Overlapping Communities and Roles in Networks with Node Attributes: Probabilistic Graphical Modeling, Bayesian Formulation and Variational Inference (Extended Abstract)*

Gianni Costa , **Riccardo Ortale**

ICAR-CNR

giovanni.costa@icar.cnr.it, riccardo.ortale@icar.cnr.it

Abstract

We study the seamless integration of community discovery and behavioral role analysis, in the domain of networks with node attributes. In particular, we focus on unifying the two tasks, by explicitly harnessing node attributes and behavioral role patterns in a principled manner. To this end, we propose two Bayesian probabilistic generative models of networks, whose novelty consists in the interrelationship of overlapping communities, roles, their behavioral patterns and node attributes. The devised models allow for a variety of exploratory, descriptive and predictive tasks. These are carried out through mean-field variational inference, which is in turn mathematically derived and implemented into a coordinate-ascent algorithm.

A wide spectrum of experiments is designed, to validate the devised models against three classes of state-of-the-art competitors using various real-world benchmark data sets from different social networking services.

1 Introduction

The integration of community discovery and role analysis can take advantage of node attributes and behavioral role patterns. To elaborate, attributes are in principle beneficial to refine communities and roles, by enabling the detection of the affiliations of nodes to such communities and roles even with few or no interactions with other nodes. In addition, behavioral role patterns are useful to model connections between nodes, in terms of both their involvement in communities and their roles in the participated communities.

Unfortunately, the interrelationship of network communities, roles with respective behavioral patterns and node attributes raises several issues. First, modeling network structure in terms of both communities and behavioral role patterns is problematic. Indeed, community discovery and role analysis provide different explanations of links. More precisely, in community discovery (that typically ignores roles), links are usually explained by community involvement. In

this regard, *pluralistic homophily* was found in [Yang and Leskovec, 2013] to be a realistic model of tie formation, in which links are more likely between nodes with a larger number of affiliations to common communities. Instead, in traditional role analysis (that generally disregards communities), links are commonly interpreted via mixed-membership block-modeling [Anderson *et al.*, 1992; Arabie *et al.*, 1978; Holland *et al.*, 1983; Rossi and Ahmed, 2015]. This amounts to considering links as interactions between node roles, whose strength is determined by underlying behavioral patterns. Second, it is not clear how node attributes are interrelated with communities, roles and respective behavioral patterns. Third, the sought-after interrelationship should be explicative and predictive of network structure.

In this paper, we aim to refine the joint analysis of overlapping communities and behavioral roles in networks, through the incorporation of node attributes and behavioral role patterns. Also, we intend to investigate whether looking simultaneously at network structure and node attributes is advantageous in terms of robustness to noise (or lack of observations) in either of the two properties. Yet, we strive to formulate plausible models explaining the interaction of node attributes with communities, roles and respective behavioral patterns.

In order to pursue the above-mentioned objectives, we develop two new model-based machine-learning approaches [Bishop, 2013; Ghahramani, 2015; Pfeffer, 2016] to the exploratory and unsupervised analysis of overlapping communities, roles and respective behavioral patterns in networks with node attributes [Yang *et al.*, 2013]. Both approaches rely on probabilistic graphical modeling [Bishop, 2006; Koller and Friedman, 2009; Murphy, 2012] to establish conditional (in)dependencies between visible and hidden aspects of networks. The visible aspects include node attributes and links. Both are treated as observed variables. The hidden aspects are treated as latent variables and include behavioral role patterns along with affiliations [Lattanzi and Sivakumar, 2009; Yang and Leskovec, 2013]. These are used to capture the extent of node involvement in communities and roles, the degree of role specificity to communities, as well as the strength of attribute relatedness to either roles or communities.

The above-mentioned conditional (in)dependencies are encoded in two Bayesian generative models of networks, i.e., SCANNER and PERISCOPE, that postulate alternative assumptions on node attributes. Under SCANNER (*StruCture*

---

and Attributes of Nodes from commuNitiEs and Roles), node attributes are affiliated to roles. Instead, under PERISCOPE (*toPology and nodE attRIbuteS from Communities and rOle PattErns*), node attributes are affiliated to communities.

