GLAD: Group Anomaly Detection in Social Media Analysis

Rose Yu
Department of Computer Science
University of Southern California
qiyu@usc.edu

Xinran He
Department of Computer Science
University of Southern California
xinranhe@usc.edu

Yan Liu
Department of Computer Science
University of Southern California
yanliu.cs@usc.edu

ABSTRACT

Traditional anomaly detection on social media mostly focuses on individual point anomalies while anomalous phenomena usually occur in groups. Therefore it is valuable to study the collective behavior of individuals and detect group anomalies. Existing group anomaly detection approaches rely on the assumption that the groups are known, which can hardly be true in real world social media applications. In this paper, we take a generative approach by proposing a hierarchical Bayes model: *Group Latent Anomaly Detection* (GLAD) model. GLAD takes both pair-wise and point-wise data as input, automatically infers the groups and detects group anomalies simultaneously. To account for the dynamic properties of the social media data, we further generalize GLAD to its dynamic extension d-GLAD. We conduct extensive experiments to evaluate our models on both synthetic and real world datasets. The empirical results demonstrate that our approach is effective and robust in discovering latent groups and detecting group anomalies.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

Keywords

anomaly detection; social media analysis; hierarchical Bayes modeling

1. INTRODUCTION

Social media provide convenient platforms for people to share, communicate, and collaborate. While people enjoy the openness and convenience of social media, many malicious behaviors, such as bullying, terrorist attack planning, and fraud information dissemination, can happen. Therefore, it is extremely important that we can detect these abnormal activities as *accurately* and *early* as possible to prevent disasters and attacks.

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623719 .

By definition, anomaly detection aims to find "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [8]. Several algorithms have been developed specifically for social media anomaly detection such as power-law models [2], spectral decomposition [18], scan statistics [15], and random walk [14, 17]. However, these algorithms only detect the individual *point anomaly*. For example, [2] proposes an "OddBall" algorithm to spot anomalous nodes in a graph. The algorithm extracts features from the egonet of the node and declares anomaly node whose features deviate from the power-law pattern.

In reality, anomaly may not only appear as an individual point, but also as a group. For instance, a group of people collude to create false product reviews or threat campaign in social media platforms; in large organizations, malfunctioning teams or insider groups closely coordinate with each other to achieve a malicious goal. Those appear as examples for another type of anomaly: *group anomaly*, which has not been thoroughly examined in social media analysis. In this work, we focus on group anomaly detection. We are interested in finding the groups which exhibit a pattern that does not conform to the majority of other groups. This problem has found its applications in galaxy identification [19], high energy particle physics [13], anomalous image detection and turbulence vorticity modeling [20].

We identify three major challenges in group anomaly detection: (i) Two forms of data coexist in social media: one is the point-wise data, which characterize the features of an individual person. The other is pair-wise relational data, which describe the properties of social ties. In social science, a fundamental axiom of social media analysis is the concept that structure matters. For example, teams with the same composition of member skills can perform very differently depending on the patterns of relationships among the members [5]. Therefore, it is important to take into account both point-wise and pair-wise data during anomaly detection. (ii) Group anomaly is usually more subtle than individual anomaly. At the individual level, the activities might appear to be normal [21]. Therefore, existing anomaly detection algorithms usually fail when the anomaly is related to a group rather than individuals. (iii) Empirical studies in social media analysis suggest the dynamic nature of individual network positions [12]. People's activities and communications change constantly over time and we can hardly know the groups beforehand. Thus developing a method that can be easily generalized to dynamic setting is critical to anomaly detection in evolving social media data.

In this paper, we take a graphical model approach to address those challenges. We propose a hierarchical model, i.e, *Group Latent Anomaly Detection* (GLAD) model, to connect two forms of data. To handle the dynamic characteristics of the social media data, we further develop a dynamic extension of GLAD: the d-GLAD model. We show that GLAD outperforms existing approaches in terms of group anomaly detection accuracy and robustness. When dealing with dynamic social networks, the dynamic extension of GLAD achieves lower false positive rate and better data fitting. The major contributions of this paper can be summarized as follows:

1. We formulate the problem of group anomaly detection in the context of social media analysis for both static and dynamic settings and articulate the three major challenges associated with the task.

2. We develop a graphical model called GLAD. GLAD can successfully discover the group structure of social media and detect group anomalies. We also generalize GLAD to its dynamic extension and provide tractable model inference algorithms.

3. We conduct thorough experiments on both synthetic and real world datasets using anomaly injections. We also construct a meaningful dataset from ACM publication dataset for rigorous evaluation. The dataset is accessible at `http://www-bcf.usc.edu/~liu32/data.html`.

