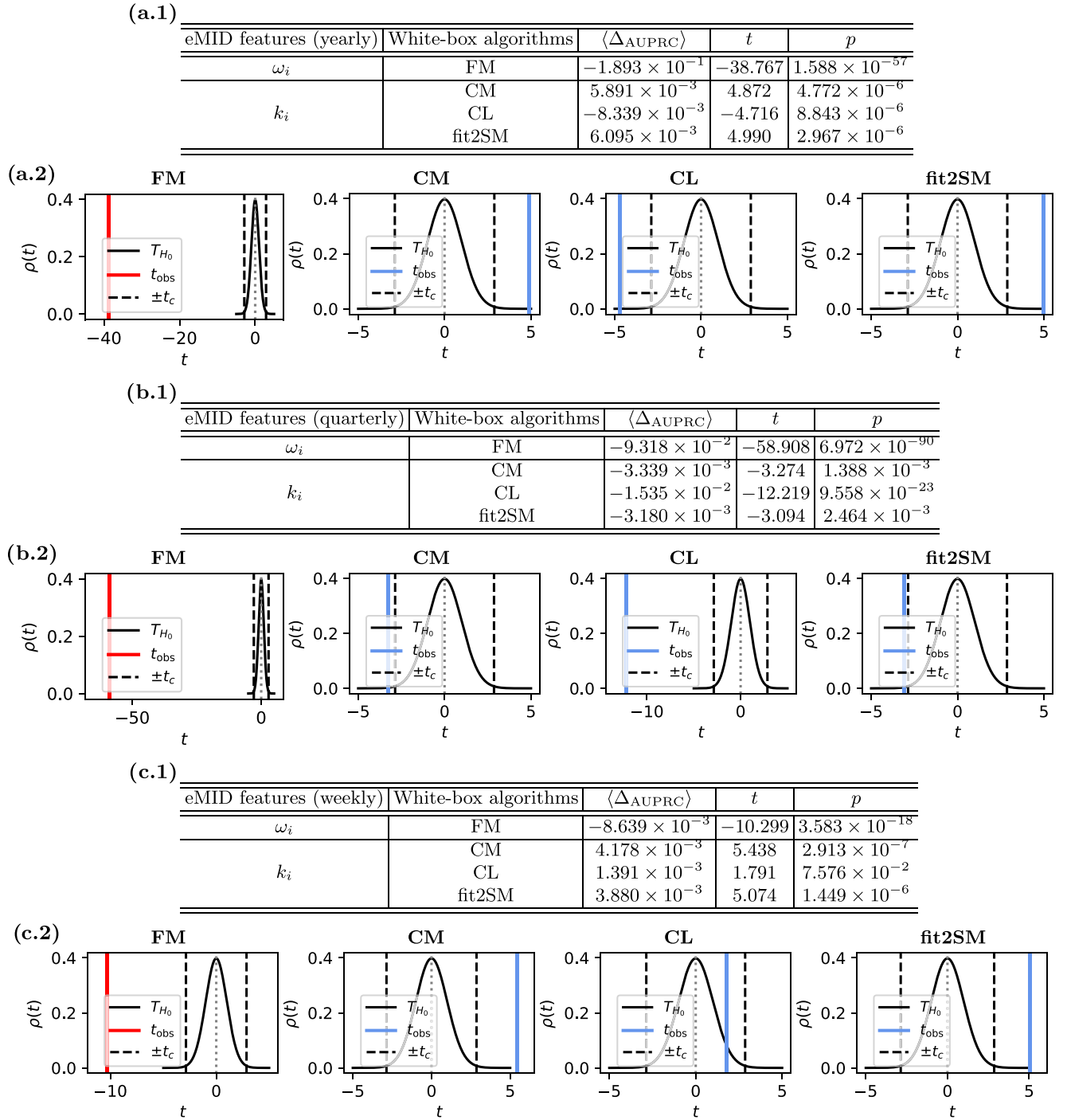


FIG. 19. (a.1)–(c.1) Statistical significance of the prediction performances of each white-box method vs its black-box counterpart (evaluated in terms of AUPRC score), for eMID when $\mathcal{E}^{\text{miss}}$ is populated with 20% of the existing links. For each pair of methods, the difference of performances is paired across all iterations for all snapshots at (a) the yearly (3 snapshots with 30 iterations each), (b) quarterly (4 snapshots with 30 iterations each), and (c) weekly (4 snapshots with 30 iterations each) aggregation levels, corresponding to a sample size of 90 points (yearly) and 120 points (quarterly and weekly). The table entries report the (sample) average difference $\bar{\Delta}$, the t -value expressing the distance of the (sample) average difference from 0, in units of the (sample) variance, and the p value. (a.2)–(c.2) Distributions representing our paired samples t -test for the AUPRC score and eMID, at the corresponding aggregation level.