The latent variables of SCANNER and PERISCOPE are estimated in [Costa and Ortale, 2022] through variational inference [Blei *et al.*, 2017; Jordan *et al.*, 1999; MacKay, 2002]. The latter enables exploratory, predictive and descriptive tasks under SCANNER and PERISCOPE. The exploratory tasks include the analysis of communities and roles with respective behavioral patterns as well as the posterior explanation of link formation. The predictive tasks comprise the prediction of unobserved links and attributes. The descriptive tasks involve the interpretation of communities and roles.

In [Costa and Ortale, 2022], SCANNER and PERIS-COPE are extensively tested quantitatively and qualitatively, through a comparative evaluation (in the above tasks) on real-world benchmark data sets from different social networking services. Their scalability is also comparatively investigated.

2 Preliminaries

The notation used in this paper is summarized below along with some fundamental concepts.

2.1 Networks, Communities, Attributes, Roles and Latent Affiliations

The network-based representation of a targeted complex system is a graph $\mathcal{G} = \{\boldsymbol{N}, \boldsymbol{E}\}$, where $\boldsymbol{N} = \{1, \ldots N\}$ is a set of nodes (numbered 1 through N) and $\boldsymbol{E} \subseteq \boldsymbol{N} \times \boldsymbol{N}$ is a set of links. Nodes correspond to entities, which interact in the network (e.g., individuals, organizations and so forth). Links represent interactions between nodes and are summarized via a binary adjacency matrix $\boldsymbol{L}$, whose generic entry $L_{n,n'}$ is 1 iff a link between nodes n and n' is observed and 0 otherwise.

We are interested in the joint modeling and analysis of three primary properties of an input network $\mathcal{G}$, i.e., its community structure, the roles of its nodes and their attributes.

Communities are structures of $\mathcal{G}$, which can be formalized as a set $\boldsymbol{C} \triangleq \{C_1, \ldots, C_K\}$ of K overlapping groups of nodes (i.e., $C_k \subseteq \boldsymbol{N}$ with $k = 1, \ldots, K$). The generic node n can participate in all communities, although with a different involvement. In particular, $\vartheta_{n,k}$ is the degree to which n is involved in the arbitrary community C_k.

Roles are a set $\boldsymbol{R} \triangleq \{R_1, \ldots, R_H\}$ of H behavioral classes (i.e., $R_h \subseteq \boldsymbol{N}$ with $h = 1, \ldots, H$)[1]. The generic node n can play all roles, although with a different attitude. Specifically, $\sigma_{n,h}$ is the extent to which n is suitable for playing the arbitrary role R_h. Besides, roles are specific to communities and the specificity of role R_h to community C_k is $\varphi_{h,k}$.

Let $\boldsymbol{A} \triangleq \{f_1, \ldots, f_A\}$ be a set of A binary attributes. The boolean indicator $F_{n,a}$ (with $1 \leq a \leq A$) is 1 iff node n exhibits attribute f_a and 0 otherwise. Notation $\boldsymbol{F} \triangleq \{F_{n,a} | n \in \boldsymbol{N}, f_a \in \boldsymbol{A}\}$ is the characterization of all nodes in terms of their attributes. In practical applications, $\boldsymbol{L}$ and $\boldsymbol{F}$ are usually (very) sparse. Attributes are related to roles to a varying degree. The relatedness of attribute f_a to role R_h is $\pi_{a,h}$.

---
[1]Notice that, in our formulation, roles are not explicitly observed as structural patterns of node connectivity in $\mathcal{G}$.

Overall, node affiliations to communities and roles are respectively denoted as $\boldsymbol{\Theta} \triangleq \{\vartheta_{n,k} | n \in \boldsymbol{N}, C_k \in \boldsymbol{C}\}$ and $\boldsymbol{\Sigma} \triangleq \{\sigma_{n,h} | n \in \boldsymbol{N}, R_h \in \boldsymbol{R}\}$. Role affiliations to communities are compactly indicated as $\boldsymbol{\Phi} \triangleq \{\varphi_{h,k} | R_h \in \boldsymbol{R}, C_k \in \boldsymbol{C}\}$. Notation $\boldsymbol{\Pi} \triangleq \{\pi_{a,h} | 1 \leq a \leq A, R_h \in \boldsymbol{R}\}$ means the relevance of all attributes to individual roles.

