Link Prediction in Bipartite Nested Networks

Medo, Matúš ; Mariani, Manuel ; Lü, Linyuan

Abstract: Real networks typically studied in various research fields—ecology and economic complexity, for example—often exhibit a nested topology, which means that the neighborhoods of high-degree nodes tend to include the neighborhoods of low-degree nodes. Focusing on nested networks, we study the problem of link prediction in complex networks, which aims at identifying likely candidates for missing links. We find that a new method that takes network nestedness into account outperforms well-established link-prediction methods not only when the input networks are sufficiently nested, but also for networks where the nested structure is imperfect. Our study paves the way to search for optimal methods for link prediction in nested networks, which might be beneficial for World Trade and ecological network analysis

DOI: https://doi.org/10.3390/e20100777

entropy

MDPI

Article

Link Prediction in Bipartite Nested Networks

Matúš Medo [1,2,3,*], Manuel Sebastian Mariani [1,4] and Linyuan Lü [1,5]

1 Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China; manuel.sebastian.mariani@gmail.com (M.S.M.); linyuan.lv@uestc.edu.cn (L.L.)
2 Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland
3 Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland
4 URPP Social Networks, Universität Zürich, 8050 Zürich, Switzerland
5 Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, China
* Correspondence: matus.medo@unifr.ch

Received: 31 July 2018; Accepted: 8 October 2018; Published: 10 October 2018

Abstract: Real networks typically studied in various research fields—ecology and economic complexity, for example—often exhibit a nested topology, which means that the neighborhoods of high-degree nodes tend to include the neighborhoods of low-degree nodes. Focusing on nested networks, we study the problem of link prediction in complex networks, which aims at identifying likely candidates for missing links. We find that a new method that takes network nestedness into account outperforms well-established link-prediction methods not only when the input networks are sufficiently nested, but also for networks where the nested structure is imperfect. Our study paves the way to search for optimal methods for link prediction in nested networks, which might be beneficial for World Trade and ecological network analysis.

Keywords: link prediction; nested networks; bipartite networks

1. Introduction

Link prediction is a popular problem in network science [1–4]. The goal of link prediction is to identify the links that are missing because of erroneous or incomplete data (such as in predicting gene interactions from available data [5]), or links that are likely to appear in the system in the course of its temporal evolution (when we speak of a network that naturally grows or otherwise evolves in time [6,7]). Note that not all networks are equally permissible to link prediction as intrinsic network randomness and link sparsity both contribute to making link prediction more difficult. The attempt to quantify network "link predictability" [8] is highly relevant in this respect.

Various classes of systems have their specificities that impact the link prediction process. For example, social networks typically feature high clustering coefficient (if person A knows B and B knows C, then it is likely that also A knows C), which implies that link prediction methods based on closing such open triangles typically perform well [2,9]. In this sense, the link prediction accuracy depends on the understanding of network specificities, that is, whether the link prediction algorithm can well reflect the corresponding mechanisms of network organization. However, even the most recent reviews of predictions in complex networks [10,11] do not specifically consider link predictions in nested networks. Our goal here is to fill the gap and provide a comparison of how various link prediction methods perform in nested networks.

A network is nested when neighborhoods of low-degree nodes tend to be subsets of neighborhoods of high-degree nodes [12,13]. Nested networks have been much studied with respect to their analysis [14], their formation [15–18], and their implications [19–21]. Nested structures are often found in ecological networks, in particular in plant–animal mutualistic networks [13] and species

geographic distribution patterns [12] (see [14,22] for reviews on ecological nested networks), as well as in country-product trade data [23].

Despite the interest of scholars from diverse fields in nestedness, only few attempts [24,25] have been made to exploit the nested topology to identify missing links. Developing effective methodologies for link prediction in nested networks can improve our understanding of at least two important classes of systems which typically exhibit a nested topology: World Trade Networks and ecological mutualistic networks. When analyzing World Trade networks, available data are often characterized by inaccurate or missing information. In the COMTRADE dataset (available at http://comtrade.un.org), for example, given the export of product α from country i to j, the export volume declared by country i often does not match the import volume declared by country j for that product [26]. In the recent Economic Complexity research field [23,27], this has led scholars to investigate the robustness of network-based metrics of country fitness and product complexity against network structural perturbation [28–30].

Analysis of ecological mutualistic networks is also affected by the fact that unobserved interactions might simply be rare and, therefore, require a longer observation time [31]. Such missing links are fundamentally different from biologically forbidden interactions that cannot physically take place due to species-specific reasons, such as size mismatch or temporal uncoupling [32]. This leads to the problem of estimating sampling bias [33] and its impact on observed network topological properties [34].

Our main contribution is two-fold. First, we provide an extensive benchmarking of link-prediction techniques on synthetic and real data that exhibit a nested structure. In line with recent developments [35,36], we consider not only networks that are overall nested, but also networks that are partitioned into blocks that internally exhibit a nested structure ("in-block nested" networks [36]). Intriguingly, we find that some well-established approaches to link prediction in bipartite networks fail on nested networks. Second, we develop and validate a link-prediction method that takes full advantage of the nested structure of the input data. Importantly, besides achieving optimal performance in perfectly nested networks, the new method performs well also on networks with imperfect nestedness structure, up to a certain number of discrepancies from a perfectly nested network.

