



HAL
open science

FDR control and FDP bounds for conformal link prediction

Gilles Blanchard, Guillermo Durand, Ariane Marandon-Carlhian, Romain Périer

► **To cite this version:**

Gilles Blanchard, Guillermo Durand, Ariane Marandon-Carlhian, Romain Périer. FDR control and FDP bounds for conformal link prediction. 2024. hal-04529648

HAL Id: hal-04529648

<https://hal.science/hal-04529648v1>

Preprint submitted on 2 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FDR control and FDP bounds for conformal link prediction

Gilles Blanchard¹, Guillermo Durand¹, Ariane Marandon², Romain Périer¹

¹ LMO, Université Paris-Saclay, gilles.blanchard@universite-paris-saclay.fr,
guillermo.durand@universite-paris-saclay.fr, romain.perier@universite-paris-saclay.fr

² Turing Institute, amarandon-carlhian@turing.ac.uk

Abstract

In Marandon (2023), the author introduces a procedure to detect true edges from a partially observed graph using a conformal prediction fashion: first computing scores from a trained function, deriving conformal p -values from them and finally applying a multiple testing procedure. In this paper, we prove that the resulting procedure indeed controls the FDR, and we also derive uniform FDP bounds, thanks to an exchangeability argument and the previous work of Marandon et al. (2022).

1 Introduction

Consider a link prediction setting where we partially observe a graph, that is, the presence or absence of an edge (borrowing to multiple testing literature, we will use the terms “true” and “false” edge, respectively) between two nodes is either observed, either masked (Lu and Zhou, 2011). Furthermore, consider that the missing data mask is known. Hence, the absence of connection between two nodes is either an observed false edge, either a masked edge (and the state of the edge, true or false, is unobserved), and we know if we are in the first or the second case. The task at hand is to provide a label (true or false) to each unobserved edge. In Marandon (2023), the author introduced the first, to our knowledge, method which aims to control the False Discovery Rate (FDR) in the given context, using tools from the Conformal Inference (CI) literature, namely conformal p -values (Vovk et al., 2005). However, because of the graph structure inducing intricate dependence in the data, the setup is different from the usual case encountered in CI: typically, in CI, the data is assumed to be exchangeable. Hence, theoretical FDR control was not proven, but was observed empirically on simulated data.

In this work, we prove that the procedure of Marandon (2023) controls the FDR under a general missing data setup and without any distributional assumptions on the complete network. The core observation to make is that even if the data is not exchangeable, the scores are, under really mild assumptions (no ties among them, almost surely). This allows to use the results of Marandon et al. (2022) to achieve FDR control. Furthermore, the exchangeability of the scores also allows to use Gazin et al. (2023) to get a uniform False Discovery Proportion (FDP) bound for the path $(R(t))_{t \in [0,1]}$, where $R(t)$ is the set of unobserved edges with conformal p -value lesser or equal than t . This uniform bound in turn allows to construct confidence post hoc bounds à la Blanchard et al. (2020).

For an extensive review of related works in link prediction, conformal inference, and multiple testing, we refer the reader to Marandon (2023, Sections 1.3, 3.1, 3.3).

The rest of the paper is organized as follows. In Section 2, we first define formally the model, the problem and the notation, before recalling the conformal link prediction procedure introduced in Marandon (2023). Then, in Section 3, we state and prove our two main theorems. We conclude in Section 4 with a short discussion of the relation to recent literature, in particular the work of Huang et al. (2023).

2 Conformal link prediction

2.1 Model and notation

Let $A^* = (A_{i,j}^*)_{1 \leq i,j \leq n}$ be the adjacency matrix of the true complete graph \mathcal{G} , $X \in \mathbb{R}^{n \times d}$ a matrix of node covariates (if available), and $\Omega = (\Omega_{i,j})_{1 \leq i,j \leq n}$ the sampling matrix such that $\Omega_{i,j} = 1$ if the interaction status (true/false) of (i, j) is observed, and 0 otherwise. We denote by A the observed adjacency matrix with $A_{i,j} = \Omega_{i,j} A_{i,j}^*$. Thus, $A_{i,j} = 1$ indicates that there is an observed true edge between i and j , whereas $A_{i,j} = 0$ indicates either the observed lack of an edge or an unreported edge. The sampling matrix Ω is assumed to be observed, so that it is known which zero-entries $A_{i,j} = 0$ correspond to observed false edges and which ones correspond to missing information. We denote by P the joint distribution of $Z^* = (A^*, X, \Omega)$, Z the observation (A, X, Ω) and \mathcal{Z} the observation space. P belongs to a model \mathcal{P} . We consider a general missing data setup in which the entries of Ω are independently generated as:

