

# Modeling Node Exposure for Community Detection in Networks

Sameh Othman<sup>1</sup>, Johannes Schulz<sup>1</sup>, Marco Baity-Jesi<sup>2</sup>, and Caterina De Bacco<sup>1</sup>

<sup>1</sup> Max Planck Institute for Intelligent Systems, Cyber Valley, Tuebingen, 72076, Germany

<sup>2</sup> Eawag (ETH), Überlandstrasse 133, CH-8600 Dübendorf, Switzerland

**Abstract.** In community detection, datasets often suffer a sampling bias for which nodes which would normally have a high affinity appear to have zero affinity. This happens for example when two affine users of a social network were not exposed to one another. Community detection on this kind of data suffers then from considering affine nodes as not affine. To solve this problem, we explicitly model the (non-)exposure mechanism in a Bayesian community detection framework, by introducing a set of additional hidden variables. Compared to approaches which do not model exposure, our method is able to better reconstruct the input graph, while maintaining a similar performance in recovering communities. Importantly, it allows to estimate the probability that two nodes have been exposed, a possibility not available with standard models.

**Keywords:** Networks, community detection, latent variable models

## 1 Introduction

Modeling the mechanisms of how nodes interact in networks is a relevant problem in many applications. In social networks, we observe a set of interactions between people, and one can use this information to cluster them into communities based on some notion of similarity [1]. Broadly speaking, the connections between users can be used to infer users' membership, and this in turns determines the likelihood that a pair of users interacts. Real networks are often sparse, people interact with a tiny amount of individuals, compared to the large set of possible interactions that they could in principle explore. Traditionally, models for community detection in networks treat an existing link as a positive endorsement between individuals: if two people are friends in a social network, this means they like each other. In assortative communities, where similar nodes are more likely to be in the same group [2,3], this encourages the algorithm to put these two nodes into the same community. On the contrary, a non-existing link influences the model to place them into different communities, as if the two non-interacting individuals were not compatible. However, many of these non-existing links (–especially in large-scale networks–) are absent because the

individuals are not aware of each other, rather than because they are not interested in interacting. This is a general problem in many network data sets: we know that interacting nodes have a high affinity, but we can not conclude the contrary about non-interacting nodes.

This problem has been explored in the context of recommender systems [4,5,6,7], where it is crucial to learn what items that a user did not consume could be of interest. In this context, items' exposure is often modeled by means of propensity scores or selection biases assigned to user-item pairs that increase the probability of rare consumption events.

It is not clear how to adapt these techniques to the case of networks of interacting individuals, hence the investigation of this problem in the context of networks is still missing. Existing approaches partially account for this by giving more weight to existing links, as in probabilistic generative models that use a Poisson distribution for modeling the network adjacency matrix [8,9,10,11]. These methods are effective, but may be missing important information contained in non-existing links.

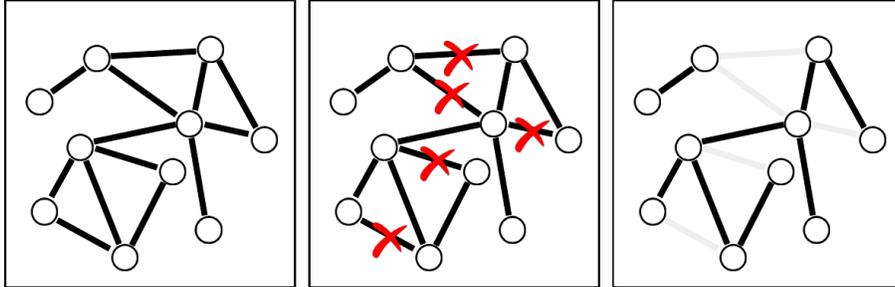
## 2 Community detection with exposure

We address this problem by considering a probabilistic formulation that assigns probabilities to pairs of nodes of being exposed or not. These are then integrated into standard probabilistic approaches for generative networks with communities. For this, as a reference model we consider MULTITENSOR [8], as it is a flexible model that takes in input a variety of network structures (*e.g.* directed or undirected networks, weighted or unweighted) and detects overlapping communities in a principled and scalable way.