2. Methods

Before detailing various link prediction methods and their evaluation procedure, we introduce the notation used in this paper. The input bipartite network consists of two sets of nodes with links running only between nodes from different groups. The sizes of the two sets are N_1 and N_2, respectively. Nodes in the two groups are labeled with Latin ($i, j, \cdots = 1, \ldots, N_1$) and Greek ($\alpha, \beta, \cdots = 1, \ldots, N_2$) indices, respectively. The network structure can be captured in the $N_1 \times N_2$ biadjacency matrix B with elements $B_{i\alpha}$ ($B_{i\alpha}$ is one if nodes i and α are connected, and zero otherwise). The sets of neighbors of nodes i and α are Γ_i and Φ_α, respectively. The sizes of these neighborhoods then define the node degree values, $k_i := |\Gamma_i|$ and $d_\alpha := |\Phi_\alpha|$. Finally, the number of links in the network is $E := \sum_i k_i = \sum_\alpha d_\alpha$.

Since the maximal possible number of links in a bipartite network is $N_1 N_2$, there are $N_1 N_2 - E$ links that are not present in the input data. In link prediction, we aim to assign score $s_{i\alpha}$, which reflects the link likelihood, to all these links. Links are then ranked by the score in decreasing order; links at the top of this ranking are the most likely candidates for "missing" links.

2.1. Link Prediction Methods

2.1.1. Preferential Attachment Index (PrefA)

The number of common neighbors of two nodes is the usual benchmark link prediction method in unipartite networks [2]. However, nodes i and α in a bipartite network cannot have any common neighbors by definition, which makes the number of common neighbors as well as popular derived metrics such as the Adamic–Adar index and the Jaccard coefficient [4] not applicable to bipartite

networks. The simplest local link prediction metric is thus the preferential attachment index $k_i d_\alpha$ where the name is of due to the close relation with the preferential attachment mechanism [37]. Notably, this simple approach was found to outperform more sophisticated algebraic link prediction methods based on the eigenvalue decomposition of the network biadjacency matrix B [38].

2.1.2. Number of Local Community Links (LCL)

While the common neighbor metric based on paths of length two between the nodes from different sets is not applicable to bipartite networks, paths of length three can exist between nodes from different sets and can be used for link prediction [39]. Since nodes i and α have k_i and d_α neighbors respectively, there can be at most $k_i d_\alpha$ links that connect them—this quantity is directly used as link prediction score by the preferential attachment (PrefA) index above. The number of local community links, by contrast, takes into account only the actually existing links between the neighbors of nodes i and α, and in this way takes the local network structure into account. In Figure 1, for example, nodes i and α have comparatively high degree, resulting in high PrefA score of the possible link between them. However, link (i, α) would be a bridge between two otherwise little connected parts of the network. From the point of view of network structure, link (i, α) thus seems little likely which manifests itself in the small number of local community links nodes i and α, and a correspondingly low LCL score.

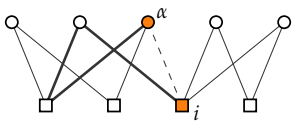

Figure 1. A toy example of a bipartite network where $N_1 = 4$ and $N_2 = 5$. Network links are shown with the solid lines. The dashed line shows the possible link between nodes i and α whose likelihood is being evaluated. The thick lines highlight the only existing link between the neighbors of nodes i and α (see the prediction method "Number of Local Community Links"). The preferential attachment index (PrefA) and LCL scores of the link between the highlighted nodes are 6 and 1, respectively.

2.1.3. Probabilistic Spreading (ProbS)

Also known under the name *mass diffusion*, probabilistic spreading is a network-based recommendation method [40] that was later generalized in many different ways [41,42]. While the original goal of the method is to produce a *personalized* list of recommended items for each individual user, these personalized recommendation scores can be also used as link prediction scores. In the case of probabilistic spreading, the recommendation score is computed by a step-wise propagation process inspired by random walk on the network. For a given target node i (for which the "recommendation" is being computed), unit resource is allocated to all nodes α connected with node i (and zero resource otherwise). We denote the initial resource vector as $f_\alpha^{(i)}$ where the superscript highlights the target node i. In the first step, the resource propagates along the network's links to all nodes j in the other set of nodes by being divided uniformly among the adjacent nodes. The resulting resource vector $g_j^{(i)}$ can be computed as

$$g_j^{(i)} = \sum_{\alpha \sim j} \frac{f_\alpha^{(i)}}{d_\alpha} \tag{1}$$

where the summation is over all α connected with j ($\alpha \sim j$). In the second step, the resource propagates along the links again in the same way, yielding the resource vector $h_\alpha^{(i)}$ in the form

$$h_\alpha^{(i)} = \sum_{j \sim \alpha} \frac{g_j^{(i)}}{k_j} \tag{2}$$

where the sum is now over all j connected with α. Score $h_\alpha^{(i)}$ can be then interpreted as the recommendation score of node α when computing recommendation for node i. We interpret it here as the link prediction score $s_{i\alpha}$.