2.2 Behavioral Role Patterns

The establishment of links between nodes is assumed to be influenced (among others) by their common community affiliations, the strength of such affiliations and the respective behavioral role patterns. The combination of such components to rule tie formation is detailed in Section 3. Here, we supplement roles with their corresponding behavioral patterns.

Let R_h be a generic role from $\boldsymbol{R}$. The behavioral role pattern associated with R_h is $\boldsymbol{\epsilon}_h \triangleq \{\epsilon_{h,h'} | R_{h'} \in \boldsymbol{R}\}$, where $\epsilon_{h,h'}$ is the unknown strength of interaction between R_h and $R_{h'}$ (with $1 \leq h' \leq H$). Intuitively, $\boldsymbol{\epsilon}_h$ (with $1 \leq h \leq H$) captures the behavior involved in the fulfillment of the social function ascribed to role R_h. Notation $\boldsymbol{\Upsilon} \triangleq \{\boldsymbol{\epsilon}_h | R_h \in \boldsymbol{R}\}$ succinctly represents the behavioral patterns of all roles.

2.3 Problem Statement

Given an input network $\mathcal{G}$, a number K of communities and a number H of roles, we aim to perform

- the unsupervised exploratory analysis of $\mathcal{G}$, to infer the unknown values of all elements of $\boldsymbol{\Theta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Upsilon}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Pi}$;

- the prediction of missing links between nodes of $\mathcal{G}$.

- the prediction of unobserved node attributes.

The above tasks are accomplished by inferring a posterior distribution over $\boldsymbol{\Theta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Pi}$ given $\mathcal{G}$ (or, equivalently, the adjacency matrix $\boldsymbol{L}$ and attribute matrix $\boldsymbol{F}$). Actually, the true posterior is approximated through variational inference in a generative network model, that explains the formation of $\mathcal{G}$ from a Bayesian probabilistic perspective. Henceforth, all elements of $\boldsymbol{\Theta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Pi}$ are treated as random variables, being unknown and not directly measurable.

3 The SCANNER Model

SCANNER is a Bayesian probabilistic generative model of overlapping communities, roles and respective behavioral patterns in networks with node attributes. SCANNER explains the observed aspects of a network $\mathcal{G}$ (i.e., topology $\boldsymbol{L}$ and node attributes $\boldsymbol{F}$), in terms of its latent aspects, namely

- the affiliations of nodes to communities and roles (i.e. $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$);

- the behavioral role patterns (i.e $\boldsymbol{\Upsilon}$);

- the relevance of roles to communities (i.e $\boldsymbol{\Phi}$);

- the relatedness of attributes to roles (i.e $\boldsymbol{\Pi}$).

Fig. 1 illustrates the conditional (in)dependencies between observed and latent aspects of $\mathcal{G}$ under SCANNER in the form of shaded and unshaded random variables, respectively.

The likelihood of observations $\boldsymbol{L}$ and $\boldsymbol{F}$ is

$$\Pr(\boldsymbol{L}, \boldsymbol{F} | \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}, \boldsymbol{\Upsilon}, \boldsymbol{\Pi}) = \Pr(\boldsymbol{L} | \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}, \boldsymbol{\Upsilon}) \cdot \Pr(\boldsymbol{F} | \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}, \boldsymbol{\Pi}) =$$

$$\prod_{n,n' \in \boldsymbol{N}} \Pr(L_{n,n'} | \boldsymbol{\vartheta}_n, \boldsymbol{\sigma}_n, \boldsymbol{\vartheta}_{n'}, \boldsymbol{\sigma}_{n'}, \boldsymbol{\Phi}, \boldsymbol{\Upsilon}) \cdot$$

$$\prod_{n,a:n \in \boldsymbol{N}, f_a \in \boldsymbol{A}} \Pr(F_{n,a} | \boldsymbol{\vartheta}_n, \boldsymbol{\sigma}_n, \boldsymbol{\Phi}, \boldsymbol{\pi}_a) \tag{1}$$

