The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Bayesian Robust Attributed Graph Clustering: Joint Learning of Partial Anomalies and Group Structure

Aleksandar Bojchevski, Stephan Günnemann
Technical University of Munich, Germany
{a.bojchevski, guennemann}@in.tum.de

Abstract

We study the problem of robust attributed graph clustering. In real data, the clustering structure is often obfuscated due to anomalies or corruptions. While robust methods have been recently introduced that handle anomalies as part of the clustering process, they all fail to account for one core aspect: Since attributed graphs consist of two views (network structure and attributes) anomalies might materialize only partially, i.e. instances might be corrupted in one view but perfectly fit in the other. In this case, we can still derive meaningful cluster assignments. Existing works only consider complete anomalies. In this paper, we present a novel probabilistic generative model (PAICAN) that explicitly models partial anomalies by generalizing ideas of Degree Corrected Stochastic Block Models and Bernoulli Mixture Models. We provide a highly scalable variational inference approach with runtime complexity linear in the number of edges. The robustness of our model w.r.t. anomalies is demonstrated by our experimental study, outperforming state-of-the-art competitors.

Introduction

Clustering of attributed graphs – a special type of multi-view data – has become an important research field (Bothorel et al. 2015), with application domains from social networks, over e-commerce, to gene analysis. By simultaneously utilizing both network structure and attribute information clustering results can be improved. In real life scenarios, these datasets are often polluted by rare occurrences, anomalies or corruptions. A spammer, for example, might be trying to connect to as many nodes as possible, inducing spurious edges and thus obscuring the real group structure in the data. Another source of anomalies are for example users on a social network obfuscating some of their attributes (age, political affiliation) on purpose due to privacy concerns. Since these anomalies hinder the cluster detection, robust attributed graph clustering methods have been proposed (Perozzi et al. 2014; Gao et al. 2010). Instead of first applying anomaly detection for attributed graphs (Akoglu, Tong, and Koutra 2015), followed by the actual clustering on the remaining data, anomaly detection and clustering are performed simultaneously. Such joint learning has shown high performance

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

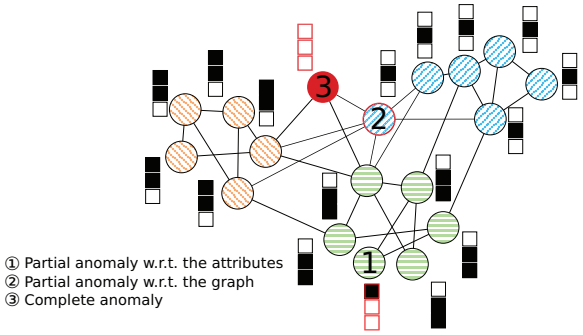


Figure 1: Three types of anomalies in attributed graphs.

for many other tasks such as regression (Rousseeuw and Leroy 2005), PCA (Wright et al. 2009; Candès et al. 2011), matrix factorization (Xiong, Chen, and Schneider 2011), and autoregression (Günnemann, Günnemann, and Faloutsos 2014). The big challenge – not sufficiently captured by the existing works – is that anomalies in attributed graphs materialize in different ways. Specifically one has to take into account challenging *camouflage behavior*: A user in a social network for example might show corrupted attributes (to, e.g., hide its identity) but still their friendship relations are normal. That is, the user is corrupted in only one of the views – we call this a *partial* anomaly. Another example of a partial anomaly is a paper in a citation network, where the content of the paper fits well in some cluster, but the relevant citations are missing. This can especially happen in a new and emerging subfield where not everyone is yet aware of the latest literature. Fig. 1 illustrates this principle in general. Node 1 fits nicely into the network structure but has completely different attributes compared to the other nodes in its group. On the other hand, node 2 perfectly fits the assigned cluster if we only look at its attributes, but it is obviously an anomaly with regards to the network structure. *The crucial observation is that we can still identify the partial anomalies' latent cluster since two sources of information are given.* In the social network example, despite the users' corrupted attributes we are still able to derive their cluster, thus, enabling downstream tasks such as e.g. targeted marketing. We observed such partial anomalies in a variety of

real-world datasets.

Existing works (Perozzi et al. 2014; Gao et al. 2010) fail in handling partial anomalies. As soon as a node is corrupted in one of the views, it is marked as an anomaly and no longer belongs to any cluster, even though it might perfectly fit in the other view. Simply speaking, the benefit of having *both* network structure and associated attributes is not taken into account for anomaly detection in existing works. Solving this limitation, we propose a model for attributed graph clustering that accounts for *partial anomalies*: i.e., a node may be corrupted in one space but not in the other. As a strong benefit of this – and in contrast to all existing works – we are still able to infer a node's group assignment even if it is (partially) corrupted. Thereby, we not only obtain more informative results, but we also enable a comparison between the nodes' observed and expected information. E.g. for node 1 we observe attributes (1,0,0) but would expect (0,1,1) due to its cluster membership.[1] Clearly, our model also handles complete anomalies such as node 3, which does not fit to any group with neither attribute nor graph space.

To realize these ideas, we propose a novel probabilistic generative model for attributed graphs, PAICAN (Partial Anomaly Identification and Clustering in Attributed Networks). We jointly model (a) the attribute and network space, as well as (b) the latent group assignments and anomaly detection by introducing a generalized, anomaly-aware Degree-Corrected Stochastic Block Models (DCSBM) combined with a Beta-Bernoulli mixture model. The main contributions of this work are:

- *Robustness and Partial Anomalies:* A novel probabilistic generative model that jointly performs clustering and anomaly detection in attributed graphs. It is the first work that realizes robust clustering for attributed graphs following a power law degree distribution, thus capturing real-life properties. Our model further takes into account that nodes might be only partially anomalous, thus, enabling us to assign partially anomalous nodes to meaningful clusters.
- *Scalable algorithm:* Using variational inference and exploiting special properties of our model we propose an algorithm with runtime complexity $\mathcal{O}(\#edges)$. Our variational formulation enables us to reason about the uncertainty of the cluster and the anomaly assignments via their posterior distributions.

Related work

Clustering attributed graphs has attracted strong attention. For a general overview we refer to (Bothorel et al. 2015). In line with the focus of this paper, we here describe primarily works with the following aspects: robustness to anomalies and principled probabilistic generative models.