2.1.4. Number of Violations of the Nestedness Property (NViol)

In [35], network nestedness was quantified using the concept of violations of the nestedness property. One first defines nodes i and j as pairwise nested if the following holds: $k_j \leq k_i : \Gamma_j \subset \Gamma_i$ (note that we adapt here the original notation to bipartite networks). In other words, two nodes are nested if the neighborhood of the higher degree node includes the neighborhood of the other node. Pairwise nestedness can be analogously formulated for nodes from the other set: $d_\beta \leq d_\alpha : \Phi_\beta \subset \Phi_\alpha$. In [35], they proceed by introducing the number of violations of the nestedness property, which is defined as the number of neighbors of the lower degree node that *are not* neighbors of the higher degree node: $V(i,j) := \Theta(k_i - k_j) \sum_\alpha B_{j\alpha}(1 - B_{i\alpha})$ where the $\Theta(x)$ is one for $x \geq 0$ and zero otherwise. The total number of violations in the network, V, is obtained by summing $V(i,j)$ over all node pairs (i,j).

In social networks where clustering coefficient is high and network links form many triangles, effective link predictions can be obtained [2] by the number of common neighbors which is equal to how many new triangles a new link introduces in the network. Having introduced a measure of network nestedness, we can reason analogously: In a nested network, links that would decrease the number of nestedness violations most are the most likely candidates for missing links. We thus compute the score of link (i,α) as

$$V_{(i,\alpha)} - V \tag{3}$$

where V is the network's original number of nestedness violations, and $V_{(i,\alpha)}$ is the number of violations *after* link (i,α) is added. Since V is the same for all links in a network, we consider directly $V_{(i,\alpha)}$ as a link prediction score. Differently from the other link prediction indices, here the lower the better.

Since the value of V changes when the biadjacency matrix is transposed, we evaluate NViol on both the original and the transposed network, and choose the approach that yields the best results in terms of AUC (of course, one is free to choose a different criterion).

2.2. Evaluation Process

Unlike [43], which considered link prediction in time-stamped country-product data, we do not aim to evaluate the prediction of *future links* in nested networks, because the notion of time is typically not defined in ecological nested networks. Instead, we adopt the usual setting of link prediction in networks without time information, where a small fraction of input data are moved in the probe and the predicted links are then compared with the probe links. In an input network with E links, $E_P := f_P E$ links are moved in the probe and the remaining $E_T := E - E_P$ links comprise training data that are used for prediction; we use $f_P = 0.1$ here. We remark, though, that the choice of the probe set by timestamps of the network's links still remains the preferable option in systems where time matters because it exposes link prediction methods to double difficulty of capturing both the systems' structural patterns, as well as their dynamical patterns [44].

Of the methods included in the evaluation, NViol is the only that is not parameter-free: It has one binary parameter that determines whether it ultimately acts on the original or transposed data. While methods with parameters are best to be evaluated using a triple training-learning-probe division [42] where the additional *learning* set is used to determine the optimal parameter values, we use for simplicity the training-probe division described above because the "minimal" parametrization of NViol is negligible in comparison with the size of the input data. The situation is very different for methods whose number of parameters scales with the input data size (as is the case for the popular matrix factorization recommendation methods [45]) where the simple training-probe evaluation typically substantially overestimates the method's actual performance.

An obtained ranking of the $N_1N_2 - E_T$ links that are not present in the data can be compared with the probe data in various ways [4]. We use four common link performance metrics: Ranking score, AUC, precision, and F_1 score. To compute the ranking score, r, the rank of all probe links in the link prediction list is averaged and subsequently normalized by the list length $N_1N_2 - E_T$. Rank score thus lies between zero and one (the lower, the better: small ranking score indicates that the probe items are placed high in the link prediction list). The second metric, AUC, is based on constructing a curve in the unit square $[0,1] \times [0,1]$ that corresponds to gradually following the link prediction list from the top (links with the highest score) to the bottom. The coordinates $[x, y]$ of a point on the curve then correspond to the fraction of non-probe and probe-links, respectively, recovered so far. The final AUC value is then obtained by computing the area under thus-constructed curve [46]. AUC lies in the range $[0,1]$; the higher the value, the better the performance. Note that the ranking score and AUC of a random link prediction list are 0.5. Any result below (in the case of r) or above (in the case of AUC) this value thus indicates that link prediction is better than random. The third metric, precision, focuses on the top L ranks of the link prediction list. If there are $n(L)$ probe links in the top L ranks, we say that the link prediction precision is $P(L) := n(L)/L$. Precision too lies in the range $[0,1]$ (the higher, the better: precision of one is achieved if all top L ranks are occupied by probe links). To evaluate precision, we use $L = 100$. Precision has a closely related counterpart, recall, which is defined as $R(L) := n(L)/E_P$. The two metrics are combined together in our fourth metric, the F_1 score, which is defined as the harmonic mean of the observed precision and recall. To make this metric parameter-free, we report the maximum F_1 score with respect to the number of top ranks L included in performance evaluation.

To factor out the randomness of the division into training and probe data, we repeat the evaluation process for 100 random training-probe divisions (in synthetic data, we use 10 independent model realizations and 10 training-probe divisions for each of them), and report the mean and the standard error of the mean for the three chosen evaluation metrics.