### 2.1 Representing exposure

Consider an  $N \times N$  *observed* network adjacency matrix  $\mathbf{A}^{(o)}$ , where  $A_{ij}^{(o)} \geq 0$  is the weight of the interaction between nodes  $i$  and  $j$ , this is the input data. For instance,  $A_{ij}^{(o)}$  could be the number of times that  $i$  and  $j$  met or exchanged messages. If a link  $A_{ij}^{(o)}$  exists, this indicates an affinity between individuals  $i$  and  $j$ , triggered by both individuals' inner preferences. If the link does not exist ( $A_{ij}^{(o)} = 0$ ), one usually assumes that this indicates a lack of affinity between  $i$  and  $j$ . However, the link might not exist simply because  $i$  and  $j$  never met. This is the case in social networks, where an *ego* might follow an *alter* because of personal preference, but this choice is subject to being exposed to the *alter* in the first place. This suggests that the event of being exposed to someone influences the patterns of interactions observed in networks. We are interested in incorporating this notion of exposure in modeling network data, and investigate how results change.

To represent this, we postulate the existence of a *ground-truth adjacency matrix*,  $\mathbf{A}^{(g)}$ , that indicates the affinity between nodes  $i$  and  $j$  regardless of



**Fig. 1.** Diagram of the exposure mechanism. On the **left** we have the full graph ( $\mathbf{A}^{(g)}$ ). We then set to zero the probability of some connections through a mask  $\mathbf{Z}$  (**center**), and reconstruct the true graph and communities based solely on the visible links,  $\mathbf{A}^{(o)}$  (**right**).

whether the two nodes were exposed to each other (Fig. 1–left). In addition, we introduce a *dilution matrix*  $\mathbf{Z}$  (red crosses in Fig. 1–center), with values  $Z_{ij} = 0, 1$  indicating whether nodes  $i$  and  $j$  were exposed ( $Z_{ij} = 1$ ) or not ( $Z_{ij} = 0$ ). The observed matrix is then the element-wise product of the ground truth network times the dilution matrix,

$$\mathbf{A}^{(o)} = \mathbf{A}^{(g)} \otimes \mathbf{Z}, \quad (1)$$

where  $\otimes$  indicates an element-by-element multiplication. A diagram of the resulting matrix is shown in Fig. 1–right. Through this representation, a zero-entry  $A_{ij}^{(o)} = 0$  can be attributed to  $A_{ij}^{(g)} = 0$  (lack of affinity),  $Z_{ij} = 0$  (lack of exposure) or both.

Standard models for community detection do not account for exposure, therefore they treat a zero-entry  $A_{ij}^{(o)} = 0$  as a signal for non-affinity. We aim at measuring both communities and exposure, given the observed data  $A_{ij}^{(o)}$ . In other words, for a given node  $i$ , we would like to estimate its community membership and for a given pair  $(i, j)$  we want to estimate the probability that they were exposed to each other. For simplicity, we show derivations for the case of undirected networks, but similar ones apply to directed ones.

## 2.2 The ground truth adjacency matrix

In our notation, we use  $\theta$  to denote the latent variables affecting community detection, *i.e.* determining the probability of observing an interaction between  $i$  and  $j$  given that they have been exposed. We will treat the case of symmetric edges,  $A_{ij}^{(g)} = A_{ji}^{(g)}$ , and provide an extension to asymmetric interactions in App. A. Following the formalism of Ref. [8], we assign a  $K$ -dimensional hidden variable  $u_i$  to every node  $i$ . Since different communities may interact in different ways, we also introduce a  $K \times K$  affinity matrix  $w$ , regulating the density of

interactions between different groups. The latent variables related to the ground truth matrix are then  $\theta = (u, w)$ .

We express the expected interaction between two nodes through a parameter

$$\lambda_{ij} = \sum_{k,q}^K u_{ik} u_{jq} w_{kq}, \quad (2)$$

and extract the elements of  $\mathbf{A}^{(g)}$  from a Poisson distribution with mean  $\lambda_{ij}$ ,

$$P(A_{ij}^{(g)} | u_i, u_j, w) = \text{Pois} \left( A_{ij}^{(g)}; \lambda_{ij} \right) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{A_{ij}^{(g)}}}{A_{ij}^{(g)}!}. \quad (3)$$

We then assume conditional independence between different pairs of edges given the latent variables  $P(\mathbf{A}^{(g)} | u, u, w) = \prod_{i < j} P(A_{ij}^{(g)} | u_i, u_j, w)$ , but this can be generalized to more complex dependencies [12,13,14]. We do not explore this here.