So far, only two approaches jointly perform clustering and anomaly detection in attributed graphs: CODA (Gao et al. 2010) and FocusCO (Perozzi et al. 2014). Both detect complete anomalies only. They do not exploit the fact that instances can be partially corrupted. CODA has the additional disadvantages of poor scalability and high sensitivity to hyper-parameter choice and initialization, thus, requiring multiple restarts. FocusCO being a semi-supervised method needs labeled data as examples of similar nodes. In contrast, our technique does not require supervision. **We compare against both techniques in our experimental study.**

Further approaches have been introduced that follow the spirit of generative models, but are *not* robust to anomalies. Note that it is not sufficient to simply treat anomalies as an additional cluster since anomalies might not show specific clustering behavior. Therefore, we cannot simply use these non-robust techniques for anomaly detection. Among the non-robust methods, PICS (Akoglu et al. 2012), CESNA (Yang, McAuley, and Leskovec 2013) and SIAN (Newman and Clauset 2016) only derive point-estimates of the learned parameters. BAGC/GBAGC (Xu et al. 2012; 2014) learns a posterior distribution over the model parameters; however, it does not account for the frequently observed power law distribution of node degrees. LSBM (Hric, Peixoto, and Fortunato 2016) uses agglomerative multilevel MCMC for inference and hence also learns a posterior distribution. So far, only SIAN, LSBM, and our PAICAN handle realistic network structure, all by relying on variants of DCSBMs. Even though none of the above approaches is able to handle scenarios of corrupted data, **we compare against PICS, BAGC, SIAN, and LSBM in our experimental study.**

Recently, the related task of multi-view anomaly detection for vector data has been proposed (Iwata and Yamada 2016), where instances which behave differently across views are detected. Our model of partial anomalies can capture such behavior – with the additional benefit of performing clustering as well. Likewise, our approach handles classical anomaly detection where instances show an overall unusual behavior.

Focusing on a different notion of robustness, various methods for subspace clustering on attributed graphs have been introduced (Günnemann et al. 2013; Günnemann, Boden, and Seidl 2012). Their goal is to derive robust clustering solutions even if subsets of the attributes are noisy. They, however, do not consider anomalies. Furthermore, while this paper is concerned with (partially) anomalous nodes, other works have proposed robust clustering methods handling corruptions in the edge structure (Bojchevski, Matkovic, and Günnemann 2017; Huang et al. 2011).

The PAICAN model

Let G be an undirected attributed graph with N nodes, and let A_{ij} be an element of the adjacency matrix $A \in \{0,1\}^{N \times N}$ of the graph. We denote with $X \in \{0,1\}^{N \times D}$ the attribute matrix where for each node i, X_i is a D-dimensional vector of binary attributes. We denote with K the number of groups to detect. An overview of the probabilistic generative model is given in Fig. 2. Note that the latent variables $\mathbf{c}$ and $\mathbf{z}$ are *shared between the graph and attribute space*.

Partial anomalies. The latent variables $\mathbf{c} = \{c_i\}_{i=1}^N$ indi-

[1] Interestingly, (Gao et al. 2010) illustrates an example similar to node 1; though, still fails to derive a cluster assignment.

cate if a node is (partially) anomalous. Given the two views in an attributed graph, anomalies might materialize in different ways, as indicated by the table below. Accordingly, we define $c_i \sim Categorical(\boldsymbol{\rho})$, $\boldsymbol{\rho} \sim Dirichlet_4(\boldsymbol{\beta})$ where $\boldsymbol{\rho}$ is the usual Dirichlet prior. To simplify the notation we introduce the following two shortcuts: $c_i^A = 0$ if $c_i \in \{0, 2\}$, else 1; indicating whether the node is good or anomalous w.r.t. the graph and similarly regarding the attribute space: $c_i^X = 0$ if $c_i \in \{0, 1\}$, else 1.

	graph	
attributes	good	anomaly
good	$c_i = 0$	$c_i = 1$
anomaly	$c_i = 2$	$c_i = 3$

The variables $\mathbf{z} = \{z_i\}_{i=1}^N$ encode the group assignment of nodes: $z_i|c_i \sim Categorical(\boldsymbol{\pi})$, $\boldsymbol{\pi} \sim Dirichlet_K(\boldsymbol{\alpha})$. Note that z_i is defined if and only if $c_i \neq 3$, that is we can only reason about the group assignment of node i if it's not a complete anomaly.

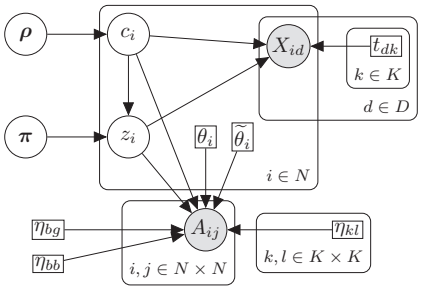


Figure 2: Probabilistic Graphical Model of PAICAN.

Graph model

To incorporate anomalies into the graph structure we propose an anomaly-aware DCSBM, as a generalization of the well-established DCSBM (Karrer and Newman 2011; Yan et al. 2014). The probability of an edge between two nodes i and j is defined as: [2]

$$A_{ij} \sim \begin{cases} Poisson(\theta_i\theta_j\eta_{z_iz_j}) & i<j, c_i^A=0, c_j^A=0 \\ Poisson(\frac{1}{2}\theta_i^2\eta_{z_iz_i}) & i=j, c_i^A=0, c_j^A=0 \end{cases} \text{Case 1}$$

$$\begin{cases} Poisson(\widetilde{\theta}_i\eta_{bg}) & i<j, c_i^A=1, c_j^A=0 \\ Poisson(\widetilde{\theta}_j\eta_{bg}) & i<j, c_i^A=0, c_j^A=1 \end{cases} \text{Case 2}$$

$$\begin{cases} Poisson(\widetilde{\theta}_i\widetilde{\theta}_j\eta_{bb}) & i<j, c_i^A=1, c_j^A=1 \\ Poisson(\frac{1}{2}\widetilde{\theta}_j^2\eta_{bb}) & i=j, c_i^A=1, c_j^A=1 \end{cases} \text{Case 3}$$

Case 1: Both nodes are good. If both considered nodes are good, we refer to a classical DCSBM. Using DCSBM as our base model we can capture diverse connection patterns and network topologies such as assortativity, homophily/heterophily, bipartite graphs, etc. The matrix of group edge probabilities called the block matrix is denoted with $\boldsymbol{\eta} \in [0, 1]^{K\times K}$, and $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^N$ is the vector representing the *latent degrees* of the nodes. Nodes with higher (latent) degree are more likely to form an edge, thus, enabling us to represent networks with a power-law degree distribution.