2. RELATED WORK

We review the related models on group anomaly detection and illustrate the motivation behind our approach.

The Multinomial Genre Model (MGM) proposed in [19] first investigates the problem following the paradigm of Latent Dirichlet Allocation (LDA) [4]. As a text processing tool, LDA assumes that each word is associated with a topic and a document is a mixture of topics. Similarly, MGM models a group as a mixture of Gaussian distributed topics with certain mixture rate and assumes there exists "best" mixture rates, corresponding to the mixture rates of normal groups. Then it conducts group anomaly detection by scoring the mixture rate likelihood of each group. One drawback of MGM is that the set of candidate mixture rates is shared globally by groups. It might leads to poor performance when groups have different sets of mixture rates. [20] further extends MGM to Flexible Genre Model (FGM) with more flexibility in the generation of topics. Specifically, the model considers the set of topic mixture rates as random variables rather than model hyper-parameters, which would adapt to diverse "genres" in groups, each of which is a typical distribution of topic mixture rates.

Another line of work takes a discriminative approach. [13] uses the same definition of group anomaly from [19]. It considers kernel embedding of the probabilistic distributions and generalizes one-class support vector machine from point anomaly detection to group anomaly detection. The proposed support measure machine (SMM) algorithm maps the distributions to a probability measure space with kernel methods, which can handle the aggregate behavior of data points.

However, existing approaches separate the group anomaly detection task into two stages: group discovery and anomaly detection. They require the group information to be given before applying the anomaly detection algorithms. For example, in [19], the Sloan Digital Sky Survey (SDSS) dataset needs to be pre-processed before feeding into MGM. The authors first construct a neighborhood graph and then treat the connected components in the graph as groups. For the application on turbulence data, the FGM model [20] considers the vertices in a local cubic region as a group. In SMM [13], the authors treat the high energy particles generated from the same collision event as a group.

The two-stage approaches identify the groups from the pair-wise data and infer the anomalies based on the point-wise data. This strategy assumes that the point-wise and pair-wise data are marginally independent. However, such independence assumption might underestimate the mutual influence between the group structure and the feature attributes. The detected group anomalies can hardly reveal the joint effect of these two forms of data. These motivate us to build an alla prima that can account for both forms of data and accomplish the tasks of group discovery and anomaly detection all at once.

Additionally, existing work can only deal with static network and fixed size groups. This is not feasible for the time-evolving nature of social media data. For example, in corporate networks, employees may switch teams from one to the other. The organization structure of a team may also change. As the dynamic setting needs to take into account the flexible group size and the changing mixture rates, we further adapt our model to the dynamic setting and formulate the problem as a change point detection task.

Group anomaly detection in social media analysis may shed light on a wide range of real world problems such as corporate restructuring, team job-hopping and political inclination shift to which our approach can apply. In section 3, we provide a formal definition of group anomaly in social media analysis. With the definition, we develop the GLAD model in section 4 and present its learning and inference algorithm. In section 5, we describe the dynamic extension of the GLAD model: the d-GLAD model, which can handle the dynamic social networks. Section 6 shows the empirical evaluation results of GLAD and d-GLAD on synthetic and real world datasets compared with existing baseline models.

3. DEFINITION OF GROUP ANOMALY

The core of our group anomaly definition lies in the collective behavior of individuals. For example, a document is a mixture of various topics and a team is a mixture of different roles. Therefore, we model the node features of each group as a mixture of components. Each component could be an article topic, a social role or a job title. Specifically, we can describe a component as either a discrete variable such as multinomial distribution or a continues variable like Gaussian distribution, depending on the data type of features. Here we use the term *role* as a general notion for the component. We assume that there are a fixed number of roles and each of which denotes a particular distribution of node features. All groups share the same set of roles but possibly with different role mixture rates. Normal groups follow the same pattern with respect to their role mixture rates, but the anomalous group has a role mixture rate that deviates from the normal pattern.

For the static GLAD model, we are interested in the distribution of the role mixture rates across the groups. According to our assumption, the mixture rates of normal groups are

more likely to appear. For groups with very rare role mixture rates, we treat them as group anomalies. One example of this type of group anomaly comes from particle physics. It is widely accepted that the dynamics of known particles are governed by the *Standard Model*, which corresponds to the normal pattern. Unknown particles would contaminate the distribution of the Standard Model. Detecting those anomalies could potentially lead to the discovery of new physical phenomenon. In practice, we first identify the normal mixture rates. Then for each learned group, we evaluate the likelihood of its observations being generated with the normal mixture rates. The lower the likelihood value is, the more anomalous the group would be.