where $\boldsymbol{\vartheta}_n = \{\vartheta_{n,k} | k \in \boldsymbol{C}\}$, $\boldsymbol{\sigma}_n = \{\sigma_{n,h} | h \in \boldsymbol{H}\}$ and $\boldsymbol{\pi}_a = \{\pi_{a,h} | h \in \boldsymbol{H}\}$. Moreover, the individual observations $\Pr(L_{n,n'} | \boldsymbol{\vartheta}_n, \boldsymbol{\sigma}_n, \boldsymbol{\vartheta}_{n'}, \boldsymbol{\sigma}_{n'}, \boldsymbol{\Phi}, \boldsymbol{\Upsilon})$ and $\Pr(F_{n,a} | \boldsymbol{\vartheta}_n, \boldsymbol{\sigma}_n, \boldsymbol{\Phi}, \boldsymbol{\pi}_a)$ are defined as Poisson distributions with suitable parameterizing rates. The Poisson distribution is well suited to binary data and beneficial to expedite inference on sparse networks [Gopalan *et al.*, 2015]. Moreover, rates naturally allow for the incorporation of the causes deemed to be determining for the emergence of links and node attributes. Formally, for each pair of nodes $n, n' \in \boldsymbol{N}$ and each attribute $f_a \in \boldsymbol{A}$, $\Pr(L_{n,n'} | \boldsymbol{\vartheta}_n, \boldsymbol{\sigma}_n, \boldsymbol{\vartheta}_{n'}, \boldsymbol{\sigma}_{n'}, \boldsymbol{\Phi}, \boldsymbol{\Upsilon}) \triangleq Poisson(L_{n,n'} | \delta_{n,n'})$ and $\Pr(F_{n,a} | \boldsymbol{\vartheta}_n, \boldsymbol{\sigma}_n, \boldsymbol{\Phi}, \boldsymbol{\pi}_a) \triangleq Poisson(F_{n,a} | \delta_{n,a})$, where

$$\delta_{n,n'} = \sum_{k=1}^{K} \sum_{h,h'=1}^{H} \vartheta_{n,k} \cdot \sigma_{n,h} \cdot \varphi_{h,k} \cdot \epsilon_{h,h'} \cdot \vartheta_{n',k} \cdot \sigma_{n',h'} \cdot \varphi_{h',k} \tag{2}$$

$$\delta_{n,a} = \sum_{k=1}^{K} \sum_{h=1}^{H} \vartheta_{n,k} \cdot \varphi_{h,k} \cdot \sigma_{n,h} \cdot \pi_{a,h} \tag{3}$$

The parameterizing rate $\delta_{n,n'}$ at Eq. 2 captures the latent interaction in $\mathcal{G}$ between n and n', through their affiliations to common communities and respective roles. Remarkably, Eq. 2 seamlessly reconciles pluralistic homophily with mixed-membership block-modeling from traditional role analysis. Hence, the formation of ties between nodes is influenced by their affiliations to shared communities, the strength of affiliation to such communities and respective roles, the specificity of roles to the participated communities as well as the patterns of interactions between roles.

The Poisson rate $\delta_{n,a}$ at Eq. 3 is meant to rule the inclusion of f_a as an attribute of node n, according to the strength with which n is affiliated to communities and roles, the specificity of roles to communities and the relevance of f_a to roles.

In compliance with conjugacy [Bishop, 2006], Gamma priors are placed on the individual random variables corresponding to the different latent aspects of $\mathcal{G}$. More precisely, for each $n \in \boldsymbol{N}$, $C_k \in \boldsymbol{C}$, $R_h, R_{h'} \in \boldsymbol{H}$ and $f_a \in \boldsymbol{A}$,

$$\Pr(\vartheta_{n,k} | \alpha_\vartheta, \beta_\vartheta) \triangleq Gamma(\vartheta_{n,k} | \alpha_\vartheta, \beta_\vartheta)$$

$$\Pr(\sigma_{n,h} | \alpha_\sigma, \beta_\sigma) \triangleq Gamma(\sigma_{n,h} | \alpha_\sigma, \beta_\sigma)$$

$$\Pr(\varphi_{h,k} | \alpha_\varphi, \beta_\varphi) \triangleq Gamma(\varphi_{h,k} | \alpha_\varphi, \beta_\varphi)$$

$$\Pr(\pi_{a,h} | \alpha_\pi, \beta_\pi) \triangleq Gamma(\pi_{a,h} | \alpha_\pi, \beta_\pi)$$

$$\Pr(\epsilon_{h,h'} | \alpha_\epsilon, \beta_\epsilon) \triangleq Gamma(\epsilon_{h,h'} | \alpha_\epsilon, \beta_\epsilon)$$

where α_ψ and β_ψ with $\psi \in \{\vartheta, \sigma, \varphi, \pi, \epsilon\}$ are, respectively, shape and rate hyperparameters.