Case 2: Only one node is anomalous. If exactly one of the nodes is anomalous (e.g. a spammer i tries to establish a connection with a regular user j), we argue as follows: As in a DCSBM, there might be anomalous nodes which try to establish more connections than other anomalies. Thus, it is reasonable to account for different *(latent) degrees of anomalies*, indicated by $\widetilde{\boldsymbol{\theta}} = \{\widetilde{\theta}_i\}_{i=1}^N$. However, since the anomalous connection itself is often originated by the anomaly – i.e. a normal user in a social network is not really interested in establishing a connection to a spammer – a high latent degree θ_j of the good node should not be taken into account. Meaning, an anomaly does not specifically prefer nodes with a high degree but uniformly establishes connections to other nodes. Accordingly, only $\widetilde{\theta}$ of the anomalous node is considered. Additionally, similar to $\boldsymbol{\eta}$, the parameter η_{bg} denotes the base probability of an edge between any anomalous/bad and good node.

Case 3: Both nodes are anomalous. If both nodes are anomalous, we do not assume any specific clustering behavior. Instead we assume a basic connectivity model which takes into account the nodes' latent degrees $\widetilde{\boldsymbol{\theta}}$ as well as some base probability η_{bb} that denotes the probability of any two anomalous nodes forming an edge.

Discussion of $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$. The graph model defined above is not only intuitive but also fulfills two interesting properties: The Maximum-Likelihood estimate of $\widetilde{\theta}_i$ corresponds to the *observed degree* of the corruption. Similarly, the MLE for θ_i is the number of 'good' neighbors of node i (i.e. i's degree w.r.t. the good nodes; anomalies are excluded).[3]

Attribute model

We use an (anomaly-aware) Bernoulli mixture model (BMM). Let $\mathbf{t} \in [0, 1]^{K\times D}$ be the matrix of mixture probabilities, where t_{dk} represent the probability of attribute d having a value of 1 for the nodes in group k, we obtain $X_{id} \sim \begin{cases} Bernouli(t_{dz_i}) & c_i^X = 0 \\ Bernouli(0.5) & c_i^X = 1 \end{cases}$. If the node is good ($c_i^X = 0$) this is a standard BMM. Otherwise we can draw no conclusions about the distribution and we pick the least informative parameter for the Bernoulli distribution of 0.5. The inferred probabilities t_{dk}, yield insight into the importance of different attributes for different groups, and in the context of text

[2] Since we consider undirected graphs, we only need to consider $i \leq j$. As discussed in (Yan et al. 2014), in the sparse regime, the Poisson distribution represents the Bernoulli model well and simplifies the derivations. Note that the well established DCSBM and follow up works are indeed based on the usual (non-truncated) Poisson distribution. Thus, we use the same in our work.

[3] From a generative perspective, a node is either good or anomalous (regarding the graph structure). Thus, for each instance either only $\widetilde{\theta}_i$ or θ_i is used. Hence, in principle, we can combine both vectors $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$ to a single one. However, since later in our learning procedure we compute each node's posterior distribution (i.e. each node is good/anomalous with a specific probability), it is beneficial to model both variables separately.

data, $\mathbf{t}$ takes the role of a topic distribution. It is trivial to extend this work to numerical attributes via e.g. Gaussian Mixture Models.

Posterior inference

We are interested in the posterior distribution of the latent variables $\mathbf{z}$ and $\mathbf{c}$ as well as point-estimates for the remaining parameters – MAP estimates for the latent variables $(\boldsymbol{\pi}, \boldsymbol{\rho})$ and MLE for $(\boldsymbol{\eta}, \eta_{bg}, \eta_{bb}, \boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}, \mathbf{t})$. For inference, we employ a mean-field variational approximation (MFVI), i.e. we learn a variational distribution q aiming to maximize the evidence lower bound (ELBO) (Bishop 2006): $\mathcal{L} = \mathbb{E}_q[\log p(A, X, \mathbf{z}, \mathbf{c}| \ldots)] - \mathbb{E}_q[\log q(\mathbf{z}, \mathbf{c})]$. Our coordinate ascent MFVI algorithm has closed form locally optimal updates and is theoretically guaranteed to converge to a local optimum. We use the following mean-field family:

$$q(\mathbf{z}, \mathbf{c}|\boldsymbol{\psi}, \boldsymbol{\phi}) = \prod_i q(z_i|\psi_i) \prod_i q(c_i|\phi_i)$$

$$s.t.\ q(z_i|\psi_i) \sim Categorical(\psi_i), \quad q(c_i|\phi_i) \sim Categorical(\phi_i)$$

where the free variational parameters $\psi_i \in [0,1]^K, \phi_i \in [0,1]^4$ satisfy $\sum_{k=1}^K \psi_{ik} = 1, \sum_{m=0}^3 \phi_{im} = 1$. As shortcuts for later use, we define $\phi_{i0}^A = \phi_0 + \phi_2$, $\phi_{i1}^A = \phi_1 + \phi_3$, $\phi_{i0}^X = \phi_0 + \phi_1$, $\phi_{i1}^X = \phi_2 + \phi_3$, denoting whether node i is corrupted or not in graph/attribute space.