3. Data

3.1. Synthetic Data

To create synthetic nested networks, we adapt to a bipartite setting the construction that was used in [36] to test the detection of simultaneous modular and nested structure in data. We first present this construction in the case without modular structure. The first step is to establish a perfectly nested biadjacency matrix where no nestedness violations are present. To this end, one introduces the contour

$$y = 1 - (1 - x^{1/\xi})^{\xi} \tag{4}$$

in the unit square $[0,1] \times [0,1]$ (see Figure 2); the unit square can be then mapped onto the $N_1 \times N_2$ biadjacency matrix by setting $y := 1 - i/N_1$ (in this way we follow the usual matrix notation and the first row corresponding to $i = 1$ maps onto the top part of the unit square) and $x := \alpha/N_2$. All biadjacency matrix elements "above" the contour are then set to one, and the remaining elements are to zero. As ξ increases, the degree distribution corresponding to the such-created biadjacency matrix becomes more heterogeneous and the network becomes more sparse . While the resulting biadjacency matrix is nested for any $\xi > 0$, the range $\xi \in [1,5]$ is considered in [36]. After creating a perfectly nested network, noise is introduced by moving each of the initial links with probability p to a random node pair (i, α) that is not yet connected by a link. When $p = 1$, the initial nested structure completely disappears and the resulting network is fully random.

The above-described nested networks with noise are in [36] further generalized to *in-block nested* networks where the network consists of N_B blocks whose internal connections are more dense than the connections between different blocks (hence the network has a community structure [47]), and each block separately has a nested structure. In the model, this is achieved by forming perfectly nested blocks, adding intrablock noise as described above, and finally adding interblock noise by moving

each link with probability $\mu(N_B - 1)/N_B$ to a randomly chosen node from the original block and a randomly chosen node from another block. When $p = \mu = 0$, we obtain unperturbed nested structure in each block and the blocks are mutually disconnected. When $\mu = 1$, then the density of links between the blocks is the same as the density of links within each block (i.e., the community structure vanishes). Synthetic data for various model settings are shown in Figure 2.

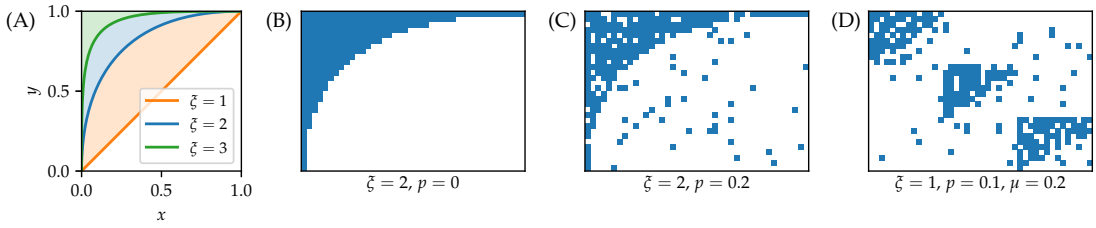

Figure 2. An illustration of synthetic nested networks. (**A**) Nestedness contours for various values of the ξ parameter. (**B**–**D**) Nested networks with $N_1 = 30$ and $N_2 = 42$: Perfectly nested network (**B**), nested network with low noise (**C**), and in-block nested network with three blocks and low noise (parameter values are specified in the panels). The results are averaged over 10 model realizations and 10 independently chosen probe sets for each realization.

3.2. Real Data

We use six datasets from the ecological networks database http://www.web-of-life.es/ (from the available datasets, all with at least 1000 links are used). These datasets are referred to by their original identifiers (M_SD_022, M_PL_021, M_PL_044, M_PL_015, M_PL_057, and M_PL_062). The five datasets whose identifiers contain PL capture plant-pollinator interactions. M_SD_022 captures fruit-frugivores interactions in two bird communities in south-eastern Brazil.

We further use country-product datasets from two different years (2001 and 2009). These datasets are obtained from the detailed data on the export volumes (measured in dollars) of various categories of products by individual countries (the data has been obtained from https://atlas.media.mit.edu/en/resources/data/). In the first step, the total export of each product category for each country is computed. To represent these data in binary form (a country either exports a product or not), the weighted country-product links are typically filtered [23,26,27] using the concept of *Revealed Comparative Advantage* (RCA) which quantifies how much a country exports a product with respect to the country's total export and the product's total export. All country-product pairs with RCA above one are represented as links in the resulting bipartite network (the RCA of one indicates that the country exports as much as one would expect from the country size and the overall export of the given product). The resulting datasets are labeled CP-2001 and CP-2009 (CP stands for country-product, the number denotes the export data year).

Basic statistical properties of the datasets are summarized in Table 1. To quantify the nested structure of the datasets, we use the density of nestedness property violations ϱ_V introduced in [35]. This quantity is obtained by diving the actual number of nestedness property violations in the network, V, with the maximum number of violations V_m where V_m is obtained by summing $V_m(i,j) := \Theta(k_i - k_j)\min(k_j, N_2 - k_i)$ over all node pairs (i,j). To evaluate whether a network has a nested structure, we compare the obtained ϱ_V value with the mean $\mu(\varrho_V)$ and the standard deviation $\sigma(\varrho_V)$ of ϱ_V observed on the network randomized using the classical Configuration Model [37]. The difference $\varrho_V - \mu(\varrho_V)$ is negative for all eight real networks (meaning that the original networks are more nested—as measured by ϱ_V—than their randomized counterparts), and the z-score $[\varrho_V - \mu(\varrho_V)]/\sigma(\varrho_V)$ shows the statistical significance (at $p < 0.05$) for six out of eight networks (all except for M_PL_044 and M_PL_062). Note, however, that the observed statistical significance is not indicative of whether these networks are suitable candidates for link prediction with NVIol that assumes nested structure in the data. The

Country-Product networks, for example, are deemed significantly nested but their values ϱ_V and $\mu(\varrho_V)$ are actually close to each other (they differ by 0.0003).