Notably, the above Gamma distributions enforce nonnegativity both on affiliation strengths and behavioral role patterns. This improves the interpretability of SCANNER by
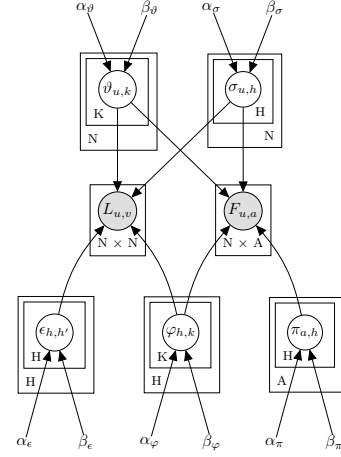

Figure 1: Directed graphical representation of SCANNER

favoring sparseness in its representation. Additionally, it ensures that the overall strength of affiliation for each node, role and attribute does not amount to 1, thus avoiding the typical inconvenient of conventional mixed-membership modeling. Therefore, a very strong affiliation of any node, role or attribute does not imply a corresponding drop in the overall strength of all other affiliations of that node, role or attribute.

Under SCANNER, a network $\mathcal{G}$ is assumed to be generated through a sequence of interactions among the random variables of Fig. 1. As detailed in Fig. 2, such a process consists of six steps. Steps I through IV preliminarily deal with the realization of the latent affiliations and behavioral role patterns. Steps V through VI eventually generate $\mathcal{G}$ through the realization of the observed links and node attributes. More precisely, at steps I, II and III, the generative process samples the strength of the latent affiliations covered in Section 2.1 from the respective Gamma priors. At step IV, another Gamma prior is again sampled to draw the behavioral role patterns of Section 2.2. At step V, the presence/absence of a link between any two nodes is sampled from the Poisson distribution, that is placed over $\boldsymbol{L}$ as the link data likelihood. Lastly, at step VI, the presence/absence of attribute f_a in the characterization of node n is sampled from the Poisson distribution, which is placed over $\boldsymbol{F}$ as the attribute data likelihood.

4 Tasks

Fitting SCANNER to an input network $\mathcal{G}$ enables the accomplishment of a wide variety of exploratory, predictive and descriptive tasks, including those enumerated at Section 2.3 (please see [Costa and Ortale, 2022] for further details).

4.1 Exploratory Network Analysis

Node affiliations reveal the underlying organization of $\mathcal{G}$ into overlapping communities and roles. The degree to which node n participates in community C_k is $\vartheta^*_{n,k} \triangleq E[\vartheta_{n,k}]$. Moreover, the extent to which node n is suitable for playing role R_h is $\sigma^*_{n,h} \triangleq E[\sigma_{n,h}]$.

Community structure and roles allow the posterior explanation of links. Let n and n' be two connected no-

I) For each node $n \in \boldsymbol{N}$
- For each community $C_k \in \boldsymbol{C}$
  - draw the degree $\vartheta_{n,k}$ to which n is involved in community C_k, i.e., $\vartheta_{n,k} \sim \mathsf{Gamma}(\vartheta_{n,k}|\alpha_\vartheta, \beta_\vartheta)$.
- For each role $R_h \in \boldsymbol{R}$
  - draw the extent $\sigma_{n,h}$ to which n is suitable to play role R_h, i.e., $\sigma_{n,h} \sim \mathsf{Gamma}(\sigma_{n,h}|\alpha_\sigma, \beta_\sigma)$.

II) For each role $R_h \in \boldsymbol{R}$
- For each community $C_k \in \boldsymbol{C}$
  - draw the relevance of role R_h to community C_k, i.e., $\varphi_{h,k} \sim \mathsf{Gamma}(\varphi_{h,k}|\alpha_\varphi, \beta_\varphi)$.