For the dynamic d-GLAD model, we emphasize on the temporal aspect of the data and detect the change of the role mixture rate within the groups. For instance, in scientific area, it is valuable to study the evolution of research topics and detect the bursty time periods. In the dynamic setting, since the structure of groups change as well as their role mixture rates, detecting groups with rare mixture rate no long applies. Therefore, we think of the task as a change point detection problem and aim to detect the groups whose mixture rates change drastically from the previous time stamps. Compared with GLAD, we not only need to decide whether a group is anomaly or not, but also need to specify when the group appears anomalous.

Even though we use slightly different definitions of group anomaly for the GLAD model and the d-GLAD model, the key ideas behind our definitions are the same. Both definitions build upon the notion of role mixture rate, which essentially requires a precise inference of both the group membership and role identity for each individual in the group.

4. THE GLAD MODEL

Suppose that we are given a social network with N nodes. We observe node activities $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ and their communications $\mathbf{Y} = \{Y_{1,1}, Y_{1,2}, \ldots, Y_{N,N}\}$. $X_p \in \mathbb{R}^V$ consists of V entries, denoting a feature vector of V dimensions. $Y_{p,q} \in \{0, 1\}$ is a binary valued variable, indicating the pairwise relationship of nodes. These two forms of data are our inputs. Our goal is to analyze these data jointly and declare the group that has irregular role mixture rate as anomaly. In the following sections, we first describe the motivation for our hierarchical Bayes model and provide its generative process and the plate notation. Then we derive the inference algorithm using the variational Bayesian approach.

4.1 Model Specification

We model a social network with N individuals. Assuming that each person p is associated with a group identity G_p and a role identity R_p. By groups, we mean the clusters that capture the similarity suggested by the pair-wise communications. By roles, we refer to the mixture components that categorize the point-wise feature values of the nodes. For simplification, we fix the number of groups as M and the number of roles as K. Figure 1 shows the plate notation for the GLAD model. The motivation for us to assume two identities for an individual comes from the controversial viewpoints of what is the right metric for a community. In community detection literature [7], some argue that a community is the one that has dense communications within clusters while others suggest that people in the same community should share common activity features. We get

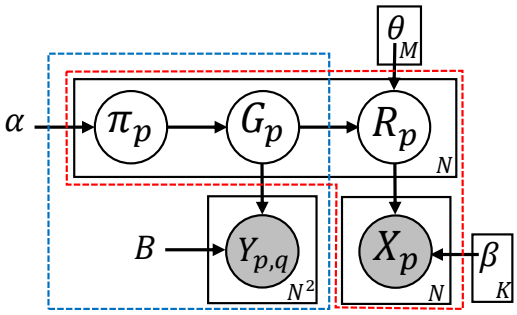


Figure 1: Plate notation for the Group Latent Anomaly Detection (GLAD) model. Shaded circles are observations, blank circles are latent variables and the variables without a circle are model parameters. The blue rectangular resembles MMSB. The red polygon integrates the generating process of LDA.

around the controversy by recognizing the arguments of both sides and assume two types of latent structures coexist in the data.

For each person p, he joins a group according to the membership probability distribution π_p. We impose a Dirichlet prior on the membership distribution. It is well known that the Dirichlet distribution is conjugate to the multinomial distribution. As we will show later, when dealing with latent variables, the Dirichlet prior facilitates the learning and inference of the model. We assume the pair-wise link $Y_{p,q}$ between person p and person q depends on the group identities of both p and q with the parameter B. Furthermore, we model the dependency between the group and the role using a multinomial distribution parameterized by a set of role mixture rate $\{\theta_{1:M}\}$. The role mixture rate characterizes the constitution of the group: the proportion of the population that plays the same role in the group. Finally, we model the activity feature vector of the individual X_p as the dependent variable of his role with parameter set $\{\beta_{1:K}\}$. Table 1 summarizes the notations used in our model.

We specify the generative process of the GLAD model in Algorithm 1. Our model unifies the ideas from both the Mixture Membership Stochastic Block (MMSB) model [1] and the Latent Dirichlet Allocation (LDA) model [4]. As shown in Figure 1, the blue dashed rectangular on the left side resembles MMSB which models the formation of groups using link information. The red dashed polygon integrates the generating process of LDA which is often used for topic extraction from documents. By assuming mixture of groups and roles, we allow each person to have multiple roles and multiple group memberships. Without loss of generality, we assume that the activity data have discrete values and choose to model X_p with a multinomial distribution. When the activity data are distributed in other forms, we can easily adapt GLAD to model other type of X_p.