with

$$\bar{\Delta} = \frac{\sum_{i=1}^n \Delta_i}{n} \quad (\text{F3})$$

and

$$s = \sqrt{\frac{\sum_{i=1}^n (\Delta_i - \bar{\Delta})^2}{n-1}}, \quad (\text{F4})$$

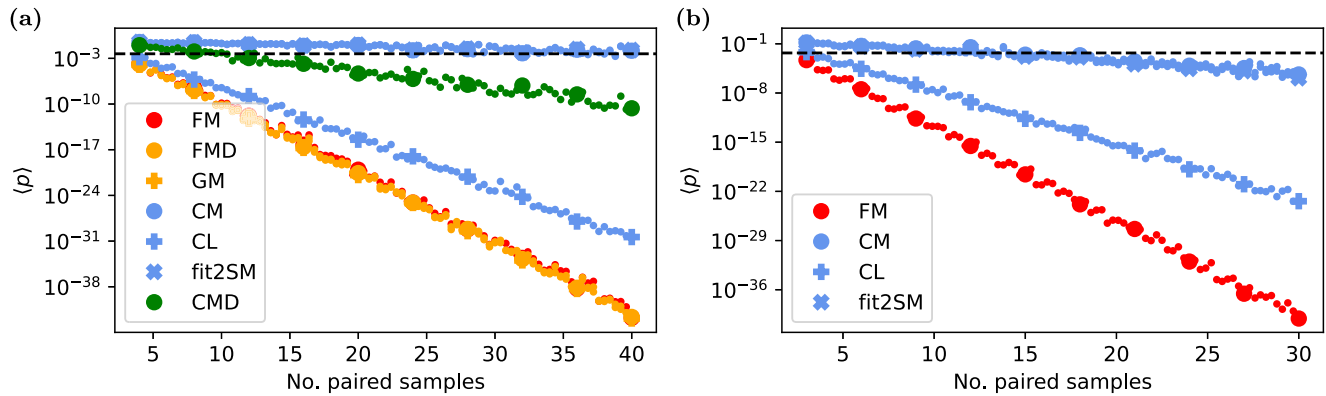


FIG. 20. (a) Sensitivity analysis of the results concerning the statistical significance of the prediction performances of our algorithms for the year 2000 of the WTW (AUROC score). $\mathcal{E}^{\text{miss}}$ is populated with 10% of the existing links. For each pair of methods, the difference of performances is paired across the number of samples indicated on the x axis. (b) Sensitivity analysis of the results concerning the statistical significance of the prediction performances of our algorithms for the year 2000 of eMID (AUROC score). $\mathcal{E}^{\text{miss}}$ is populated with 20% of the existing links. For each pair of methods, the difference of performances is paired across the number of samples indicated on the x axis. Our results are robust across the cardinality of samples.

obeys a Student's t -distribution with $n - 1$ degrees of freedom—for all practical purposes, a Normal distribution (i.e., a Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$).

For what concerns the WTW, the paired samples t -test is carried out by pooling together the 40 observations (i.e., training set–test set splits) per year, amounting at 440 observations in total. While Fig. 16 suggests that black-box methods always perform significantly better than their white-box counterparts, Fig. 17 reveals the picture to be more nuanced: In this second case, in fact, a clear winner cannot be unambiguously identified as (1) black-box models perform better than the white-box counterparts not taking structural information as input; (2) the former and the latter perform in a statistically undistinguishable way when “mixed” structural information

is taken as input; and (3) white-box models (typically) perform better than black-box counterparts when either the plain degree sequence or nonlinear structural information is taken as input.

For what concerns eMID, instead, the paired samples t -test is carried out by pooling together the 40 observations (i.e., training set–test set splits) per year, amounting at 440 observations in total. As Figs. 18 and 19 suggest, the presence of a winner depends on the level of aggregation as well; the finer the aggregation, the higher the chance that white-box models are favored.

Finally, as the sensitivity analysis depicted in Fig. 20 confirms, our results are quite robust across the cardinality of samples.