III) For each attribute $f_a \in \boldsymbol{A}$
- For each role $R_h \in \boldsymbol{R}$
  - draw the relatedness of attribute f_a to role R_h, i.e., $\pi_{a,h} \sim \mathsf{Gamma}(\pi_{a,h}|\alpha_\pi, \beta_\pi)$.

IV) For each role $R_h \in \boldsymbol{R}$
- For each role $R_{h'} \in \boldsymbol{R}$
  - draw the strength of interaction between role R_h and role $R_{h'}$, i.e., $\epsilon_{h,h'} \sim \mathsf{Gamma}(\epsilon_{h,h'}|\alpha_\epsilon, \beta_\epsilon)$.

V) For each node $n \in \boldsymbol{N}$
- For each node $n' \in \boldsymbol{N}$
  - draw the presence/absence of a link $L_{n,n'}$ between n and n', i.e., $L_{n,n'} \sim \mathsf{Poisson}(L_{n,n'}|\delta_{n,n'})$, where $\delta_{n,n'}$ is the rate defined by Eq. 2.

VI) For each node $n \in \boldsymbol{N}$
- For each attribute $f_a \in \boldsymbol{A}$
  - draw whether or not attribute f_a characterizes node n, i.e., $F_{n,a} \sim \mathsf{Poisson}(F_{n,a}|\delta_{n,a})$, where $\delta_{n,a}$ is the rate defined by Eq. 3.

Figure 2: The generative process under SCANNER

des of $\mathcal{G}$, i.e., such that $\langle n, n' \rangle \in \boldsymbol{E}$ and, consequently, $L_{n,n'} = 1$. Also, assume that notation $z_{n,n'}^{(k,h,h')} = 1$ means that $z_{n,n'}^{(k,h,h')}$ is the only Poisson count in $\boldsymbol{z}_{n,n'}$ amounting to 1, whereas all other counts are zero. The probability $P(z_{n,n'}^{(k,h,h')} = 1|L_{n,n'}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}, \boldsymbol{\Upsilon}, \boldsymbol{\Phi})$ that $\langle n, n' \rangle$ was established with n and n' playing, respectively, roles R_h and $R_{h'}$ within community k is approximated by $E[\tau_{n,n'}^{(k,h,h')}]$, according to the mean-field approximation. Therefore, $\langle n, n' \rangle$ can be reasonably explained by the triple (k^*, h^*, h'_*), such that $(k^*, h^*, h'_*) = \arg\max_{(k,h,h')} E[\tau^{(k,h,h')}]$.

Behavioral patterns supplement roles, by unveiling their functioning in the fulfillment of the respective social functions. Here, the strength of interaction between roles R_h and $R_{h'}$ is $\epsilon_{h,h'}^* \triangleq E[\epsilon_{h,h'}]$.

4.2 Predictive Analysis

The emergence of missing links and node attributes is forecast by associating them with a respective score, that is used to rank their prospective observation. In particular, the score $s_{n,n'}$ associated with unconnected pairs of nodes n and n' is the posterior expectation of the corresponding Poisson-distribution rate. More precisely, $s_{n,n'} \triangleq E[\delta_{n,n'}]$, where $\delta_{n,n'}$ is defined by Eq. 2.

Analogously, the score $s_{n,a}$ regarding the unobserved attribute f_a of node n is the posterior expectation of the respective Poisson-distribution rate. Specifically, $s_{n,a} \triangleq E[\delta_{n,a}]$, where $\delta_{n,a}$ is defined by Eq. 3.

4.3 Descriptive Analysis

SCANNER supports the interpretation of roles and communities by discriminatory characterizations, therefore roles can be characterized by related node attributes. In fact the re-

latedness of attribute f_a to role R_h is $\pi_{a,h}^* \triangleq E[\pi_{a,h}]$. A discriminatory characterization of R_h is, hence, any suitable subset of strongly related attributes.

Communities can be characterized by relevant role affiliations (directly) and pertinent node attributes (indirectly). The relevance of role R_h to community C_k is $\varphi_{h,k}^* \triangleq E[\varphi_{h,k}]$. In addition, the pertinence of attribute f_a to community C_k is $\omega_{a,k}^* \triangleq E[\sum_{h=1}^H \pi_{a,h}\varphi_{h,k}]$. Accordingly, C_k can be discriminatorily characterized in terms of highly relevant roles and/or especially pertinent attributes.