$$\Omega_{i,j} | A^*, X \sim \mathcal{B}(w_0 \mathbb{1}_{A_{i,j}^*=0} + w_1 \mathbb{1}_{A_{i,j}^*=1}), \quad 1 \leq i, j \leq n$$

for some unknown sampling rates w_0, w_1 . This type of missing data setup is called double standard sampling (Tabouy et al., 2020; Sportisse et al., 2020) and is a generalization of the case where the entries of Ω are i.i.d, that is more relevant for practical applications (see Section 2 in Marandon (2023) and the references therein for a discussion of this assumption with regards to the literature). Let us introduce the following notations:

- We denote by $\mathcal{D}_{\text{test}}(Z) = \{(i, j) : \Omega_{i,j} = 0\}$ the set of non-sampled (or missing) node pairs and by $\mathcal{D}(Z) = \{(i, j) : \Omega_{i,j} = 1\}$ the set of sampled pairs, with $\mathcal{D}^0 = \{(i, j) \in \mathcal{D} : A_{i,j}^* = 0\}$ the set of observed false edges and $\mathcal{D}^1 = \{(i, j) \in \mathcal{D} : A_{i,j}^* = 1\}$ the set of observed true edges. We refer to $\mathcal{D}_{\text{test}}(Z)$ as the test set.
- We denote by $\mathcal{H}_0 = \{(i, j) : \Omega_{i,j} = 0, A_{i,j}^* = 0\}$ the (unobserved) set of false edges in the test set and $\mathcal{H}_1 = \{(i, j) : \Omega_{i,j} = 0, A_{i,j}^* = 1\}$ the (unobserved) set of true edges in the test set.

In link prediction, one is interested in classifying the unobserved node pairs of $\mathcal{D}_{\text{test}}$ into true edges and false edges, or in other words, selecting a set of unobserved node pairs to be declared as true edges, based on the observed graph structure. Define a selection procedure as a (measurable) function $R = R(Z)$ that returns a subset of $\mathcal{D}_{\text{test}}$ corresponding to the indices (i, j) where an edge is declared. For any such procedure R , the False Discovery Rate (FDR) of R is defined as the average of the False Discovery Proportion (FDP) of R under the model parameter $P \in \mathcal{P}$, that is,

$$\text{FDR}(R) = \mathbb{E}_{Z^* \sim P}[\text{FDP}(R)], \quad \text{FDP}(R) = \frac{\sum_{i \in \mathcal{H}_0} \mathbb{1}_{i \in R}}{1 \vee |R|}.$$

Similarly, the true discovery rate (TDR) is defined as the average of the true discovery proportion (TDP), that is,

$$\text{TDR}(R) = \mathbb{E}_{Z^* \sim P}[\text{TDP}(R)], \quad \text{TDP}(R) = \frac{\sum_{i \in \mathcal{H}_1} \mathbb{1}_{i \in R}}{1 \vee |\mathcal{H}_1|}.$$

The aim considered here is to build a procedure R that controls the FDR while having a TDR (measuring the *power* of the procedure) as large as possible.

Notably, a well-known procedure controlling the FDR is the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) which, given m p -values p_1, \dots, p_m , sorts them: $0 = p_{(0)} \leq p_{(1)} \leq \dots \leq p_{(m)}$, defines $\hat{k} = \max \{k \in \llbracket 0, m \rrbracket : p_{(k)} \leq \alpha \frac{k}{m}\}$, and rejects the p -values smaller than $\alpha \frac{\hat{k}}{m}$.

Let us emphasize that, in our setting, \mathcal{H}_0 is random, which is not entirely classical in the multiple testing literature, but it is also not unusual, and can be traced back to, at least, the mixture model presented in Efron et al. (2001).