### 2.3 The observed adjacency matrix

The observed adjacency matrix depends on whether two nodes were exposed or not, through the matrix  $\mathbf{Z}$ . If  $Z_{ij} = 1$ , the two nodes are exposed, and the edge comes from the ground truth matrix, *i.e.*  $P(A_{ij}^{(o)} | Z_{ij} = 1, \theta) = P(A_{ij}^{(g)} | \theta) = \text{Pois}(A_{ij}^{(g)}; \lambda_{ij})$ . If  $Z_{ij} = 0$ , then  $A_{ij}^{(o)} = 0$  regardless of  $\lambda_{ij}$ . Therefore, the elements of  $\mathbf{A}^{(o)}$  are extracted from the distribution

$$P(A_{ij}^{(o)} | Z_{ij}, \theta) = \text{Pois}(A_{ij}^{(o)}; \lambda_{ij})^{Z_{ij}} \delta(A_{ij}^{(o)})^{1-Z_{ij}}. \quad (4)$$

Since  $Z_{ij}$  is binary, we assign it a Bernoulli prior with parameter  $\mu_{ij}$ ,

$$P(\mathbf{Z} | \mu) = \prod_{i < j} P(Z_{ij} | \mu_{ij}) = \prod_{i < j} (\mu_{ij})^{Z_{ij}} (1 - \mu_{ij})^{1-Z_{ij}}. \quad (5)$$

The parameter  $\mu_{ij}$  will depend on some latent variable related to nodes  $i$  and  $j$ . There are several possible choices for that. Here, we consider a simple setting:

$$\mu_{ij} = \mu_i \mu_j, \quad (6)$$

$$\mu_i \in [0, 1], \quad (7)$$

This allows to keep the number of parameters small and has an easy interpretation. In fact, the parameter  $\mu_i$  acts as the propensity of an individual to be exposed to others: the higher its value, the higher the probability that node  $i$  will be exposed to other nodes. This way of modeling exposure only adds one more parameter per node, allowing for heterogeneous behaviors among users while keeping the model compressed. The full set of variables that need to be inferred consists of the  $u$ , the  $w$  and the  $\mu$  variables, which amounts to  $NK + K^2 + N$  parameters, which is one order of magnitude smaller than the  $N^2$  elements of  $\mathbf{A}^{(o)}$ .

## 2.4 Inference and Expectation-Maximization

Given the data  $\mathbf{A}^{(o)}$ , our goal is to first determine the values of the parameters  $\theta$ , which fixes the relationship between the hidden indicator  $Z_{ij}$  and the data, and then to approximate  $Z_{ij}$  given the estimated  $\theta$ .

We perform this using statistical inference as follows. Consider the posterior distribution  $P(\mathbf{Z}, \theta | \mathbf{A}^{(o)})$ . Since the dilution  $\mathbf{Z}$  is independent from the parameters  $\theta$  and all the edges are considered conditionally independent given the parameters, Bayes' formula gives

$$P(\mathbf{Z}, \theta | \mathbf{A}^{(o)}) = \frac{P(\mathbf{A}^{(o)} | \mathbf{Z}, \theta) P(\mathbf{Z} | \mu) P(\theta)}{P(\mathbf{A}^{(o)})}. \quad (8)$$

Summing over all the possible indicators we have:

$$P(\theta | \mathbf{A}^{(o)}) = \sum_{\mathbf{Z}} P(\mathbf{Z}, \theta | \mathbf{A}^{(o)}) = \prod_{i < j} \sum_{Z_{ij}=0,1} P(Z_{ij}, \theta | \mathbf{A}^{(o)}) \quad , \quad (9)$$

which is the quantity that we need to maximize to extract the optimal  $\theta$ . It is more convenient to maximize its logarithm, as the two maxima coincide. We use Jensen's inequality:

$$\log P(\theta | \mathbf{A}^{(o)}) = \log \sum_{\mathbf{Z}} P(\mathbf{Z}, \theta | \mathbf{A}^{(o)}) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{P(\mathbf{Z}, \theta | \mathbf{A}^{(o)})}{q(\mathbf{Z})} := \mathcal{L}(q, \theta, \mu) \quad (10)$$

where  $q(\mathbf{Z})$  is any distribution satisfying  $\sum_{\mathbf{Z}} q(\mathbf{Z}) = 1$ , we refer to this as the variational distribution.