5 The PERISCOPE Model

SCANNER assumes that node attributes are directly affiliated to roles. By contrast, PERISCOPE is a new model conceived to formulate and investigate an important and otherwise unexplored conjecture, i.e., the direct affiliation of node attributes to communities instead of roles. The directed graphical representation of the conditional (in)dependencies between observed and latent variables under PERISCOPE is reported in [Costa and Ortale, 2022] along with the corresponding generative process. Here, we only emphasize that, under PERISCOPE, the inclusion of f_a among the attributes of node n is governed by the below rate

$$\delta_{n,a} \triangleq \sum_{k=1}^K \sum_{h=1}^H \vartheta_{n,k} \cdot \varphi_{h,k} \cdot \sigma_{n,h} \cdot \pi_{a,k} \tag{4}$$

6 Conclusions

We presented two new approaches to the exploratory and unsupervised analysis of communities and behavioral roles in networks, whose novelty consists in accounting for node attributes and behavioral role patterns. The devised approaches are based on as many Bayesian probabilistic latent-factor models of networks, namely SCANNER and PERISCOPE. Both assume a generative process, in which the observed network topology and node attributes are statistically dependent on latent network aspects. These include node affiliations to overlapping communities and roles, role affiliations to communities, attribute affiliation either to roles (under SCANNER) or to communities (under PERISCOPE) as well as behavioral role patterns.

An extensive experimentation is carried out on real-world benchmark data sets in [Costa and Ortale, 2022]. Tests reveal that the devised approaches are more accurate in community discovery, link prediction and attribute prediction than several competitors. The devised approaches are also found more accurate in community discovery and link prediction, despite noise or lack of observations in network structure or node attributes. This finding substantiates that the unified modeling and processing of network structure and node attributes makes both the approaches more robust to perturbations. Besides accuracy and sensitivity to perturbations, the scalability of our approaches is also studied in [Costa and Ortale, 2022].

References

[Anderson *et al.*, 1992] Carolyn J. Anderson, Stanley Wasserman, and Katherine Faust. Building stochastic blockmodels. *Social Networks*, 14(1-2):137 – 161, 1992.

[Arabie *et al.*, 1978] Phipps Arabie, Scott A. Boorman, and Paul R Levitt. Constructing blockmodels: How and why. *Journal of Mathematical Psychology*, 17(1):21 – 63, 1978.

[Bishop, 2006] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Bishop, 2013] Christopher M. Bishop. Model-based machine learning. *Philosophical Transactions of the Royal Society A*, page 371:20120222, 2013.

[Blei *et al.*, 2017] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859 – 877, 2017.

[Costa and Ortale, 2022] Gianni Costa and Riccardo Ortale. Overlapping communities and roles in networks with node attributes: Probabilistic graphical modeling, bayesian formulation and variational inference. *Artificial Intelligence*, 302: 103580, 2022.

[Ghahramani, 2015] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452 – 459, 2015.

[Gopalan *et al.*, 2015] Prem Gopalan, Jake M. Hofman, and David M. Blei. Scalable recommendation with hierarchical poisson factorization. In *UAI*, pages 326 – 335, 2015.

[Holland *et al.*, 1983] Paul W. Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137, 1983.

[Jordan *et al.*, 1999] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183 – 233, 1999.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models. Principles and Techniques*. The MIT Press, 2009.

[Lattanzi and Sivakumar, 2009] Silvio Lattanzi and D. Sivakumar. Affiliation networks. In *ACM STOC*, pages 427 – 434, 2009.

[MacKay, 2002] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2002.

[Murphy, 2012] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[Pfeffer, 2016] Avi Pfeffer. *Practical Probabilistic Programming*. Manning Publications, 2016.

[Rossi and Ahmed, 2015] Ryan A. Rossi and Nesreen K. Ahmed. Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(04):1112 – 1131, 2015.

[Yang and Leskovec, 2013] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: A non-negative matrix factorization approach. In *ACM WSDM*, pages 587 – 596, 2013.

[Yang *et al.*, 2013] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *ICDM*, pages 1151 – 1156, 2013.