- [1] T. Squartini, G. Caldarelli, G. Cimini, A. Gabrielli, and D. Garlaschelli, Reconstruction methods for networks: The case of economic and financial systems, *Phys. Rep.* **757**, 1 (2018).
- [2] L. Lü and T. Zhou, Link prediction in complex networks: A survey, *Physica A* **390**, 1150 (2011).
- [3] D. Liben-Nowell and J. Kleinberg, The link-prediction problem for social networks, *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019 (2007).
- [4] G. Berlusconi, F. Calderoni, N. Parolini, M. Verani, and C. Piccardi, Link prediction in criminal networks: A tool for criminal intelligence analysis, *PLoS One* **11**, e0154244 (2016).
- [5] M. Jalili, Y. Orouskhani, M. Asgari, N. Alipourfard, and M. Perc, Link prediction in multiplex online social networks, *R. Soc. Open Sci.* **4**, 160863 (2017).
- [6] F. Parisi, G. Caldarelli, and T. Squartini, Entropy-based approach to missing-links prediction, *Appl. Network Sci.* **3**, 17 (2018).
- [7] L. Mungo, F. Lafond, P. Astudillo-Estévez, and J. D. Farmer, Reconstructing production networks using machine learning, *J. Econ. Dyn. Control* **148**, 104607 (2023).
- [8] M. Bardoscia, P. Barucca, S. Battiston, F. Caccioli, G. Cimini, D. Garlaschelli, F. Saracco, T. Squartini, and G. Caldarelli, The physics of financial networks, *Nat. Rev. Phys.* **3**, 490 (2021).
- [9] L. N. Ialongo, C. de Valk, E. Marchese, F. Jansen, H. Zmarrou, T. Squartini, and D. Garlaschelli, Reconstructing firm-level interactions in the Dutch input–output network from production constraints, *Sci. Rep.* **12**, 11847 (2022).
- [10] A. L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286**, 509 (1999).
- [11] T. Zhou, L. Lü, and Y.-C. Zhang, Predicting missing links via local information, *Eur. Phys. J. B* **71**, 623 (2009).
- [12] L. Katz, A new status index derived from sociometric analysis, *Psychometrika* **18**, 39 (1953).
- [13] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Comput. Networks ISDN Syst.* **30**, 107 (1998).
- [14] P. Y. Chebotarev and E. Shamis, The matrix-forest theorem and measuring relations in small social groups, *Autom. Remote Control* **58**, 1505 (1997).
- [15] G. Jeh and J. Widom, SimRank: A measure of structural-context similarity, in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)* (Association for Computing Machinery, New York, NY, 2002), pp. 538–543.