Inequality (10) is saturated when

$$q(\mathbf{Z}) = \frac{P(\mathbf{Z}, \theta | \mathbf{A}^{(o)})}{\sum_{\mathbf{Z}} P(\mathbf{Z}, \theta | \mathbf{A}^{(o)})}, \quad (11)$$

hence this choice of  $q$  maximizes  $\mathcal{L}(q, \theta, \mu)$  with respect to  $q$ . Further maximizing it with respect to  $\theta$  gives us the optimal latent variables. This can be done in an iterative way using Expectation-Maximization (EM), alternating between maximizing with respect to  $q$  using Eq. (11) and then maximizing  $\mathcal{L}(q, \theta, \mu)$  with respect to  $\theta$  and  $\mu$ .

To obtain the updates for the parameters we need to derive the equations that maximize  $\mathcal{L}(q, \theta, \mu)$  with respect to  $\theta$  and  $\mu$  and set these derivatives to

zero. This leads to the following closed-form updates:

$$u_{ik} = \frac{\sum_j Q_{ij} A_{ij} \sum_q \rho_{ijkq}}{\sum_j Q_{ij} \sum_q u_{jq} w_{kq}} \quad (12)$$

$$w_{kq} = \frac{\sum_{i,j} Q_{ij} A_{ij} \rho_{ijkq}}{\sum_{i,j} Q_{ij} u_{ik} u_{jq}} \quad (13)$$

$$\rho_{ijkq} = \frac{u_{ik} u_{jq} w_{kq}}{\sum_{k,q} u_{ik} u_{jq} w_{kq}} \quad (14)$$

$$\mu_i = \frac{\sum_j Q_{ij}}{\sum_j \frac{(1-Q_{ij}) \mu_j}{(1-\mu_i \mu_j)}} \quad , \quad (15)$$

where we defined  $Q_{ij} = \sum_{\mathbf{Z}} Z_{ij} q(\mathbf{Z})$  the expected value of  $Z_{ij}$  over the variational distribution.

As  $\mu_i$  appears on both sides of Eq. (15), this can be solved with root-finding methods bounding  $\mu_i$  to the interval  $[0, 1]$ , to be compatible as a parameter of the Bernoulli prior.<sup>3</sup>

Finally, to evaluate  $q(Z)$ , we substitute the estimated parameters inside Eq. (8), and then into Eq. (11) to obtain:

$$q(\mathbf{Z}) = \prod_{i < j} Q_{ij}^{Z_{ij}} (1 - Q_{ij})^{(1-Z_{ij})} \quad , \quad (16)$$

where

$$Q_{ij} = \frac{\text{Pois}(A_{ij}; \lambda_{ij}) \mu_{ij}}{\text{Pois}(A_{ij}; \lambda_{ij}) \mu_{ij} + \delta(A_{ij}) (1 - \mu_{ij})} \quad . \quad (17)$$

In other words, the optimal  $q(\mathbf{Z})$  is a product  $\prod_{i < j} q_{ij}(\mathbf{Z}_{ij})$  of Bernoulli distributions  $q_{ij}$  with parameters  $Q_{ij}$ . This parameter is also a point-estimate of the exposure variable, as for the Bernoulli distribution  $Q_{ij} = \mathbb{E}_q[\mathbf{Z}_{ij}]$ .

The algorithmic EM procedure then works by initializing at random all the parameters and then iterating Eqs. (12) to (15) for fixed  $q$ , and the calculating Eq. (17) given the other parameters, and so on until convergence of  $\mathcal{L}$ . The function  $\mathcal{L}$  is not convex, hence we are not guaranteed to converge to the global optimum. In practice, one needs to run the algorithm several times with different random initial parameters' configurations and then select the run that leads to best values of  $\mathcal{L}$ . In the following experiments we use 5 of such realizations.

### 3 Results

We test our algorithm on synthetic and real data, and compare it to its formulation without exposure, *i.e.* the MULTITENSOR algorithm described in Ref. [8]. In the following, we refer to our algorithm as EXP, and we use NoEXP for the algorithm that does not utilize exposure.

<sup>3</sup> In practice, we limit the domain of  $\mu_i$  to the interval  $[\epsilon, 1 - \epsilon]$ , where  $\epsilon$  is a small hyperparameter chosen to avoid numerical overflows of  $\mathcal{L}$ . To maintain the model interpretable in terms of exposure, at the end of the optimization we set to zero each  $\mu_i \equiv \epsilon$  and to one each  $\mu_i \equiv 1 - \epsilon$ .