2.2 Conformal link prediction procedure

We briefly recall the procedure which is described in details in Marandon (2023, Section 3.2). Assume that we have at hand a score estimator function $g : \mathcal{Z} \rightarrow \mathbb{R}^{n \times n}$ that takes as input an observation $z \in \mathcal{Z}$ and returns a score matrix $(S_{i,j})_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$ that scores the likeliness of a true edge between couples of nodes. The function g corresponds to a link prediction algorithm that would be trained on the ‘‘training’’ dataset z . Any off-the-shelf link prediction algorithm can be used, up to a specification detailed in the next paragraph. Our training set has to be different from a ‘‘calibration’’ set on which we will compute scores and compare them to the scores computed on the test set to build the conformal p -values. We build this calibration set \mathcal{D}_{cal} of false edges by sampling uniformly $\ell = |\mathcal{D}_{\text{cal}}|$ edges without replacement from the set of observed false edges \mathcal{D}^0 . The algorithm is then trained on $Z_{\text{train}} = (A, X, \Omega_{\text{train}})$ where $(\Omega_{\text{train}})_{i,j} = 0$ if $(i, j) \in \mathcal{D}_{\text{cal}}$ and $\Omega_{i,j}$ otherwise. That is, formally, we simply compute $g(Z_{\text{train}})$. The scores $S_{i,j} = g(Z_{\text{train}})_{i,j}$ are then computed for all $(i, j) \in \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$. Conformal p -values $(p_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}}$ are then computed following

$$p_{i,j} = \frac{1}{\ell + 1} \left(1 + \sum_{(u,v) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S_{i,j} \leq S_{u,v}\}} \right), \quad (i, j) \in \mathcal{D}_{\text{test}} \quad (1)$$

and finally the BH procedure is applied to them. Those steps are recapped in Algorithm 1.

Although g can be any measurable function, we now discuss a general shape of g mild and commonly used. Let us introduce, for a given $K \in \{1, \dots, n\}$, and any adjacency matrix a ,

$$w(a)_{i,j} = (a_{i,\bullet}, a_{j,\bullet}, a_{i,\bullet}^2, a_{j,\bullet}^2, \dots, a_{i,\bullet}^K, a_{j,\bullet}^K),$$

where $a_{i,\bullet}^k = (a_{i,u}^k)_{1 \leq u \leq n}$ for $1 \leq k \leq K$, that is, $a_{i,\bullet}^k$ is the i -th row of a to the power k . For $z = (a, x, \omega) \in \mathcal{Z}$, the commonly used form for $g(z)$ is then

$$g(z)_{i,j} = h((w(a)_{i,j}, x_{N(i,K),\bullet}, x_{N(j,K),\bullet}); \{(w(a)_{u,v}, x_{N(u,K),\bullet}, x_{N(v,K),\bullet}, a_{u,v}), \omega_{u,v} = 1\}) \quad (2)$$

with $h(\cdot; \cdot)$ some real-valued measurable function, $N(i, K)$ the K -hop neighborhood of node i , and $x_{S,\bullet}$ is the submatrix $(x_{u,l})_{\substack{u \in S \\ 1 \leq l \leq d}}$ for any subset $S \in \llbracket 1, n \rrbracket$. Note that the second argument of h means exactly that g is trained only on the information provided by the non-missing edges. Also note that because $(\Omega_{\text{train}})_{i,j} = 0$ for all $(i, j) \in \mathcal{D}_{\text{cal}}$, Ω_{train} treats the edges of \mathcal{D}_{cal} as missing. Hence $g(Z_{\text{train}})$ indeed is trained on a set of observed edges that is disjoint from the edges of the calibration set, thus avoiding overfitting.

This formulation indeed is mild and encompasses numerous link prediction algorithms. For example the simple common neighbors heuristic (Lu and Zhou, 2011) given by $g(z)_{i,j} = a_{i,\bullet}^\top a_{j,\bullet}$. The larger class of Empirical Risk Minimization (ERM) methods (like in Bleakley et al. (2007) or Zhang and Chen (2018)) also generally fall into this form. Take a class \mathcal{F} of measurable functions taking values in $[0, 1]$, that can be thought of as binary classifiers. It could be a class of (graph) neural networks for example. Also let \mathcal{L} a measurable loss function,

Algorithm 1 Conformal link prediction

Input: Observation $Z = (A, X, \Omega)$: (Observed) adjacency matrix A , node covariate matrix X , sampling matrix Ω ; LP algorithm g ; sample size ℓ of the reference set

Define $\mathcal{D} = \{(i, j) : \Omega_{ij} = 1\}$; $\mathcal{D}_{\text{test}} = \{(i, j) : \Omega_{ij} = 0\}$; $\mathcal{D}^0 = \{(i, j) \in \mathcal{D} : A_{ij} = 0\}$

1. Sample \mathcal{D}_{cal} of size ℓ uniformly without replacement from \mathcal{D}^0

2. Learn g on “masked” observation $Z_{\text{train}} = (A, X, \Omega_{\text{train}})$, with Ω_{train} given by

$$(\Omega_{\text{train}})_{i,j} = \begin{cases} 0 & \text{if } (i, j) \in \mathcal{D}_{\text{cal}}, \\ \Omega_{i,j} & \text{otherwise.} \end{cases}$$

3. For each $(i, j) \in \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$, compute the score $S_{i,j} = g(Z_{\text{train}})_{i,j}$

4. For each $(i, j) \in \mathcal{D}_{\text{test}}$, compute the conformal p -value $p_{i,j}$ given by (1)

5. Apply BH using $(p_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}}$ as input p -values, providing a rejection set $R(Z) \subset \mathcal{D}_{\text{test}}$.