Given our model, the ELBO decomposes as follows:

$$\mathcal{L} = \underbrace{\mathbb{E}_q[\log p(A|\mathbf{z}, \mathbf{c}, \boldsymbol{\eta}, \eta_{bg}, \eta_{bb}, \boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})]}_{:=\mathcal{L}_\mathcal{A}} + \underbrace{\mathbb{E}_q[\log p(X|\mathbf{z}, \mathbf{c}, \mathbf{t})]}_{:=\mathcal{L}_\mathcal{X}} \quad (1)$$
$$+ \mathbb{E}_q[\log p(\mathbf{z}|\mathbf{c}, \boldsymbol{\pi})] + \mathbb{E}_q[\log p(\mathbf{c}|\boldsymbol{\rho})] - \mathbb{E}_q[\log q(\mathbf{z}, \mathbf{c})]$$

The last four terms are straightforward (see supp. material) and can all be evaluated in linear time w.r.t. the number of nodes and dimensions. For $\mathcal{L}_\mathcal{A}$ we obtain:

$$\mathcal{L}_\mathcal{A} = \sum_{i<j} \Big[\sum_{k,l} \psi_{ik}\psi_{jl}\phi_{i0}^A\phi_{j0}^A \big(A_{ij} \log(\theta_i\theta_j\eta_{kl}) - \theta_i\theta_j\eta_{kl}\big)$$
$$+ \phi_{i1}^A\phi_{j0}^A \big(A_{ij} \log(\widetilde{\theta}_i\eta_{bg}) - \widetilde{\theta}_i\eta_{bg}\big) + \phi_{i0}^A\phi_{j1}^A \big(A_{ij} \log(\widetilde{\theta}_j\eta_{bg}) - \widetilde{\theta}_j\eta_{bg}\big)$$
$$+ \phi_{i1}^A\phi_{j1}^A \big(A_{ij} \log(\widetilde{\theta}_i\widetilde{\theta}_j\eta_{bb}) - \widetilde{\theta}_i\widetilde{\theta}_j\eta_{bb}\big)\Big] + \sum_i \Big[\sum_k \psi_{ik}\phi_{i0}^A \big($$
$$A_{ii} \log(\frac{1}{2}\theta_i^2\eta_{kk}) - \frac{1}{2}\theta_i^2\eta_{kk}\big) + \phi_{i1}^A \big(A_{ii} \log(\frac{1}{2}\widetilde{\theta}_i^2\eta_{bb}) - \frac{1}{2}\widetilde{\theta}_i^2\eta_{bb}\big)\Big] \quad (2)$$

While this term seems to be quadratic in the number of nodes, which is impractical for large networks, we will derive a method that is **linear** in the number of edges.

Variational expectation-maximization

We use variational expectation-maximization (EM) (Bishop 2006). That is, we use an iterative update scheme: in the variational E-step we find the optimal variational parameters of q (Eq. 4 - 5); and in the variational M-step we compute MAP/ML estimates for the remaining parameters regarding the ELBO (Eq. - 7); repeated until convergence. Due to space restrictions, we present here the equations for graphs without self-loops ($A_{ii} = 0$). Full derivations and proofs are available in the supp. material.

We first note **one core result** which is crucial to obtain linear complexity in the number of edges: Given the MLE/MAP estimates as derived in the M-Step, it holds:

$$\textstyle\sum_j \sum_l \psi_{jl}\phi_{j0}^A\theta_j\eta_{kl} = 1, \forall k \quad (3)$$

This result helps to obtain an efficient computation of the first term in Eq. (2).

E-Step: Update of ψ (i.e. z) and ϕ (i.e. c). We employ coordinate ascent, i.e. we optimize each variational parameter while holding the others fixed. In this case, we can derive closed form updates for the optimal parameters (see (Bishop 2006) (Ch. 10)). The optimal variational parameters for ψ_{ik} are:

$$\psi_{ik}^{new} \propto \exp\Big(\phi_{i0}^A \Big[\sum_{j\in\mathcal{N}_i} \phi_{j0}^A \sum_l \psi_{jl} \log(\theta_i\theta_j\eta_{kl}) - \theta_i - \frac{1}{2}\theta_i^2\eta_{kk} +$$
$$\theta_i^2\phi_{i0}^A \sum_l \psi_{il}\eta_{kl}\Big] + \phi_{i0}^X \sum_d \log \mathrm{Ber}(X_{id}|t_{dk}) + (1-\phi_{i3}) \log\pi_k\Big) \quad (4)$$

Here, we defined $\mathcal{N}_i$ as the set of neighbors of i and used the result of Eq. 3. Normalizing them to 1, i.e. $\sum_k \psi_{ik}^{new} = 1$, gives the final update.

Similarly, for the anomaly assignments ϕ_{im}:

$$\phi_{i0}^{new} \propto \exp(\hat{\phi}_{i0}^A + \hat{\phi}_{i0}^X + \log\rho_0)$$
$$\phi_{i1}^{new} \propto \exp(\hat{\phi}_{i1}^A + \hat{\phi}_{i0}^X + \log\rho_1)$$
$$\phi_{i2}^{new} \propto \exp(\hat{\phi}_{i0}^A + \hat{\phi}_{i1}^X + \log\rho_2)$$
$$\phi_{i3}^{new} \propto \exp(\hat{\phi}_{i1}^A + \hat{\phi}_{i1}^X + \log\rho_3 - \sum_k \psi_{ik}\log\pi_k) \quad (5)$$

Here, the updates are based on the following terms regarding the attribute space:

$$\hat{\phi}_{i0}^X = \sum_k \psi_{ik}\Big(\sum_d \log \mathrm{Ber}(X_{id}|t_{dk})\Big) \quad \hat{\phi}_{i1}^X = D\log(0.5)$$

and regarding the graph space:

$$\hat{\phi}_{i0}^A = \sum_{j\in\mathcal{N}_i} \phi_{j0}^A \sum_{kl} \psi_{ik}\psi_{jl} \log(\theta_i\theta_j\eta_{kl}) - \theta_i(1-\theta_i\phi_{i0}^A \sum_{kl} \psi_{ik}\psi_{jl}\eta_{kl})$$
$$+ \sum_{j\in\mathcal{N}_i} \phi_{j1}^A \log(\widetilde{\theta}_j\eta_{bg}) - \eta_{bg}(\widetilde{\theta}^B - \phi_{i1}^A\widetilde{\theta}_i) - \frac{1}{2}\theta_i^2 \sum_k \psi_{ik}\eta_{kk}$$
$$\hat{\phi}_{i1}^A = \log(\widetilde{\theta}_i\eta_{bg}) \sum_{j\in\mathcal{N}_i} \phi_{j0}^A - \eta_{bg}\widetilde{\theta}_i(g - \phi_{i0}^A)$$
$$+ \sum_{j\in\mathcal{N}_i} \phi_{j1}^A \log(\widetilde{\theta}_i\widetilde{\theta}_j\eta_{bb}) - \widetilde{\theta}_i\eta_{bb}(\widetilde{\theta}^B - \phi_{i1}^A\widetilde{\theta}_i) - \frac{1}{2}\widetilde{\theta}_i^2\eta_{bb}$$

where we defined $g = \sum_i \phi_{i0}^A$ and $\widetilde{\theta}^B = \sum_i \phi_{i1}^A\widetilde{\theta}_i$. The *crucial observation* is that the terms g and $\widetilde{\theta}^B$ can be maintained incrementally, i.e. after updating the parameters of node i both terms can be recomputed in constant time.