### 3.1 Synthetic data

Synthetic data experiments are particularly interesting, because we can validate our model performances on the ground truth values. The creation of a synthetic dataset follows the generative model described in Sect. 2.1:

1. For a graph with  $N = 500$  nodes, we generate the latent parameters  $\theta$  and  $\mu$  as follows. We draw overlapping communities by sampling  $u_i$  from a Dirichlet distribution with parameter  $\alpha_k = 1, \forall k$ ; we choose an assortative  $w$  by selecting the off-diagonal entries to be 0.001 times smaller than the on-diagonal ones. We then vary  $K \in [3, 5, 8]$ . We draw  $\mu_i$  from a Beta distribution  $\text{Beta}(\mu_i; 2, \beta)$ , where we vary  $\beta \in [0.1, 10]$  to tune the fraction of unexposed links.
2. Sample  $\mathbf{A}^{(g)}_{ij}$  from a Poisson distribution with means  $\lambda_{ij} = \sum_{k,q} u_{ik} u_{jq} w_{kq}$ .
3. Sample  $\mathbf{Z}$  from a Bernoulli distribution of means  $\mu_{ij} = \mu_i \mu_j$ .
4. Calculate the matrix  $\mathbf{A}^{(o)} = \mathbf{A}^{(g)} \otimes \mathbf{Z}$ . This matrix has on average  $\langle k \rangle$  links per node.

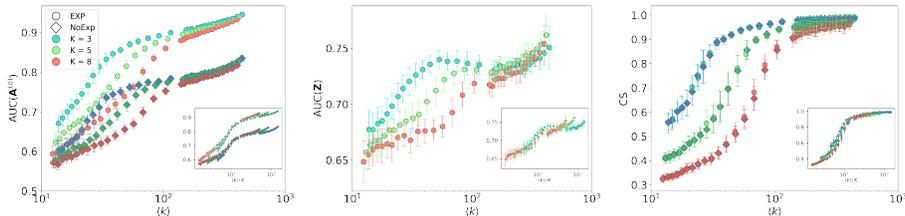
We repeat this procedure 10 times for each set of parameters to obtain different random realizations of synthetic data. We then apply the EXP and NoEXP algorithms to  $\mathbf{A}^{(o)}$  to learn the parameters and study the performance as a function of  $\langle k \rangle$ , controlling the density of observed edges.

*Reconstructing hidden links* We start by testing the ability of the model to predict missing links, a procedure often used as a powerful evaluation framework for comparing different models [15,16]. We use a 5-fold cross-validation scheme where we hide 20% of the edges in  $\mathbf{A}^{(o)}$  and train the model on the remaining 80%. Performance is then computed on the hidden 20% of the edges. As a performance evaluation metric we measure the area under the receiver operating characteristic curve (AUC) between the inferred values and the ground truth used to generate  $\mathbf{A}^{(o)}$  on the test set. The AUC is the probability that a randomly selected existing edge is predicted with a higher score than a randomly selected non-existing edge. A value of 1 means optimal performance, while 0.5 is equivalent to random guessing. As the score of an edge  $\mathbf{A}^{(o)}_{i,j}$  we use the quantity  $Q_{ij} \lambda_{ij}$  for EXP, and  $\lambda_{ij}$  for NoEXP. In both cases, these are the expected values of  $A^{(o)}_{ij}$  using the estimates of the latent parameters and, for EXP, over the inferred  $q(\mathbf{Z})$ . We find that the EXP algorithm outperforms NoEXP by a large margin, which increases as the network becomes more dense, going above 10%, as shown in Fig. 2–left. At low densities, the performance increase of the EXP algorithm is narrow for models with a large number of communities, while at large densities it becomes bigger and independent of the number of communities. This result suggests that EXP is capturing the input data better—consistently for varying dilution densities—than a model that does not account for exposure.