Output: $R(Z)$

for example the cross-entropy loss function. Then let $\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} \sum_{(u,v): \omega_{u,v}=1} \mathcal{L}(f(w(a)_{u,v}), a_{u,v})$ and finally $h(\cdot; \{(w(a)_{u,v}, a_{u,v}), \omega_{u,v} = 1\}) = \hat{f}(\cdot)$. This ERM framework indeed fits Equation (2). More examples are provided in Marandon (2023, Section 3.3).

3 FDR and FDP control guarantees

In this section, we establish that the procedure controls the FDR in finite samples and the FDP with high probability. In the sequel, we use the following notation: $k_0 = |\mathcal{D}^0|$, $m = |\mathcal{D}_{\text{test}}|$, $\mathcal{D}_{\text{train}}^0 = \mathcal{D}^0 \setminus \mathcal{D}_{\text{cal}}$.

We will consider the following assumptions:

Assumption 1. *The matrix score $g(Z_{\text{train}}) = (S_{i,j})_{1 \leq i, j \leq n}$ has no ties a.s.*

Assumption 2. *The calibration set size ℓ is a function of $k_0 = |\mathcal{D}^0|$ and \mathcal{D}^1 only.*

The main result of this section is the following:

Theorem 1. *Consider the output R of the procedure proposed in Algorithm 1 with a scoring function g of the form (2), and let Assumptions 1 and 2 hold true. Then, it holds that $\text{FDR}(R) \leq \alpha$.*

The proof relies on reformulating the proposed procedure as the one of Marandon et al. (2022) applied to a suitable ordering of the scores, which is shown to satisfy the sufficient exchangeability condition in Marandon et al. (2022) for FDR control, conditionally on a certain random variable W .

The following lemma states two key properties of the calibration scores $(S_{i,j})_{(i,j) \in \mathcal{D}_{\text{cal}}}$ and test scores $(S_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}}$.

Lemma 2. *Let W denote the r.v. given by $(k_0, A^*, \mathcal{D}^1, \mathcal{D}_{\text{train}}^0, X)$, and let Assumption 2 hold true. Then:*

(i) $(S_{i,j})_{1 \leq i, j \leq n}$ is measurable with respect to W .

(ii) Denote, for any nonempty finite set C and $N \in \mathbb{N}^*$ with $N \leq |C|$, $\Xi(C, N)$ the distribution that corresponds to sampling uniformly and without replacement an N -sized subset of

the set C . Then it holds:

$$\mathcal{D}_{cal} | W \sim \Xi(\mathcal{E}^0 \setminus \mathcal{D}_{train}^0, \ell), \quad (3)$$

$$\mathcal{H}_0 = (\mathcal{E}^0 \setminus \mathcal{D}_{train}^0) \setminus \mathcal{D}_{cal}, \quad (4)$$

$$\mathcal{D}_{test} = (\llbracket 1, n \rrbracket^2 \setminus (\mathcal{D}^1 \cup \mathcal{D}_{train}^0)) \setminus \mathcal{D}_{cal}, \quad (5)$$

where $\mathcal{E}^0 = \{(i, j) \in \llbracket 1, n \rrbracket^2 : A_{i,j}^* = 0\}$.

Proof. Point (i) comes from $(S_{i,j})_{1 \leq i, j \leq n} = g(A, X, \Omega_{train})$ as, first, g is measurable, second, the knowledge of \mathcal{D}^1 is equivalent to the knowledge of A , and third, Ω_{train} can be rewritten as

$$(\Omega_{train})_{i,j} = \begin{cases} 1 & \text{if } (i, j) \in \mathcal{D}^1 \cup \mathcal{D}_{train}^0, \\ 0 & \text{otherwise,} \end{cases}$$

which in turn entails that $(S_{i,j})_{1 \leq i, j \leq n}$ is measurable with respect to $(\mathcal{D}^1, \mathcal{D}_{train}^0, X)$.