Overall, for each node i the updates of ψ_i and ϕ_i can be computed in linear time w.r.t. the number of its neighbors $\mathcal{N}_i$. Thus, updating *all* variables (the full E-step) can be done in **linear time w.r.t. the number edges** – and also linear in the number of dimensions.

M-Step: Update of Remaining Parameters. We first simplify $\mathcal{L}_\mathcal{A}$ by introducing some abbreviations: $d_i^G = \sum_{j\in\mathcal{N}_i} \phi_{j0}^A$, $d_i = |\mathcal{N}_i|$, $m_{bg} = \sum_{i,j} \phi_{i1}^A\phi_{j0}^A A_{ij}$, $m_{bb} = \sum_{i,j} \phi_{i1}^A\phi_{j1}^A A_{ij}$, $m_{kl} = \sum_{i\neq j} A_{ij}\psi_{ik}\psi_{il}\phi_{i0}^A\phi_{j0}^A$. We also define the degree related quantities: $D_k^G = \sum_i \theta_i\psi_{ik}\phi_{i0}^A, \forall k$ as the total degree of good nodes in cluster k and $D^B = \sum_i \widetilde{\theta}_i\phi_{i1}^A$ as total degree of bad nodes. Observe that all these terms can be computed in linear time w.r.t. the number of edges or nodes. Furthermore, as also noted in (Karrer and Newman 2011;

Yan et al. 2014), since the likelihood stays the same if we increase all $\{\theta_i \mid z_i = k\}$ by some factor, given that we also decrease $\eta_{kl}, \forall l$ by the same factor, we need constraints to ensure identifiability. Conveniently we pick $D_k^G \stackrel{!}{=} \sum_i d_i^G \psi_{ik} \phi_{i0}^A$ as constraints; and similar w.r.t. $\widetilde{\theta}$: $D^B \stackrel{!}{=} \sum_i d_i \phi_{i1}^A$.

Combining all aspects and after simplification we obtain:

$$\mathcal{L}_{\mathcal{A}} = \frac{1}{2}\Big(\sum_{k,l} m_{kl} \log \eta_{kl} - D_k^G D_l^G \eta_{kl} + m_{bb} \log \eta_{bb} + D^B D^B \eta_{bb}\Big)$$
$$+ \frac{1}{2} \sum_i \sum_{k,l} \psi_{ik}\psi_{il}\theta_i^2 \phi_{i0}^A (\phi_{i0}^A \eta_{kl} - \eta_{kk}) + m_{bg} \log \eta_{bg} - g D^B \eta_{bg}$$
$$+ \sum_i \phi_{i0}^A \log \theta_i d_i^G + \phi_{i1}^A \log \widetilde{\theta}_i d_i + \sum_i \widetilde{\theta}_i \phi_{i1}^A (1 - \phi_{i1}^A)(\eta_{bg} - \frac{1}{2}\eta_{bb})$$

We can further simplify this equation based on the following observations: If we have a rather clear decision whether a node is a graph corruption or not, i.e. $\phi_{i1}^A \to 0$ or $\phi_{i1}^A \to 1$, the term $\sum_i \widetilde{\theta}_i \phi_{i1}^A (1-\phi_{i1}^A)(\eta_{bg} - \frac{1}{2}\eta_{bb})$ evaluates to zero. Similarly, for clear clustering assignment, when $\psi_{ik} \to 1$ for a single k, the term $\frac{1}{2}\sum_i \sum_{k,l} \psi_{ik}\psi_{il}\theta_i^2 \phi_{i0}^A (\phi_{i0}^A \eta_{kl} - \eta_{kk})$ becomes zero. This is indeed what we observed for real data. Besides, while most terms in $\mathcal{L}_{\mathcal{A}}$ grow quadratically with N (e.g. $D_k^G D_l^G$), these terms grow only linearly. Thus, removing them only introduces an error of at most $\frac{1}{N}$. Therefore, for large graphs we can safely drop both terms, since the error they introduce approaches zero in the limit case. We provide further justification in the supplementary material. Overall, we get:

$$\mathcal{L}_{\mathcal{A}} = \Big[\frac{1}{2}\sum_{k,l} m_{kl} log \eta_{kl} - \frac{1}{2} D_k^G D_l^G \eta_{kl} +$$
$$\sum_i \phi_{i0}^A d_i^G \log \theta_i + \phi_{i1}^A d_i \log \widetilde{\theta}_i + m_{bg} \log \eta_{bg} +$$
$$\frac{1}{2} m_{bb} \log \eta_{bb} - \frac{1}{2} D^B D^B \eta_{bb} - g D^B \eta_{bg}\Big] \cdot \Big(1 + \mathcal{O}(\tfrac{1}{N})\Big)$$

Using this in the ELBO – and taking the identifiability constraints via Lagrange multipliers into account – the MAP/ML estimates can now be computed by setting the gradient to zero. We obtain

$$\theta_i = d_i^G \,, \widetilde{\theta}_i = d_i \,, t_{dk} = \frac{\sum_i r_{ik} X_{id}}{R_k}$$
$$\eta_{kl} = \frac{m_{kl}}{D_k^G D_l^G} \,, \eta_{bg} = \frac{m_{bg}}{D^B{}_g} \,, \eta_{bb} = \frac{m_{bb}}{D^B D^B} \tag{6}$$

Where we have defined $r_{ik} = \phi_{i0}^X \psi_{ik}$ as expected responsibilities and $R_k = \sum_i r_{ik}$ as expected fraction of ones in the cluster k. The MAP estimates are

$$\pi_k = \frac{\sum_i (1-\phi_{i3})\psi_{ik} + \alpha_k}{\sum_i (1-\phi_{i3}) + \sum_k \alpha_k} \text{ and } \rho_m = \frac{\sum_i \phi_{im} + \beta_m}{N + \sum_m \beta_m} \tag{7}$$

Using these closed form estimates the full M-step is **linear in the number of edges** as well.