4.2 Inference and Learning

Inference requires us to compute the posterior distributions of the latent variables given the data. The normalizing term of the posterior distribution involves the calculation of the marginal likelihood of the data for which we resort to variational EM algorithms [10].

Table 1: Notations of GLAD

Symbol	Description
α	Dirichlet prior parameter
π_p	membership distribution of person p
$Y_{p,q}$	pair-wise communication between p and q
B	global block probability among groups
G_p	group of p
R_p	role of p
$\theta_{1:M}$	role mixture rate for M groups
X_p	activity of p
$\beta_{1:K}$	activity distribution for K roles

Algorithm 1 Generative Process of GLAD

for individual $p = 1 \to N$ **do**
  Draw membership distribution $\pi_p \sim \text{Dir}(\alpha)$
  Draw $G_p \sim \text{Multinomial}(\pi_p)$
  for individual $q = 1 \to N$ **do**
    Sample $Y_{p,q} \sim \text{Bernoulli}\ (G_p^T B G_q)$
  end for
  Draw $R_p \sim \text{Multinomial}(R_p|\theta_{1:M}, G_p)$
  Draw $X_p \sim \text{Multinomial}(X_p|\beta_{1:K}, R_p)$
end for

Denote the set of model parameters as $\Theta = \{\alpha, B, \theta_{1:M}, \beta_{1:K}\}$, the set of observed variables as $v = \{X_{1:N}, Y_{1:N}\}$, and the set of the hidden variables as $h = \{\pi_{1;N}, G_{1:N}, R_{1:N}\}$. Our aim is to estimate the posterior distribution $p(h, \Theta|v)$. We can first write out the complete joint likelihood of observed and latent variables as follows:

$$\begin{aligned} p(v, h|\Theta) &= \prod_p p(\pi_p|\alpha) \times \prod_{p,q} p(Y_{p,q}|G_p, G_q, B) \\ &\times \prod_p p(X_p|R_p, \beta_{1:K}) p(R_p|G_p, \theta_{1:M}) p(G_p|\pi_p). \end{aligned}$$

Computing the maximizer for the marginal likelihood of the data $p(v|\Theta) = \int_h p(v, h|\Theta) dh$ requires the integration over all the latent variables in the equation above, which is intractable [1]. Therefore, we apply the variational Bayesian approach [10] to perform the inference approximately. The essence of the variational Bayesian approach is to choose a variational distribution $q(h)$ to approximate the actual posterior distribution, so that the Kullback-Leibler divergence (KL-divergence) between $p(h, \Theta|v)$ and its approximation $q(h)$ is minimized.

Rewriting the marginal log likelihood and plugging in the variational distribution, we have

$$\log p(v|\Theta) = D_{KL}(p||q) + \langle \log p(v, h|\Theta) \rangle_q - \langle \log q(h) \rangle_q,$$

where we use $\langle f \rangle_p$ to represent the expectation of the function f with respect to the distribution p. Since the marginal likelihood $\log p(v|\Theta)$ is invariant to the choice of q, minimizing the KL-divergence $D_{KL}(p||q)$ is equivalent to maximizing the last two terms $\langle \log p(v, h|\Theta) \rangle_q - \langle \log q(h) \rangle_q$. In practice, we choose $q(h)$ to be factorized over the latent variables with free parameters $\Delta = \{\gamma_{1:N}, \mu_{1:N}, \lambda_{1:N}\}$ as follows:

$$q(h|\Delta) = \prod_p q(\pi_p|\gamma_p) q(R_p|\mu_p) q(G_p|\lambda_p).$$

Our goal is to find the optimal set of free parameters that provides a variational distribution closest to the actual posterior. Then our problem is to maximize the objective function formulated as follows subject to probability constraints:

$$\begin{aligned} \Delta^\star &= \arg\max_\Delta \langle \log p(v, h|\Theta) \rangle_q - \langle \log q(h|\Delta) \rangle_q \\ &= \arg\max_\Delta L(v, h, \Theta, \Delta). \end{aligned}$$