*Guessing unexposed links* Our algorithm not only allows us to predict missing edges but also gives interpretable estimates of the probability of exposure between nodes. These probabilities follow naturally from the posterior distribution

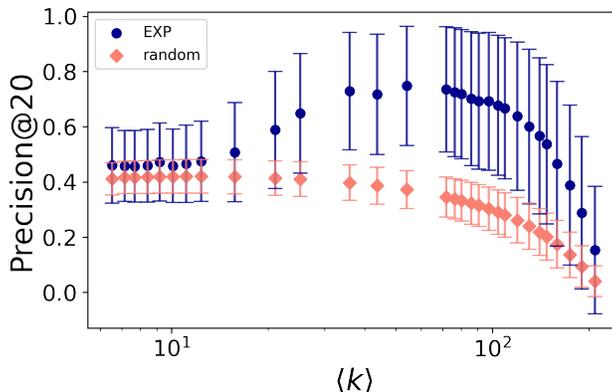
on  $\mathbf{Z}$ , which is the Bernoulli distribution in Eq. (16). Standard algorithms as NoEXP cannot estimate this. We can use the mean value  $Q_{ij}$  as in Eq. (17) as a score of an edge to compute the AUC between inferred and ground truth values of  $\mathbf{Z}$ , analogously to what was done for reconstructing  $\mathbf{A}^{(o)}$ . We report in Fig. 2–center the ability of EXP to reconstruct the matrix  $\mathbf{Z}$ , *i.e.* to infer which edges were removed in the dilution step. The AUC varies between 0.65 and 0.75, well above the random baseline of 0.5. We notice how the values increase as the density of connection increases, but stay above 0.65 even at small density values, where reconstruction is more challenging.

*Inferring communities* In Fig. 2–right, we can see that EXP and NoEXP show similar performances in reconstructing communities. From this plot we can also notice how reconstruction improves for larger densities and fewer communities. The similar performances may be due to selecting a simple prior as in Eq. (6). For a more structured prior, the inferred communities would likely change and potentially improve. Given this similar community detection abilities but the better predictive power in reconstructing  $\mathbf{A}^{(o)}$ , we argue that the learned  $Q_{ij}$ 's are important to boost prediction compared to a model that does not properly account for exposure. This is true even for a simple prior.



**Fig. 2.** Performance of the EXP and NoEXP algorithms on synthetic data. The matrix  $\mathbf{A}^{(g)}$  has  $K = 3, 5, 8$  communities and  $N = 500$ . The exposure mask  $Z$  is extracted from a binomial distribution with parameter  $\mu_{ij} = \mu_i \mu_j$ . **Left:** AUC between the inferred values and the ground truth used to generate  $\mathbf{A}^{(o)}$ . **Center:** AUC of the reconstruction of the exposure mask  $\mathbf{Z}$ . **Right:** Cosine similarity between inferred and ground truth communities. **Inset:** We show the same data as in the main plots by rescaling the average number of links by the number of communities.

*Dependence on the number of communities* All of these metrics exhibit a scaling w.r.t. the variable  $\langle k \rangle / K$ , as can be seen in the insets of Fig. 2. This suggests that the curves seem to be independent of the number of communities when accounting for this rescaling. Thus observing the behavior for one particular value of  $K$  should be informative enough to understand how the model behaves for various densities.



**Fig. 3.** Suggesting unexposed compatible nodes. For each node  $i$ , we suggest the 20 links with highest  $\lambda_{ij}$  inferred by our algorithm from the non-observed links where  $A_{ij}^{(o)} = 0$ . We show the P@20 averaged across all nodes and compare with a uniform-at-random baseline (random) where 20 nodes are selected at random among the available ones. Error bars are standard deviations. Here we use a synthetic network generated as in Sect. 3.1 with  $N = 500$  and  $K = 5$ .

*Suggesting good matches* Since the EXP algorithm is good at predicting which nodes were removed from the original graph (Fig. 2–center), we can use this to address the following question: Is the EXP algorithm able to suggest two nodes that have high affinity despite not having any connection? In other words, we are asking whether we are able to find links that in  $\mathbf{A}^{(o)}$  are absent, but have a high expected value in  $\mathbf{A}^{(g)}$ . To test this ability, we take for each node  $i$ : a) all the possible neighbors  $j$  such that  $A_{ij}^{(o)} = 0$ ; b) select among them the 20 with the largest inferred affinity  $\lambda_{ij}$ ; and c) check how many of those are present in  $\mathbf{A}^{(g)}$ . We call Precision@20 (P@20) the fraction of links which were correctly inferred, averaging across all nodes. In Fig. 3 we show that for intermediate dilution values, the P@20 reaches around 80%, and outperforms random guessing at any value of the dilution. Notice that random guessing is not constant in  $\langle k \rangle$ . This is because this depends on the number of missing links in  $\mathbf{A}^{(o)}$ , and those depend both on the density of  $\mathbf{A}^{(g)}$  and on the dilution mask  $\mathbf{Z}$ . Specifically, P@20 of the random baseline goes as  $(\langle k \rangle_g - \langle k \rangle) / (N - \langle k \rangle)$ , where  $\langle k \rangle_g$  is degree of  $\mathbf{A}^{(g)}$ . This is a decreasing function of  $\langle k \rangle$ , for  $\langle k \rangle_g < N$ .