Table 1. Basic properties of the real datasets used to evaluate link prediction methods: Number of rows (N_1), columns (N_2), edges (E), and the density of edges [$\varrho_E := E/(N_1 N_2)$].

Dataset	N_1	N_2	E	ϱ_E
M_SD_022	207	110	1121	0.05
M_PL_015	131	666	2933	0.03
M_PL_021	91	677	1193	0.02
M_PL_044	110	609	1125	0.02
M_PL_057	114	883	1920	0.02
M_PL_062	456	1044	15,255	0.03
CP-2001	169	781	17,639	0.13
CP-2009	168	774	17,739	0.14

4. Results

Results on model data with no community structure ($B = 1$) are shown in Figure 3. There is a number of points to note:

1. As ξ grows and the networks' nested structure thus becomes more pronounced, differences between the methods grow.
2. NViol is generally the best-performing method with respect to the metrics r and AUC that take the whole link prediction list into account. Upon a closer inspection of link prediction lists produced by the respective methods, the advantage of NViol is due to its ability to place well also links connecting low-degree nodes that the other methods miss due to their general bias towards high-degree nodes. With NViol, though, probe links adjacent to low-degree nodes are not among the top 100 and hence do not contribute to the method's precision, yet they rank much better than where other methods are used. If we would increase the number of top ranks included in precision evaluation from 100 to 200 or 300, NViol would have an edge also in this metric.
3. As the randomization parameter p grows, PrefA eventually outperforms NViol in terms of link prediction precision. High precision improvement with respect to LCL's precision for $\xi = 5$ are due to the generally low precision achieved for the sparse networks produced at $\xi = 5$ (which is made further worse by introducing the noise when $p > 0$).
4. ProbS outperforms LCL but lacks behind PrefA and NViol. This is expected because ProbS is based on a "personalized" recommendation algorithm; with no communities in the data, there is no place for personalization and thus ProbS's merits cannot manifest themselves. The situation becomes radically different when there is more than one nested block in the data (see below).

To illustrate how the relative performance of the methods changes when there is more than one nested block in the data, we choose the simplest case with two blocks and no links between the blocks ($N_B = 2$, $\mu = 0$). As can be seen in Figure 4, the relative performance of PrefA and NViol lowers with respect to the one-block case, in particular in terms of precision which is now worse than the precision achieved by LCL. While NViol remains the best method in terms of the ranking score and AUC, ProbS is clearly best in terms of precision. The reason for the worsening performance of PrefA and NViol is that they both ignore the block structure of the network. If, for example, nodes i and α have high degree, PrefA will assign them high score even if they belong to different blocks. By contrast, ProbS and LCL explore the network locally, and thus naturally obey the blocks' boundaries. To help PrefA and NViol overcome the challenge poised by the presence of multiple blocks, community detection [47] could be first applied to detect blocks in the input network. The very recent method for detecting the structure of nested networks with block-wise nested structure is particularly relevant in this respect [36].