For (ii), we reason conditionally to $(A^*, k_0, \mathcal{D}^1)$ and assume those to be fixed. Because Ω are i.i.d. on \mathcal{E}^0 conditionally to (A^*, X) , recalling $\mathcal{D}^0 = \{(i, j) \in \mathcal{E}^0 : \Omega_{i,j} = 1\}$ and recalling the definition of \mathcal{D}_{cal} as a sub-sampling without replacement of ℓ elements from \mathcal{D}^0 , it follows that

$$\begin{aligned} \mathcal{D}^0 | (A^*, k_0, \mathcal{D}^1, X) &\sim \Xi(\mathcal{E}^0, k_0); \\ \mathcal{D}_{cal} | (A^*, k_0, \mathcal{D}^1, \mathcal{D}^0, X) &\sim \Xi(\mathcal{D}^0, \ell); \\ \mathcal{D}_{train}^0 &= \mathcal{D}^0 \setminus \mathcal{D}_{cal}; \\ \mathcal{H}_0 &= \mathcal{E}^0 \setminus \mathcal{D}^0; \\ \mathcal{D}_{test} &= \llbracket 1, n \rrbracket^2 \setminus (\mathcal{D}^0 \cup \mathcal{D}^1). \end{aligned}$$

It is easy to check that the above joint distribution of $(\mathcal{D}^0, \mathcal{D}_{cal}, \mathcal{D}_{train}^0, \mathcal{D}_{test}, \mathcal{H}_0)$ conditional to $(A^*, k_0, \mathcal{D}^1)$ is equivalently represented as

$$\begin{aligned} \mathcal{D}_{train}^0 | (A^*, k_0, \mathcal{D}^1, X) &\sim \Xi(\mathcal{E}^0, k_0 - \ell); \\ \mathcal{D}_{cal} | (A^*, k_0, \mathcal{D}^1, \mathcal{D}_{train}^0, X) &\sim \Xi(\mathcal{E}^0 \setminus \mathcal{D}_{train}^0, \ell); \\ \mathcal{D}^0 &= \mathcal{D}_{cal} \cup \mathcal{D}_{train}^0; \\ \mathcal{H}_0 &= (\mathcal{E}^0 \setminus \mathcal{D}_{train}^0) \setminus \mathcal{D}_{cal}; \\ \mathcal{D}_{test} &= (\llbracket 1, n \rrbracket^2 \setminus (\mathcal{D}^1 \cup \mathcal{D}_{train}^0)) \setminus \mathcal{D}_{cal}. \end{aligned}$$

This implies the claim. □

The next lemma introduces a suitable (random) ordering of the scores in order to formally grant the sufficient score exchangeability condition of Marandon et al. (2022).

Lemma 3. *Use the same notation and assumptions as in Lemma 2, denote $m_0 = |\mathcal{H}_0|$. For given W , let σ_0 be a purely random ordering of $\mathcal{E}^0 \setminus \mathcal{D}_{train}^0$ (i.e. a random element drawn uniformly from $\text{Bij}(\{1, \dots, \ell + m_0\}, \mathcal{E}^0 \setminus \mathcal{D}_{train}^0)$) and σ_1 an arbitrary ordering of \mathcal{H}_1 (an element of $\text{Bij}(\{1, \dots, |\mathcal{H}_1|\}, \mathcal{H}_1)$), e.g. lexical ordering. Let:*

$$\begin{aligned} \mathcal{D}_{cal}^{\sigma_0} &= \{\sigma_0(i), i = 1, \dots, \ell\}; \\ \mathcal{H}_0^{\sigma_0} &= \{\sigma_0(i), i = \ell + 1, \dots, \ell + m_0\}; \\ \mathcal{D}_{test}^{\sigma_0} &= \mathcal{H}_0^{\sigma_0} \cup \{\sigma_1(i), i = 1, \dots, |\mathcal{H}_1|\}; \\ S'_i &= S_{\sigma_0(i)}, i = 1, \dots, \ell + m_0; \\ S'_{\ell + m_0 + j} &= S_{\sigma_1(j)}, j = 1, \dots, |\mathcal{H}_1|. \end{aligned}$$

Then the distribution of $(\mathcal{D}_{\text{cal}}^{\sigma_0}, \mathcal{D}_{\text{test}}^{\sigma_0}, \mathcal{H}_0^{\sigma_0})$ conditionally to W is identical to that of $(\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}}, \mathcal{H}_0)$ given by (3)-(4), and the scores $(S'_i)_{1 \leq i \leq \ell+m_0}$ form an exchangeable family of random variables (conditionally to W , hence also unconditionally).