### 3.2 Real data

To test our algorithm on real data, we use the American College Football Network (ACFN) dataset provided in Ref. [17], which represents the schedule of Division I games for the season of the year 2000. Each node in the data set corresponds to a team, and each link is a game played between teams. Teams are grouped in conferences, and each team plays most of its games within a same



**Fig. 4.** **Left:** Performance in predicting missing links of the EXP and NoEXP algorithms on the ACFN dataset. Different marker shapes correspond to different numbers of communities, while blue (red) markers denote instances where EXP (NoEXP) has better performance than NoEXP (EXP). There is a total of 150 markers, denoting 5 folds repeated for 10 random seeds for each value of  $K$ . **Right:** Top 10 games that are recommended by the EXP algorithm with  $K = 11$ , which were not played in the ACFN data set. Different colors indicate different conferences.

conference (though not all teams within a conference encounter each other). Conferences group teams of similar level, but another main criterion is geographic distance. Therefore, this dataset has a community structure which is not based on affinity. Here, affinity indicates that teams are of similar level, and therefore should play in the same conference, if conferences were based solely on affinity.

We randomly hide 20% of the links in the ACFN and check how well the EXP and NoEXP algorithms are able to reconstruct which links are missing. We run the algorithm with various number of communities  $K = 9, 11, 13$ , finding the best result at  $K = 11$ , which is also the number of conferences in the dataset. In Fig. 4–left we show a scatter plot of the AUC trial-by-trial. This reveals a superior performance of the EXP method which outperforms NoEXP in 142 out of 150 trials (5 fold per 10 random seeds for each of  $K = 9, 11, 13$ ). This suggest that EXP is better capturing the data.

In Fig. 4–right we show the top 10 recommendations that we can extract from the EXP algorithm by taking, among the links missing from  $\mathbf{A}^{(o)}$ , those with smallest predicted exposure  $Q_{ij}$  and the highest affinity  $\lambda_{ij}$ . Although, in the absence of ground truth, we are not able to assess the validity of these suggestions, we note that all the suggested links represent unplayed games within the the same conference and that games within teams in different conferences were ranked lower.

## 4 Conclusions

In networks, nodes that would enjoy a high mutual affinity often appear disconnected for reasons that are independent of affinity. This is the case, for example, with people or entities in social networks that have never met, or due to some kind of sampling bias. This introduces a sampling bias in the datasets used for

community detection. We studied this problem through a general framework, where we postulate that affinity in terms of compatibility of communities is not enough in order to explain the existence of a link, but rather a mechanism of exposure between nodes should be taken into account as well.

We proposed a principled probabilistic model, EXP, that takes into account this type of bias and is able to estimate the probability that two non-connected nodes are exposed while jointly learning what communities they belong to. We tested the EXP algorithm against a version of itself that does not account for exposure, NoEXP. On artificial data, where we could validate our results on ground truth parameters and unobserved ground truth data, we found that EXP is as good as NoEXP in learning communities, but it outperforms it when it comes to reconstructing missing links. In addition, the EXP approach allows us to satisfactorily infer which links remained unexposed, an estimate that cannot be done with standard method as, for example, NoEXP. We finally tested our algorithm on a real dataset which has a hidden structure that is independent of the affinity between links, finding that also here the EXP algorithm is better at reconstructing missing links.