Experiments

There are no competing methods that can handle partial anomalies. Thus, we compare with CODA, FocusCO, PICS, BAGC, LSBM, and SIAN. To evaluate the clustering quality we use normalized mutual information (NMI). To ensure a fair evaluation of the non-robust techniques, we exclude the generated and detected complete corruptions from the NMI calculation. That is, the non-robust techniques are not penalized when they add the corruptions to specific clusters – being a big advantage. At the same time, to make sure that robust techniques do not simply mark all instances as corruptions, we evaluate the detection of anomalies based on the F_1 score. Both measures need to be high at the same time. For all methods we provide the true number K of clusters to detect, the non-deterministic methods were restarted multiple times, and we tuned the parameters required for CODA and FocusCO. For the competing methods we picked the solution achieving highest NMI, while for our approach, we simply perform several restarts with different initializations and pick the one that gives us the highest *likelihood*. The different initializations include multiple random cluster assignments, as well as cluster assignments obtained from a baseline DCSBM. Due to this set-up CODA, FocusCO, LSBM, and SIAN get a *strong benefit*. PICS and BAGC are deterministic. We additionally include a *constrained* version of PAICAN where we disable the detection of anomalies, called PAICAN C. Note that we do not compare with classical DCSBM since PAICAN C is essentially a pure attributed DCSBM (without any anomalies), which is a strictly stronger baseline. More details on the experimental set-up including all used datasets and the PAICAN source code are available in the supp. material.[4]

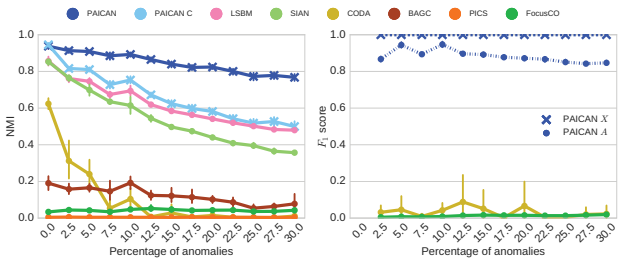
Experiments on synthetic data

To ensure a fair evaluation on synthetic data, we do *not* simply generate graphs according to our probabilistic model. Instead, we used the configuration model (Bollobás 1980): Given a desired degree sequence $\boldsymbol{\theta}$ that follows a power-law distribution $p(x) \propto x^{-\alpha}$ and density ratio $\frac{E_{in}}{E}$, where E_{in} is the number of edges within the clusters, the adjacency matrix $\mathbf{A}$ for the good nodes is generated according to the configuration model conditioned on randomly generated cluster assignments $\mathbf{z}$. We also generate anomalous nodes that form edges at random. The attribute matrix $\mathbf{X}$ for the good nodes is randomly generated given topic probabilities drawn from $Beta(0.1, 5)$. The attributes for the anomalous nodes are generated given an uninformative prior. Unless otherwise noted we generate 5K nodes, 100 attributes and 5 clusters. For each setting of the parameters we generate 10 different random synthetic datasets and report the mean and standard deviation of the relevant metric (i.e. NMI, F_1 score).

Robustness and anomaly detection. First we evaluate the robustness of the different methods. We are interested in answering the following three questions: (i) how is the clustering quality affected as we increase the percentage of anomalous nodes in the data; (ii) how many anomalous nodes we can actually detect; (iii) what is the effect of partial vs. complete anomalies.

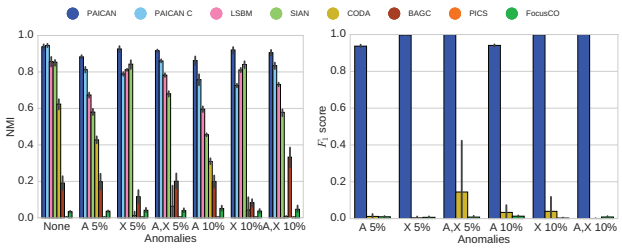
To answer the first two questions, we vary the percentage of anomalies p_a from 0% to 30%, where we distributed the anomalies randomly such that $0.45\ p_a$ are partial anomalies w.r.t. graph space, $0.45\ p_a$ w.r.t. to the attribute space, and the remaining $0.10\ p_a$ are generated as complete anomalies. The results are shown in Fig. 3. As we can see our method is robust and is able to maintain a high clustering quality despite the presence of anomalies. If we disable anomaly de-

[4] http://www.kdd.in.tum.de/PAICAN/



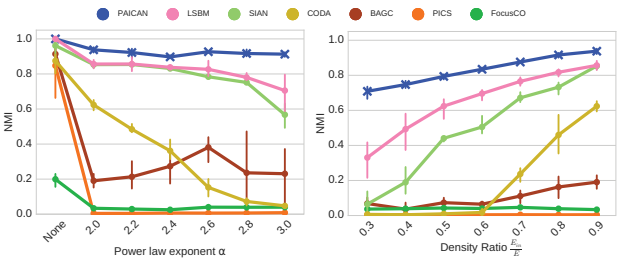
(a) NMI (b) F_1 score

Figure 3: Clustering and anomaly detection performance on synthetic data. PAICAN performs best.



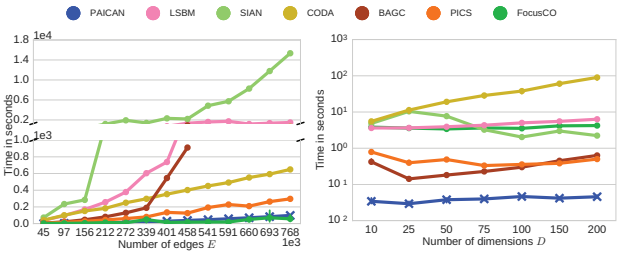
(a) NMI (b) F_1 score

Figure 4: Clustering and anomaly detection performance on synthetic data. PAICAN performs best.



(a) Degree distribution (b) Density ratio

Figure 5: Effect of degree distribution and density ratio on clustering quality. PAICAN clearly performs best.