- [16] W. Liu and L. Lü, Link prediction based on local random walk, *Europhys. Lett.* **89**, 58007 (2010).
- [17] A. Clauset, C. Moore, and M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks, *Nature (London)* **453**, 98 (2008).
- [18] R. Guimerà and M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, *Proc. Natl. Acad. Sci. USA* **106**, 22073 (2009).
- [19] F. Adriaens, A. Mara, J. Lijffijt, and T. De Bie, Block-approximated exponential random graphs, in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, Piscataway, NJ, 2020), pp. 70–80.
- [20] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer, Learning probabilistic relational models, in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)* (Morgan Kaufmann, San Francisco, CA, 1999), Vol. 2, pp. 1300–1307.
- [21] K. Yu, W. Chu, S. Yu, V. Tresp, and Z. Xu, Stochastic relational models for discriminative link prediction, in *Advances in Neural Information Processing Systems 19*, edited by B. Schölkopf, J. Platt, and T. Hofmann (MIT Press, Cambridge, MA, 2007), pp. 1553–1560.
- [22] The terms “endogenous” and “exogenous” are borrowed from economics to distinguish between features that are structural in nature, like the degrees (“endogenous”, i.e., “inherent to the structure”) and features that, instead, are not, like the gross domestic products and the geographic distances (“exogenous”, i.e., “not inherent to the structure”).
- [23] A. Polanco and M. E. J. Newman, Drug-disease networks and drug repurposing, *PLoS Comput. Biol.* **21**, e1013595 (2025).
- [24] V. Krishnamurthy, M. Faloutsos, M. Chrobak, L. Lao, J. H. Cui, and A. G. Percus, Reducing large internet topologies for faster simulations, in *NETWORKING 2005. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, edited by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, R. Boutaba, K. Almeroth, R. Puigjaner, S. Shen, and J. P. Black, Lecture Notes in Computer Science (Springer, Berlin, 2005), pp. 328–341.
- [25] W. Isard, Location theory and trade theory: Short-run analysis, *Q. J. Econ.* **68**, 305 (1954).
- [26] J. Tinbergen, *Shaping the World Economy: Suggestions for an International Economic Policy* (Twentieth Century Fund, New York, 1962).
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [28] J. H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Stat.* **29**, 1189 (2001).
- [29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017), pp. 3146–3154.
- [30] J. Park and M. E. J. Newman, Statistical mechanics of networks, *Phys. Rev. E* **70**, 066117 (2004).
- [31] G. Bianconi, The entropy of randomized network ensembles, *Europhys. Lett.* **81**, 28005 (2008).
- [32] T. Squartini and D. Garlaschelli, Analytical maximum-likelihood method to detect patterns in real networks, *New J. Phys.* **13**, 083001 (2011).
- [33] A. Fronczak, Statistical mechanics of the international trade network, *Phys. Rev. E* **85**, 056113 (2012).
- [34] T. Squartini, R. Mastrandrea, and D. Garlaschelli, Unbiased sampling of network ensembles, *New J. Phys.* **17**, 023052 (2015).
- [35] F. Saracco, R. Di Clemente, A. Gabrielli, and T. Squartini, Randomizing bipartite networks: The case of the World Trade Web, *Sci. Rep.* **5**, 10595 (2015).
- [36] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli, The statistical physics of real-world networks, *Nat. Rev. Phys.* **1**, 58 (2019).
- [37] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Principles of statistical mechanics of uncorrelated random networks, *Nucl. Phys. B* **666**, 396 (2003).
- [38] J. Park and M. E. J. Newman, Origin of degree correlations in the Internet and other networks, *Phys. Rev. E* **68**, 026112 (2003).
- [39] D. Garlaschelli and M. I. Loffredo, Maximum likelihood: Extracting unbiased information from complex networks, *Phys. Rev. E* **78**, 015101(R) (2008).
- [40] N. Vallarano, M. Bruno, E. Marchese, G. Trapani, F. Saracco, G. Cimini, M. Zanon, and T. Squartini, Fast and scalable likelihood maximization for exponential random graph models with local constraints, *Sci. Rep.* **11**, 15227 (2021).
- [41] M. Di Vece, Deterministic, quenched, and annealed parameter estimation for heterogeneous network models, *Phys. Rev. E* **108**, 054301 (2023).
- [42] M. Marzi, F. Giuffrida, D. Garlaschelli, and T. Squartini, Reproducing the first and second moment of empirical degree distributions *Phys. Rev. Res.* **8**, 013141 (2026).
- [43] In a sense, the link prediction problem is interpreted as an instance of the network reconstruction one [1,49], the aim of which is precisely that of inferring a graph structure from a set of observed quantities—although network reconstruction usually deals with *less information* to predict *more aggregate* properties [see also Fig. 1(b)].
- [44] F. Picciolo, T. Squartini, F. Ruzzenenti, R. Basosi, and D. Garlaschelli, The role of distances in the World Trade Web, in *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS)* (IEEE Computer Society, Los Alamitos, CA, 2012), pp. 784–792.
- [45] G. Bianconi, Information theory of spatial network ensembles, in *Handbook on Entropy, Complexity and Spatial Dynamics* (Edward Elgar Publishing, Cheltenham, UK, 2021), pp. 61–96.
- [46] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Scale-free networks from varying vertex intrinsic fitness, *Phys. Rev. Lett.* **89**, 258702 (2002).
- [47] D. Garlaschelli and M. I. Loffredo, Fitness-dependent topological properties of the World Trade Web, *Phys. Rev. Lett.* **93**, 188701 (2004).
- [48] K. S. Gleditsch, Expanded trade and GDP data, *J. Conflict Resolut.* **46**, 712 (2002).

- [49] G. Cimini, R. Mastrandrea, and T. Squartini, *Reconstructing Networks* (Cambridge University Press, Cambridge, 2021).
- [50] G. Cimini, T. Squartini, D. Garlaschelli, and A. Gabrielli, Systemic risk analysis on reconstructed economic and financial networks, *Sci. Rep.* **5**, 15758 (2015).
- [51] J. Park and M. E. J. Newman, Solution of the two-star model of a network, *Phys. Rev. E* **70**, 066146 (2004).
- [52] M. Bolfe, F. L. Metz, E. Guzmán-González, and I. P. Castillo, Analytic solution of the two-star model with correlated degrees, *Phys. Rev. E* **104**, 014147 (2021).