Proof. Straightforward since a (set-valued) random variable $F \sim \Xi(C, N)$ can be represented by a set whose elements are N successive draws from C without replacement, in turn equivalent to (i) drawing uniformly an ordering of C ; (ii) taking the N first elements according to that order. \square

Proof of Theorem 1. Let us consider the same construction and notation as in Lemma 3 and define additionally

$$\tilde{p}_{i,j} = \frac{1}{\ell+1} \left(1 + \sum_{(u,v) \in \mathcal{D}_{\text{cal}}^{\sigma_0}} \mathbb{1}_{\{S_{i,j} \leq S_{u,v}\}} \right), \quad (i,j) \in \mathcal{D}_{\text{test}}^{\sigma_0};$$

observe that it is merely definition (1) with $\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}}$ replaced by $\mathcal{D}_{\text{cal}}^{\sigma_0}, \mathcal{D}_{\text{test}}^{\sigma_0}$.

We recall the definition of the FDP of a rejection set R :

$$\text{FDP}(R, \mathcal{H}_0) = \frac{|R \cap \mathcal{H}_0|}{|R| \vee 1},$$

where for precision of notation, we have specified explicitly the set of null hypotheses \mathcal{H}_0 on the left-hand side, because it is random in our observation model.

For a rejection procedure $R((p_{i,j}))$ only depending on the family of p -values, observe that we have

$$\text{FDP}(R((p_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}}), \mathcal{H}_0) \stackrel{\mathcal{L}}{=} \text{FDP}(R((\tilde{p}_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}^{\sigma_0}}), \mathcal{H}_0^{\sigma_0}), \quad (6)$$

where $\stackrel{\mathcal{L}}{=}$ represents equality in distribution. Namely, by inspection of the definition of the family p -values (1), notice that conditionally to W those only depend on $(\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}})$ (since the scores are measurable with respect to W , see Lemma 2). As a consequence, the left-hand side of (6) only depends on $(\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{test}}, \mathcal{H}_0)$, and the right-hand-side is obtained by replacing those by $(\mathcal{D}_{\text{cal}}^{\sigma_0}, \mathcal{D}_{\text{test}}^{\sigma_0}, \mathcal{H}_0^{\sigma_0})$, which according to Lemma 3 have the same joint distribution conditional to W . Hence (6) holds conditionally to W (hence also unconditionally). We therefore focus on the right-hand side of (6) from now on.

For any set of p -values $(p_i)_{i \in \mathcal{I}}$, denote by $R_{\text{BH}}((p_i)_{i \in \mathcal{I}}) \subseteq \mathcal{I}$ the rejection set of the BH procedure; it takes the form

$$R_{\text{BH}}((p_i)_{i \in \mathcal{I}}) = \left\{ i \in \mathcal{I} : p_i \leq t_{\text{BH}}(\{p_i, i \in \mathcal{I}\}) \right\},$$

where t_{BH} is a threshold that depends only on the (multi)set of p -values $\{p_i, i \in \mathcal{I}\}$. Let us introduce the set of r.v. p'_1, \dots, p'_m such that

$$p'_a = \frac{1}{\ell+1} \left(1 + \sum_{b=1}^{\ell} \mathbb{1}_{S'_{\ell+a} \leq S'_b} \right), \quad 1 \leq a \leq m, \quad (7)$$

with the r.v. $(S'_a)_{1 \leq a \leq m}$ defined in Lemma 3.

It can be readily checked that

$$p'_a = \tilde{p}_{\sigma_0(\ell+a)}, \quad 1 \leq a \leq m_0; \quad p'_{m_0+a} = \tilde{p}_{\sigma_1(a)}, \quad 1 \leq a \leq |\mathcal{H}_1|. \quad (8)$$

As a consequence, we have that

$$\begin{aligned}
\text{FDP} \left(R_{BH}((\tilde{p}_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}^{\sigma_0}}), \mathcal{H}_0^{\sigma_0} \right) &= \frac{|\mathcal{H}_0^{\sigma_0} \cap R_{BH}((\tilde{p}_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}^{\sigma_0}})|}{|R_{BH}((\tilde{p}_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}^{\sigma_0}})| \vee 1} \\
&= \frac{\sum_{(i,j) \in \mathcal{H}_0^{\sigma_0}} \mathbb{1}\{\tilde{p}_{i,j} \leq t_{BH}(\{p_{i,j}, (i,j) \in \mathcal{D}_{\text{test}}^{\sigma_0}\})\}}{|R_{BH}((\tilde{p}_{i,j})_{(i,j) \in \mathcal{D}_{\text{test}}^{\sigma_0}})| \vee 1} \\
&= \frac{\sum_{a=1}^{m_0} \mathbb{1}\{p'_a \leq t_{BH}(\{p'_b, 1 \leq b \leq m\})\}}{|R_{BH}((p'_b)_{1 \leq b \leq m})| \vee 1} \tag{9} \\
&= \text{FDP} \left(R_{BH}((p'_a)_{1 \leq a \leq m}), \{1, \dots, m_0\} \right), \tag{10}
\end{aligned}$$

where (9) is due to (8) and the definitions in Lemma 3.