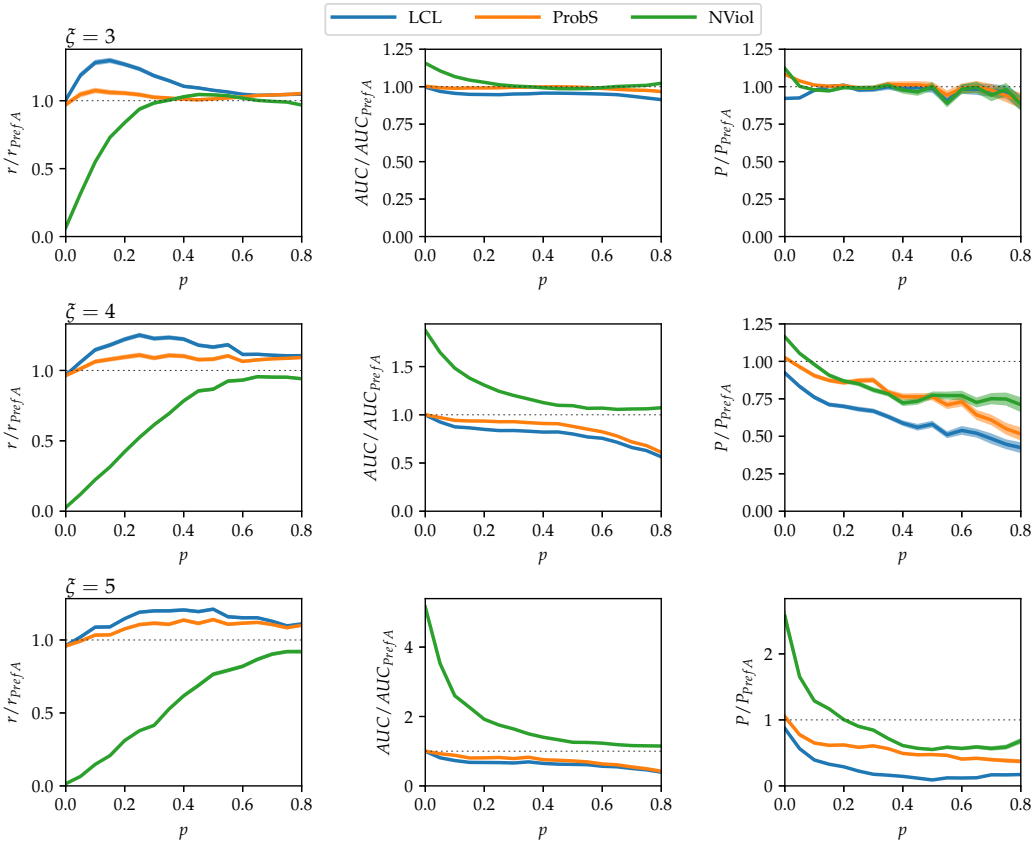

Figure 3. Link prediction results on model data with $N_1 = 100$, $N_2 = 200$, and no community structure. To remove the strong dependency of method performance on the randomization parameter p, the shown results are scaled with the results of the simplest PrefA method.

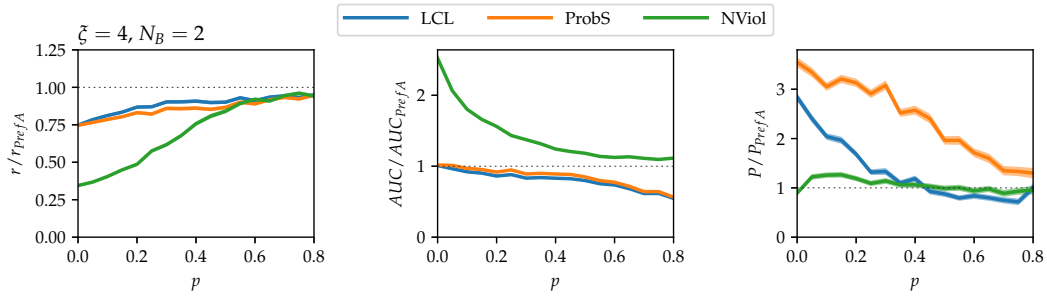

Figure 4. Link prediction results on model data with $N_1 = 100$, $N_2 = 200$, $N_B = 2$, $\mu = 0$ (two blocks, no links between the blocks). As in Figure 3, results are again scaled with the results of the simplest PrefA method.

Table 2 summarizes the performance of the evaluated methods on the chosen real datasets. In terms of precision, LCL and ProbS are the best methods. Similarly to Figure 3, global ranking metrics r and AUC show a more diverse pattern with NViol being the best method by a large margin in three datasets (ProbS is the best in the remaining ones). The performance gap between the link prediction methods that do not consider the local network structure (PrefA and NViol) and the network neighborhood-based methods (LCL and ProbS) in some networks (M_PL_015 and M_PL_062) suggest that these networks have a more pronounced block structure which puts the PrefA and NViol in disadvantage. Similarly to Figure 4, link prediction with NViol could be helped in these cases by combining it with detection of the block structure.

Table 2. Mean link prediction results on real datasets. Best performance values for a given method and metric are highlighted with bold. Results are averaged over 100 independently chosen probe sets. Standard error of the mean is less than 0.005 in all cases. If NViol produces best AUC on transposed data, it is labeled as NViol$^{\top}$.

M_SD_022

method	r	AUC	P	F_1
PrefA	0.20	0.80	0.11	0.13
LCL	0.19	0.80	0.14	0.15
ProbS	**0.17**	**0.82**	**0.15**	**0.16**
NViol$^{\top}$	0.19	0.81	0.11	0.12

M_PL_057

method	r	AUC	P	F_1
PrefA	0.38	0.61	0.12	0.10
LCL	0.34	0.59	0.14	0.12
ProbS	0.33	0.61	**0.16**	**0.13**
NViol	**0.16**	**0.84**	0.12	0.10

M_PL_015

method	r	AUC	P	F_1
PrefA	0.23	0.77	0.12	0.09
LCL	0.19	0.80	0.20	0.14
ProbS	**0.18**	**0.82**	**0.25**	**0.18**
NViol	0.24	0.75	0.13	0.08

M_PL_062

method	r	AUC	P	F_1
PrefA	0.20	0.80	0.05	0.05
LCL	0.17	0.83	0.12	0.07
ProbS	**0.16**	**0.84**	**0.15**	**0.08**
NViol$^{\top}$	0.20	0.80	0.04	0.05

M_PL_021

method	r	AUC	P	F_1
PrefA	0.49	0.49	0.06	0.07
LCL	0.41	0.46	**0.09**	0.09
ProbS	0.40	0.48	**0.09**	**0.10**
NViol	**0.16**	**0.84**	0.06	0.05

CP-2001

method	r	AUC	P	F_1
PrefA	0.23	0.78	0.05	0.10
LCL	0.21	0.79	**0.18**	**0.13**
ProbS	**0.19**	**0.81**	0.14	0.12
NViol	0.24	0.76	0.06	0.10

M_PL_044

method	r	AUC	P	F_1
PrefA	0.47	0.52	0.05	0.06
LCL	0.38	0.45	**0.07**	**0.07**
ProbS	0.37	0.46	0.06	**0.07**
NViol	**0.21**	**0.79**	0.04	0.05

CP-2009

method	r	AUC	P	F_1
PrefA	0.24	0.77	0.07	0.10
LCL	0.22	0.79	**0.22**	**0.13**
ProbS	**0.20**	**0.80**	0.13	0.12
NViol	0.25	0.75	0.08	0.10

5. Discussion

To summarize, our paper compared the performance of existing link prediction techniques to identify missing links in networks that exhibit nestedness at both macroscopic [14] and mesoscopic [36] scale. For both kinds of structures, we found that a method that exploits the nested structure of such systems (NViol) outperforms existing methods not only for perfectly nested structures, but also for imperfect nested structures up to a certain level of departure from perfect nestedness. The new NViol method performs well also on some real nested datasets where it is the second most successful method after ProbS.