(a) edges (b) attributes

Figure 6: Runtime vs. number of edges and attributes. PAICAN scales linearly.

tection (PAICAN C), the quality drop is more evident. Similarly for LSBM and SIAN we can see a clear decrease of the performance as the percentage of anomalies increases. Considering Fig. 3b, we can answer the second question: Here, we plot the F_1 score w.r.t. the ground-truth anomalies. Since PAICAN is able to distinguish between graph and attribute corruptions, we can even analyze its performance in detail (e.g. PAICAN A indicates the F_1 score regarding the graph corruptions). We observe that PAICAN is slightly better at detecting attribute corruptions, though, in any case clearly outperforms the competitors.[5] Finally, to answer the third question, we analyze in Figs. 4a and 4b how the methods behave when nodes are partially anomalous. As before, we examine the NMI and F_1 score for 0%, 5% and 10% anomalies – here generating either only graph anomalies (A), attribute anomalies (X), or complete anomalies (A, X). Again PAICAN performs consistently and significantly better.

Degree distribution and density ratio. Despite the fact that most real-world networks have power-law like degree distributions, many (attributed) graph clustering methods are not equipped to properly handle such scenarios. To illustrate this effect we generate data where we vary the power-law exponent to values often encountered in real-world networks ($2.0 \leq \alpha \leq 3.0$) (Chakrabarti and Faloutsos 2006). We also include the simple case of uniform 'blocky' clusters, i.e. all degrees are the same. Fig. 5a shows the results. Our method clearly outperforms all competitors and is not sensitive to the degree distribution and furthermore demonstrates high stability as shown by the low standard deviation across different runs.

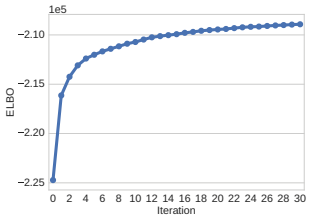
We also explore how the methods behave w.r.t the density ratio $\frac{E_{in}}{E}$. We see on Fig. 5b that most methods start failing as soon as ratio of intra-cluster edges becomes too small, with PAICAN being able to handle the disassortative case the best.

Runtime complexity. The complexity of our method is linear in the number of edges and dimensions. Fig. 6 confirms this result. BAGC and LSBM do not scale linear w.r.t. the number of edges, while CODA does not perform well when increasing the number of attributes (note the log scale on Fig. 6b). SIAN has the worst scaling out of all the methods even though performs relatively well w.r.t. NMI. All of the methods except CODA are not affected by the number of attributes.

Experiments on real-world data

Dataset description. We used six attributed graph datasets, available in the supp. material, along with detailed description. **CORA** is a well-known citation network (N=2708, E=5429, D=1434) The **LAZEGA LAWYERS** dataset contains different networks among attorneys with some categorical attributes for each of them. We use the friendship network and binarize the attributes to obtain N=71, E=575,

[5] Note that PAICAN is the only method able to handle data with both power-law distributed degrees *and* anomalies (see also Fig. 5). Therefore, although FocusCO and CODA can detect anomalies in principle, they fail to detect most of them in the power-law distributed case. They are relatively better for the less common case of 'blocky' clusters, however PAICAN still outperforms them (see Fig. 1 in the supp. material).



(a) Convergence

Dataset	Runtime
Lawyers	0.01s
HVR	0.01s
Parliament	0.01s
Cora	0.15s
DBLP	4.46s
Soc. Papers	25.56s

(b) Runtime

Figure 7: Convergence and runtime on real-world data. PAICAN converges in small number of iterations.

D=70. **HVR** (N=307, E=6526, D=6) is a dataset consisting of several networks of highly recombinant malaria parasite genes. Similar to (Newman and Clauset 2016) we analyze the HVR 6 subnetwork and use the Cys-PoLV (CP) labels as ground-truth clusters. **DBLP** (N=40k, E=418k, D=28) is a co-authorship network of computer science researchers, with the attributes signifying conferences at which authors have published. We created an **AMAZON** co-purchase attributed graph where the attributes are binary product category indicators. We form the dataset from a random subset of products and select the largest connected component (N=29k, E=850k, D=4643). We introduce the **PARLIAMENT** dataset where nodes are French parliament members having an edge if they cosigned a bill together, while their attributes indicate their constituency (N=451, E=11646, D=108). Here we consider political parties as ground-truth clusters. We also create a new **SOCIALPAPERS** dataset where nodes represent biomedical papers forming edges if they are frequently mentioned by the same users on social media (N=20k, E=2mio, D=96). The attributes designate the paper's subjects (e.g. psychology, neurology), and journals are considered as ground-truth communities. The data was collected using the Altmetric API (Adie and Roe 2013).

Ground-truth evaluation. The table below shows the NMI achieved by PAICAN and the competing methods on datasets with ground-truth labels. As we can see PAICAN consistently outperforms the competitors. The non-robust LSBM performs relatively well for most but not all datasets. CODA shows promising results for some datasets, but suffers from scaling issues.

	CODA	FocusCO	BAGC	PICS	LSBM	SIAN	PAICAN
Lawyers	0.50	0.28	0.14	0.27	0.50	0.58	**0.66**
Parliament	0.06	0.00	0.53	0.47	0.77	0.73	**0.78**
Cora	d.n.f.	0.13	0.15	0.04	0.52	0.39	**0.53**
Social Papers	d.n.f.	0.25	0.17	0.10	0.50	d.n.f.	**0.52**
HVR	0.71	0.50	0.18	0.44	0.83	0.77	**0.89**

Table 1: Comparison of NMI for real-world datasets.

Convergence and runtime. We examine the convergence of our algorithm by studying the value of the ELBO per iteration. Fig. 7a shows the evolution of the ELBO for the CORA dataset per iteration. PAICAN quickly converges after a few iterations showing the effectiveness of our variational method. Overall, PAICAN easily handles large graphs as the runtime statistics (seconds per iteration) on the real-world data in Table 7b show.

Figure 8: Topics of the clusters in Amazon data.