Under Assumption 1, Theorem 3.3. of Marandon et al. (2022) entails that conditionally on W , the r.v. $(p'_a)_{1 \leq a \leq m_0}$ are each marginally super-uniform and the r.v. $(p'_a)_{1 \leq a \leq m}$ are PRDS on $\{1, \dots, m_0\}$. It follows from Benjamini and Yekutieli (2001) that the conditional expectation with respect to W of the right term in (10) is below α (thus this holds also for the unconditional expectation). This concludes the proof. \square

Furthermore, the same arguments as above (representing the distributions of null scores as a conditionally exchangeable distribution) also allow us to leverage recent results of Gazin et al. (2023) to get a control of the false discovery proportion (FDP) with high probability, and which hold uniformly over the choice of the rejection threshold.

Theorem 4. *Consider the same setting and assumptions as in Theorem 1, but instead of applying the BH procedure as the last step of Algorithm 1, denote $R(t)$ as the rejection set obtained as the set of edges having conformal p -value less than t , for $t \in [0, 1]$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ with respect to the sampling matrix Ω and the draw of \mathcal{D}_{cal} , it holds*

$$P \left(\forall t \in [0, 1] : \text{FDP}(R(t)) \leq \frac{m}{1 \vee |R(t)|} (1 + \lambda(m, \ell, \delta)) \right) \geq 1 - \delta, \tag{11}$$

where $\lambda(m, \ell, \delta)$ is the $1 - \delta$ quantile of a certain universal probability distribution $P_{m, \ell}$, and satisfies

$$\lambda(m, \ell, \delta) \leq \left(\frac{\log \delta^{-1} + \log(1 + 2\sqrt{\pi}(m \wedge \ell))}{m \wedge \ell} \right)^{\frac{1}{2}}. \tag{12}$$

Proof. Following the same argument as in the proof of Theorem 1, we use the fact that

$$((p_{i,j})_{j \in \mathcal{D}_{\text{test}}}, \mathcal{H}_0) \stackrel{\mathcal{L}}{=} ((\tilde{p}_{i,j})_{j \in \mathcal{D}_{\text{test}}^{\sigma_0}}, \mathcal{H}_0^{\sigma_0}).$$

Hence, since the event considered in the claim (11) only depends on $((p_{i,j})_{j \in \mathcal{D}_{\text{test}}}, \mathcal{H}_0)$, its probability is unchanged when those are replaced by $((\tilde{p}_{i,j})_{j \in \mathcal{D}_{\text{test}}^{\sigma_0}}, \mathcal{H}_0^{\sigma_0})$.

Consider now the p -values $(p'_i)_{1 \leq i \leq m}$ given by explicit reindexing of $(\tilde{p}_{i,j})_{j \in \mathcal{D}_{\text{test}}^{\sigma_0}}$ via (7)-(8); again, since the event in (11) only depends on the p -values and on \mathcal{H}_0 as (multi)sets and not of the ordering of their indexation, we can use the reindexing given by $(p'_i)_{1 \leq i \leq m}$ (and the corresponding $\mathcal{H}'_0 = \{1, \dots, m_0\}$) without changing the probability. Since the reindexed p -values $(p'_i)_{1 \leq i \leq m}$ can now be seen as conformal p -values based on exchangeable scores without ties $(S'_i)_{1 \leq i \leq \ell+m}$, the claimed bound then holds, conditionally to W , as a direct application of Corollary 4.1 and Theorem 2.4 of Gazin et al. (2023) (with some straightforward estimates to simplify slightly the final expression). Thus the claimed probability bound also holds unconditionally. \square

Remark 1. Observe that the uniformity with respect to t allows for a post hoc (i.e. data-dependent) choice of the rejection threshold, rather than having to set a target level α for the FDR in advance, see Blanchard et al. (2020). This is of particular interest since the output of a link prediction procedure is often presented to the user as a list of edges ordered by their scores rather than a fixed rejection set; the user can then monitor by themselves the guaranteed FDP bound in dependence of the threshold and decide of an adequate trade-off based on possibly additional criteria (e.g. concerning the geometry of rejected edges, or using additional information).