The principled approach that we used based on statistical inference is general. It can be made more specific depending on the application at hand. For example, we considered the simple case where exposure only depends on each individual’s propensity towards being exposed. However, this could depend on a more fine structure of society, and we could think of introducing an exposure mechanism that mimics the presence of communities which are independent of affinity (*e.g.* different schools, or different classes in a school). Allowing for community-dependent exposure has the potential to better mimic the kind of dilution that occurs in many real datasets. This can also apply to the AFCN dataset, where a better way to model exposure may be one that allows a structure that is able to account for different conferences or geographical regions. We leave this for future work. Additionally, exposure could be driven by covariate information on nodes, as also used in recommender systems [4]. This could be integrated using variants of community detection methods that account for this extra information [18,19,20]. Exposure could also change through time, and it could also have some dependence on the structure of  $\mathbf{A}^{(g)}$ . These are all interesting avenues for future work.

## References

1. Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.
2. Mark Newman. *Networks*. Oxford university press, 2018.
3. Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.
4. Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 951–961. International World Wide Web Conferences Steering Committee, 2016.

5. Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 279–287, 2018.
6. Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 115–124, 2016.
7. Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services*, 7(3):1–115, 2015.
8. Caterina De Bacco, Eleanor A Power, Daniel B Larremore, and Cristopher Moore. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317, 2017.
9. Brian Ball, Brian Karrer, and Mark EJ Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.
10. Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.
11. Aaron Schein, Mingyuan Zhou, David Blei, and Hanna Wallach. Bayesian poisson tucker decomposition for learning the structure of international relations. In *International Conference on Machine Learning*, pages 2810–2819. PMLR, 2016.
12. Hadiseh Safdari, Martina Contisciani, and Caterina De Bacco. Generative model for reciprocity and community detection in networks. *Physical Review Research*, 3(2):023209, 2021.
13. Martina Contisciani, Hadiseh Safdari, and Caterina De Bacco. Community detection and reciprocity in networks by jointly modeling pairs of edges. *arXiv preprint arXiv:2112.10436*, 2021.
14. Hadiseh Safdari, Martina Contisciani, and Caterina De Bacco. Reciprocity, community detection, and link prediction in dynamic networks. *Journal of Physics: Complexity*, 3(1):015010, 2022.
15. David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.
16. Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.
17. Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.
18. Martina Contisciani, Eleanor A Power, and Caterina De Bacco. Community detection with node attributes in multilayer networks. *Scientific reports*, 10(1):1–16, 2020.
19. Mark EJ Newman and Aaron Clauset. Structure and inference in annotated networks. *Nature communications*, 7:11863, 2016.
20. Oscar Fajardo-Fontiveros, Roger Guimerà, and Marta Sales-Pardo. Node meta-data can produce predictability crossovers in network inference problems. *Physical Review X*, 12(1):011010, 2022.

## A Asymmetric links

Here, we show how to extend the EXP model to asymmetric networks ( $A_{ij} \neq A_{ji}$ ), while still maintaining a symmetric structure for the exposure ( $\mu_{ij} = \mu_{ji} = \mu_i \mu_j$ ).

We introduce an extra set of  $N \times K$ -dimensional hidden variables,  $v_{ik}$ , which have the same structure of the  $u_{ik}$ . The latent variable set is now  $\theta = (u, v, w)$ . This allows us to replace Eq. (2) with an asymmetric Poisson parameter

$$\lambda_{ij} = \sum_{k,q}^K u_{ik} v_{jq} w_{kq}. \quad (18)$$

As a result, Eqs. (12), (13), (14) and (15) are replaced by

$$u_{ik} = \frac{\sum_j Q_{ij} A_{ij} \sum_q \rho_{ijkq}}{\sum_j Q_{ij} \sum_q v_{jq} w_{kq}} \quad (19)$$

$$v_{iq} = \frac{\sum_j Q_{ij} A_{ij} \sum_k \rho_{ijkq}}{\sum_j Q_{ij} \sum_k v_{jq} w_{kq}} \quad (20)$$

$$w_{kq} = \frac{\sum_{i,j} Q_{ij} A_{ij} \rho_{ijkq}}{\sum_{i,j} Q_{ij} u_{ik} v_{jq}} \quad (21)$$

$$\rho_{ijkq} = \frac{u_{ik} v_{jq} w_{kq}}{\sum_{k,q} u_{ik} v_{jq} w_{kq}} \quad (22)$$

$$\mu_i = \frac{\sum_j Q_{ij}}{\sum_j \frac{(1-Q_{ij}) \mu_j}{(1-\mu_i \mu_j)}}, \quad (23)$$

where now we also take into account the updates for the new hidden variables,  $v_{ik}$ .