One of the challenges raised by our work is how to best combine the detection of blocks that exhibit an internal nested structure [36] with link prediction. By definition, no links would be predicted between the blocks if the blocks were considered in isolation; however, such rigid assumption is likely to be suboptimal for "mixed" topologies where a given amount of inter-block links exist. Developing a well-performing link-prediction that takes into account this aspect is a challenge for future research.

More generally, our work shows that if we know that a given system exhibit a given structural pattern, we can exploit this information to design a competitive link-prediction techniques. In this respect, we envision that optimal methods for link prediction might be based on a two-step process: First, we learn the topology of the network in hand; second, we adopt a prediction method that is optimal for the detected topology. Once fully developed, such a link prediction method could help us understand how best to measure the degree to which a network is nested. As discussed

in Section 3.2, the detection of a statistically significant structural pattern does not necessarily have practical implications. By contrast, whether or not a nestedness-based link prediction method can outperform other benchmark link prediction methods is a very practical way of assessing the value of a network's nested structure. Work in this direction can help us to better understand not only the nested structure of networks, but also other specific network topologies such as the core-periphery structure, for example.

Author Contributions: Conceptualization, M.M., M.S.M. and L.L.; Methodology, M.M. and L.L.; Analyzing data and results, M.M., M.S.M. and L.L.; Visualization, M.M.; Writing—original draft, M.M.; Writing—review & editing, M.M., M.S.M. and L.L. We acknowledge Zhuoming Ren for the help in obtaining the world trade data.

Funding: This work is supported by the National Natural Science Foundation of China (Nos. 11622538, 61673150), and the Zhejiang Provincial Natural Science Foundation of China (No. LR16A050001). M.S.M. acknowledges financial support from the University of Zurich through the URPP Social Networks.

Acknowledgments: We acknowledge the Science Strength Promotion Programme of UESTC, Chengdu.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Getoor, L.; Diehl, C.P. Link mining: A survey. *ACM SIGKDD Explor. Newsl.* **2005**, *7*, 3–12.
2. Liben-Nowell, D.; Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Tech.* **2007**, *58*, 1019–1031.
3. Guimerà, R.; Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 22073–22078.
4. Lü, L.; Zhou, T. Link prediction in complex networks: A survey. *Physical A* **2011**, *390*, 1150–1170.
5. Warde-Farley, D.; Donaldson, S.L.; Comes, O.; Zuberi, K.; Badrawi, R.; Chao, P.; Franz, M.; Grouios, C.; Kazi, F.; Lopes, C.T.; et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **2010**, *38*, W214–W220.
6. Holme, P.; Saramäki, J. Temporal networks. *Phys. Rep.* **2012**, *519*, 97–125.
7. Liao, H.; Mariani, M.S.; Medo, M.; Zhang, Y.C.; Zhou, M.Y. Ranking in evolving complex networks. *Phys. Rep.* **2017**, *689*, 1–54.
8. Lü, L.; Pan, L.; Zhou, T.; Zhang, Y.C.; Stanley, H.E. Toward link predictability of complex networks. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 2325–2330.
9. Al Hasan, M.; Zaki, M.J. A survey of link prediction in social networks. In *Social Network Data Analytics*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 243–275.
10. Ren, Z.M.; Zeng, A.; Zhang, Y.C. Structure-oriented prediction in complex networks. *Phys. Rep.* **2018**, *750*, 1–51.
11. Squartini, T.; Caldarelli, G.; Cimini, G.; Gabrielli, A.; Garlaschelli, D. Reconstruction methods for networks: The case of economic and financial systems. *arXiv* **2018**, arXiv:1806.06941.
12. Patterson, B.D.; Atmar, W. Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biol. J. Linn. Soc.* **1986**, *28*, 65–82.
13. Bascompte, J.; Jordano, P.; Melián, C.J.; Olesen, J.M. The nested assembly of plant–animal mutualistic networks. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 9383–9387.
14. Ulrich, W.; Almeida-Neto, M.; Gotelli, N.J. A consumer's guide to nestedness analysis. *Oikos* **2009**, *118*, 3–17.
15. König, M.D.; Tessone, C.J. Network evolution based on centrality. *Phys. Rev. E* **2011**, *84*, 056108.
16. Suweis, S.; Simini, F.; Banavar, J.R.; Maritan, A. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature* **2013**, *500*, 449.
17. König, M.D.; Tessone, C.J.; Zenou, Y. Nestedness in networks: A theoretical model and some applications. *Theor. Econ.* **2014**, *9*, 695–752.
18. Valverde, S.; Piñero, J.; Corominas-Murtra, B.; Montoya, J.; Joppa, L.; Solé, R. The architecture of mutualistic networks as an evolutionary spandrel. *Nat. Ecol. Revolut.* **2018**, *2*, 94–99.
19. 'Bastolla, U.; Fortuna, M.A.; Pascual-García, A.; Ferrera, A.; Luque, B.; Bascompte, J. The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature* **2009**, *458*, 1018–1020.
20. Allesina, S.; Tang, S. Stability criteria for complex ecosystems. *Nature* **2012**, *483*, 205–208.