From the results of Gazin et al. (2023), the “universal probability” $P_{m,\ell}$ appearing in Theorem 4 is that of the supremum deviation from identity of the empirical cdf for the color histogram of a Pólya urn scheme with ℓ draws from an urn initially containing $(m + 1)$ balls of different colors. It can be easily approximated by Monte-Carlo simulation to get a sharper estimate. The explicit estimate (12) has the advantage of clearly showing an exponential concentration of order $\mathcal{O}(\sqrt{\log((m \wedge \ell)/\delta)/(m \wedge \ell)})$. The bound of Gazin et al. (2023) also allows the replacement of m by a specific built-in estimate \hat{m}_0 of m_0 , and a slightly sharper estimate where the logarithmic dependence in $(n \wedge m)$ can be improved to an iterated logarithm. Both of these refinements are ignored here for simplicity.

4 Discussion

Comparison to Huang et al. (2023) : Our arguments on score exchangeability bears similarity to the reasoning of Huang et al. (2023), who considered conformal prediction intervals for label values on the nodes of a graph. A first (superficial) difference is that we predict on edges rather than nodes of the graph, and that we consider FDR and FDP for edge presence/absence rather than false coverage rate prediction intervals. We do not need an analogue of Assumption 1 of Huang et al. (2023), which assumes explicitly that the score function is invariant with respect to permutation of indices of the calibration+test set. Namely, this is in fact automatically satisfied in our setting, since the training function g applied to the masked observation data Z_{train} in Algorithm 1 does not have the information of which edges are in \mathcal{D}_{cal} and which in $\mathcal{D}_{\text{test}}$.

More importantly, the nature of our results are different because, instead of fixing a threshold t and get guarantees for the FDR or FDP of $R(t)$ as in Huang et al. (2023), we consider either a (specific) data-dependant threshold t_{BH} which ensures a control of the FDR, or a bound uniform with respect to t on the FDP which allows to use any data-dependent threshold and still ensures valid bounds on the FDP. For a fixed beforehand threshold t , it is possible to follow the same line of argumentation we used to explicitly represent the scores as an exchangeable uple, and combine with the results of Marques F. (2023) instead of Marandon (2023) or Gazin et al. (2023). In this case the same control as appearing in Huang et al. (2023) is recovered. See also Gazin et al. (2023) for the relation of the distribution $P_{m,\ell}$ to the results of Marques F. (2023); Huang et al. (2023).

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

- Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. Ann. Statist., 48(3):1281–1303.
- Bleakley, K., Biau, G., and Vert, J.-P. (2007). Supervised reconstruction of biological networks with local models. Bioinformatics, 23(13):i57–i65.
- Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc., 96(456):1151–1160.
- Gazin, U., Blanchard, G., and Roquain, E. (2023). Transductive conformal inference with adaptive scores. arXiv preprint 2310.18108.
- Huang, K., Jin, Y., Candes, E., and Leskovec, J. (2023). Uncertainty quantification over graph with conformalized graph neural networks. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, Advances in Neural Information Processing Systems, volume 36, pages 26699–26721. Curran Associates, Inc.
- Lu, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. Physica A: statistical mechanics and its applications, 390(6):1150–1170.
- Marandon, A. (2023). Conformal link prediction to control the error rate. arXiv preprint arXiv:2306.14693.
- Marandon, A., Rebafka, T., Roquain, E., and Sokolovska, N. (2022). False clustering rate in mixture models. arXiv preprint arXiv:2203.02597.
- Marques F., P. C. (2023). On the universal distribution of the coverage in split conformal prediction. arXiv preprint 2303.02770.
- Sportisse, A., Boyer, C., and Josse, J. (2020). Imputation and low-rank estimation with missing not at random data. Statistics and Computing, 30(6):1629–1643.
- Tabouy, T., Barbillon, P., and Chiquet, J. (2020). Variational inference for stochastic block models from sampled data. Journal of the American Statistical Association, 115(529):455–466.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). Algorithmic learning in a random world. Springer, New York.
- Zhang, M. and Chen, Y. (2018). Link prediction based on graph neural networks. In Advances in Neural Information Processing Systems, pages 5165–5175.

Acknowledgments

A. Marandon acknowledges funding from the Turing-Roche Strategic Partnership. GB, GD and RP acknowledge funding from the grants ANR-21-CE23-0035 (ASCAI), ANR-19-CHIA-0021-01 (BISCOTTE) and ANR-23-CE40-0018-01 (BACKUP) of the French National Research Agency ANR.