Case study: Anomaly detection. As a case study for partial anomalies we analyzed the DBLP dataset. Overall, PAICAN found 37 partial attribute corruptions, 12 partial graph corruptions, and 71 complete corruptions. Since we have no anomaly ground truth we manually analyzed the detected partial anomalous nodes. As an example, the author Srinivasan Parthasarathy has been marked as anomalous in attribute space. When inspecting his ego-network he fits nicely in graph space since most of his neighbors belong to the same cluster. Inspecting his attributes however, we observed that most of his co-authors published in just a few conferences (mainly KDD, ICDM, SDM) while he published in 18 different ones (including e.g. EDBT, IJCAI). This justifies his marking as a partial attribute anomaly. We provide a plot of the ego-network, as well as further case studies on other datasets in the supp. material.

Case study: Clustering. To enable visual inspection of the clustering, we select a small subset ($N = 1549, E = 36934, D = 661$) of the AMAZON dataset. The results for $K = 15$ are visualized in Fig. 8. The learned topic distribution $\mathbf{t}$ is shown, where for easier visualization we only plot dimensions where $t_{dk} > 0.5$ for at least one cluster. Intuitively, this plot shows the 'active' categories for each cluster. For example the products in cluster $C2$ have the following most active categories [Wii U, Nintendo 3DS, PlayStation 3, Xbox 360] clearly showing a coherent cluster of prod-

ucts related to gaming consoles. Similarly, inspecting the topics of $C10$ shows products about jewelery and $C14$ cell phone cases related products. This case study clearly demonstrates that PAICAN learns meaningful clusters.

Conclusion

We proposed PAICAN, a probabilistic model for attributed graph clustering. PAICAN jointly learns the clustering structure as well as potential anomalies. In particular, exploiting the two views of information in attributed graphs, PAICAN introduces the notion of partial anomalies. For learning, we proposed a scalable variational EM algorithm, whose runtime complexity is linear in the number of edges and attributes. Our experimental study confirmed the robustness of PAICAN regarding partial and complete corruptions – state-of-the-art competitors are consistently outperformed. As future work, we aim to investigate methods for automatically selecting the number of clusters using, e.g., principles of nonparametric Bayesian modeling, as well as extensions to numerical attributes.

Acknowledgments

This research was supported by the German Research Foundation, Emmy Noether grant GU 1409/2-1, and by the Technical University of Munich - Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement no 291763, co-funded by the European Union.

References

Adie, E., and Roe, W. 2013. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing* 26(1):11–17.

Akoglu, L.; Tong, H.; Meeder, B.; and Faloutsos, C. 2012. PICS: parameter-free identification of cohesive subgroups in large attributed graphs. In *SIAM SDM*, 439–450.

Akoglu, L.; Tong, H.; and Koutra, D. 2015. Graph based anomaly detection and description: a survey. *DMKD* 29(3):626–688.

Bishop, C. M. 2006. Pattern recognition. *Machine Learning* 128.

Bojchevski, A.; Matkovic, Y.; and Günnemann, S. 2017. Robust spectral clustering for noisy data: Modeling sparse corruptions improves latent embeddings. In *ACM SIGKDD*, 737–746.

Bollobás, B. 1980. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *EJC* 1(4):311–316.

Bothorel, C.; Gomez, J. D. C.; Matteo, M.; and Micenkova, B. 2015. Clustering attributed graphs: models, measures and methods. *Network Science* 3(03):408–444.

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis? *Journal of the ACM* 58(3):11.

Chakrabarti, D., and Faloutsos, C. 2006. Graph mining: Laws, generators, and algorithms. *ACM computing surveys (CSUR)* 38(1):2.

Gao, J.; Liang, F.; Fan, W.; Wang, C.; Sun, Y.; and Han, J. 2010. On community outliers and their efficient detection in information networks. In *ACM SIGKDD*, 813–822.

Günnemann, S.; Färber, I.; Raubach, S.; and Seidl, T. 2013. Spectral subspace clustering for graphs with feature vectors. In *IEEE ICDM*, 231–240.

Günnemann, S.; Boden, B.; and Seidl, T. 2012. Finding density-based subspace clusters in graphs with feature vectors. *DMKD* 25(2):243–269.

Günnemann, S.; Günnemann, N.; and Faloutsos, C. 2014. Detecting anomalies in dynamic rating data: A robust probabilistic model for rating evolution. In *ACM SIGKDD*, 841–850.

Hric, D.; Peixoto, T. P.; and Fortunato, S. 2016. Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X* 6(3):031038.

Huang, H.; Yoo, S.; Qin, H.; and Yu, D. 2011. A robust clustering algorithm based on aggregated heat kernel mapping. In *IEEE ICDM*, 270–279.

Iwata, T., and Yamada, M. 2016. Multi-view anomaly detection via robust probabilistic latent variable models. In *NIPS*, 1136–1144.

Karrer, B., and Newman, M. E. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E* 83(1):016107.

Newman, M. E., and Clauset, A. 2016. Structure and inference in annotated networks. *Nature Communications* 7.

Perozzi, B.; Akoglu, L.; Sánchez, P. I.; and Müller, E. 2014. Focused clustering and outlier detection in large attributed graphs. In *ACM SIGKDD*, 1346–1355.

Rousseeuw, P. J., and Leroy, A. M. 2005. *Robust regression and outlier detection*, volume 589. John Wiley.

Wright, J.; Ganesh, A.; Rao, S. R.; Peng, Y.; and Ma, Y. 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2080–2088.

Xiong, L.; Chen, X.; and Schneider, J. 2011. Direct robust matrix factorizatoin for anomaly detection. In *IEEE ICDM*, 844–853. IEEE.

Xu, Z.; Ke, Y.; Wang, Y.; Cheng, H.; and Cheng, J. 2012. A model-based approach to attributed graph clustering. In *ACM SIGMOD*, 505–516.

Xu, Z.; Ke, Y.; Wang, Y.; Cheng, H.; and Cheng, J. 2014. GBAGC: A general bayesian framework for attributed graph clustering. *TKDD* 9(1):5:1–5:43.

Yan, X.; Shalizi, C.; Jensen, J. E.; Krzakala, F.; Moore, C.; Zdeborová, L.; Zhang, P.; and Zhu, Y. 2014. Model selection for degree-corrected block models. *Journal of Statistical Mechanics* 2014(5):P05007.

Yang, J.; McAuley, J. J.; and Leskovec, J. 2013. Community detection in networks with node attributes. In *IEEE ICDM*, 1151–1156.