21. Rohr, R.P.; Saavedra, S.; Bascompte, J. On the structural stability of mutualistic systems. *Science* **2014**, *345*, 1253497.
22. Montoya, J.M.; Pimm, S.L.; Solé, R.V. Ecological networks and their fragility. *Nature* **2006**, *442*, 259–264.
23. Tacchella, A.; Cristelli, M.; Caldarelli, G.; Gabrielli, A.; Pietronero, L. A new metrics for countries' fitness and products' complexity. *Sci. Rep.* **2012**, *2*, 723.
24. Maron, M.; Mac Nally, R.; M. Watson, D.; Lill, A. Can the biotic nestedness matrix be used predictively? *Oikos* **2004**, *106*, 433–444.
25. Bustos, S.; Gomez, C.; Hausmann, R.; Hidalgo, C.A. The dynamics of nestedness predicts the evolution of industrial ecosystems. *PLoS ONE* **2012**, *7*, e49393.
26. Tacchella, A.; Mazzilli, D.; Pietronero, L. A dynamical systems approach to gross domestic product forecasting. *Nat. Phys.* **2018**, *14*, 861–865.
27. Cristelli, M.; Tacchella, A.; Pietronero, L. The heterogeneous dynamics of economic complexity. *PLoS ONE* **2015**, *10*, e0117174.
28. Battiston, F.; Cristelli, M.; Tacchella, A.; Pietronero, L. How metrics for economic complexity are affected by noise. *Complex. Econ.* **2014**, *3*, 1–22.
29. Mariani, M.S.; Vidmer, A.; Medo, M.; Zhang, Y.C. Measuring economic complexity of countries and products: which metric to use? *Eur. Phys. J. B* **2015**, *88*, 293.
30. Wu, R.J.; Shi, G.Y.; Zhang, Y.C.; Mariani, M.S. The mathematics of non-linear metrics for nested networks. *Phys. A Stat. Mech. Appl.* **2016**, *460*, 254–269.
31. Olesen, J.M.; Bascompte, J.; Dupont, Y.L.; Elberling, H.; Rasmussen, C.; Jordano, P. Missing and forbidden links in mutualistic networks. *Proc. R. Soc. Lond. B Biol. Sci.* **2010**, doi:10.1098/rspb.2010.1371.
32. Bascompte, J.; Jordano, P. *Mutualistic Networks*; Princeton University Press: Princeton, NJ, USA, 2013.
33. Vázquez, D.P.; Poulin, R.; Krasnov, B.R.; Shenbrot, G.I. Species abundance and the distribution of specialization in host–parasite interaction networks. *J. Anim. Ecol.* **2005**, *74*, 946–955.
34. Nielsen, A.; Bascompte, J. Ecological networks, nestedness and sampling effort. *J. Ecol.* **2007**, *95*, 1134–1141.
35. Grimm, A.; Tessone, C.J. Analysing the sensitivity of nestedness detection methods. *Appl. Netw. Sci.* **2017**, *2*, 37.
36. Solé-Ribalta, A.; Tessone, C.J.; Mariani, M.S.; Borge-Holthoefer, J. Revealing in-block nestedness: Detection and benchmarking. *Phys. Rev. E* **2018**, *97*, 062302.
37. Newman, M. *Networks: An Introduction*; Oxford University Press: Oxford, UK, 2010.
38. Kunegis, J.; De Luca, E.W.; Albayrak, S. The link prediction problem in bipartite networks. In Proceedings of the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, Dortmund, Germany, 28 June–2 July 2010; Springer: Berlin/Heidelberg, Germany, 2010, pp. 380–389.
39. Daminelli, S.; Thomas, J.M.; Durán, C.; Cannistraci, C.V. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New J. Phys.* **2015**, *17*, 113037.
40. Zhou, T.; Ren, J.; Medo, M.; Zhang, Y.C. Bipartite network projection and personal recommendation. *Phys. Rev. E* **2007**, *76*, 046115.
41. Lü, L.; Medo, M.; Chi, H.Y.; Zhang, Y.C.; Zhang, Z.K.; Zhou, T. Recommender systems. *Phys. Rep.* **2012**, *519*, 1–49.
42. Yu, F.; Zeng, A.; Gillard, S.; Medo, M. Network-based recommendation algorithms: A review. *Phys. A Stat. Mech. Appl.* **2016**, *452*, 192–208.
43. Vidmer, A.; Zeng, A.; Medo, M.; Zhang, Y.C. Prediction in complex systems: The case of the international trade network. *Phys. A Stat. Mech. Appl.* **2015**, *436*, 188–199.
44. Vidmer, A.; Medo, M. The essential role of time in network-based recommendation. *EPL* **2016**, *116*, 30007.
45. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *8*, 30–37.
46. Swets, J.A. Information retrieval systems. *Science* **1963**, *141*, 245–250.
47. Fortunato, S.; Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **2016**, *659*, 1–44.