The objective function L, by plugging in the joint likelihood and the variational distribution and taking expectations, is given by

$$\begin{aligned} & L(v, h, \Theta, \Delta) \\ = & \sum_p \langle \log p(X_p|R_p, \beta_{1:K}) \rangle_q \\ + & \sum_p \langle \log p(R_p|G_p, \theta_{1:M}) \rangle_q + \sum_p \langle \log p(G_p|\pi_p) \rangle_q \\ + & \sum_{p,q} \langle \log p(Y_{p,q}|G_p, G_q, B) \rangle_q + \sum_p \langle \log p(\pi_p|\alpha) \rangle_q \\ - & \sum_p \langle \log q(\pi_p|\gamma_p) \rangle_q \\ - & \sum_p \langle \log q(R_p|\mu_p) \rangle_q - \sum_p \langle \log q(G_p|\lambda_p) \rangle_q. \end{aligned}$$

We follow a variational EM procedure in order to maximize $L(v, h, \Theta, \Delta)$ over Δ. Basically we iteratively update the free parameters by taking the derivative of the Lagrange function of the objective L over one parameter at a time given the value of others from the last iteration. Since $\{Y_{p,q}\}$ is symmetric, the objective function will result in a quadratic term with respect to λ_p. Taking the derivative over the variational parameter would not have a closed form solution. A simple workaround is by assuming constant probability for the generation of $\{Y_{p,p}\}$. We omit the tedious derivations and only present the final update formulas of each of the free parameters, as shown in Algorithm 2. For convenience, we denote $f(Y_{p,q}, B_{m,n}) = Y_{p,q} \log B_{m,n} + (1 - Y_{p,q}) \log(1 - B_{m,n})$.

Algorithm 2 Variational Inference for GLAD

initialize $\gamma_{p,m} := 1/M$
initialize $\mu_{p,k} := 1/K$
initialize $\lambda_{p,m} := 1/M$
repeat
  for $p = 1 \to N$, $m = 1 \to M$ $k = 1 \to K$ **do**
    $\gamma_{p,m} = \alpha_m + \lambda_{p,m}$
    $\lambda_{p,m} = \exp\{\sum_k \log \theta_{m,k} \mu_{p,k} + \psi(\gamma_{p,m}) - \psi(\sum_n \gamma_{p,n}) + \sum_{q \neq p} \sum_n \lambda_{q,n} \cdot f(Y_{p,q}, B_{m,n})\}$
    $\mu_{p,k} = \exp\{\sum_v \log \beta_{v,k} X_{p,v} + \sum_m \log \theta_{m,k} \lambda_{p,m}\}$
  end for
until convergence

For the parameter estimation, we apply the empirical Bayes method on the variational likelihood. We maximize the Lagrange function of $L(v, h, \Theta, \Delta)$ over model parameters $\Theta = \{\alpha, B, \theta_{1:M}, \beta_{1:K}\}$. Due to the compounding of Dirichlet distribution, there is no close form solution for the maximizer w.r.t α. We apply the Newton-Raphson method to reach a numerical solution. The resulting parameter updating functions for α and B are the same as those of MMSB [1] and the parameters β and θ can be estimated as follows:

$$\beta_{v,k} = \frac{\sum_p X_{p,v}\mu_{p,k}}{\sum_{v,p} X_{p,v}\mu_{p,k}} \quad \theta_{m,k} = \frac{\sum_p \mu_{p,k}\lambda_{p,m}}{\sum_{k,p} \mu_{p,k}\lambda_{p,m}}.$$

With our definition of group anomaly in section 3, we score the group anomalousness using $-\sum_{p\in G}\langle \log p(R_p|\Theta)\rangle_p$. The most anomalous group will have the highest anomaly score. We approximate the true log likelihood with the variational log likelihood to get $-\sum_{p\in G}\langle \log p(R_p|\Theta)\rangle_q$.

A limitation of GLAD is that it only models the static network. This might be restrictive if we want to further consider dynamic networks. Besides the anomaly group whose mixture rate deviates significantly from other groups, we are also interested to study how the mixture rate evolves over time. Fortunately, GLAD can be easily extended to account for this dynamics. This leads to the dynamic extension of the GLAD model, which will be discussed in the next section.

5. DYNAMIC EXTENSION OF GLAD

We now generalize the GLAD model to take into account the dynamics in the social media. We refer the dynamic extension of GLAD as the d-GLAD model. To be consistent with our description for GLAD in section 4, we start with the model specification and then provide the model inference algorithm using both the variational Bayesian method and the Monte Carlo sampling technique.

5.1 Model Specification

Generalization of GLAD to d-GLAD stems from the template models [23], which use the model for a particular time stamp as a template, duplicate it over time and connect temporal components sequentially. Similarly, we can adapt GLAD to the dynamic setting by making a copy of GLAD for each time point. To simplify the model, we assume that the latent factors including role R_p, group G_p and mixture rate $\{\theta_{1:M}\}$ change over time but the membership distribution $\{\pi_p\}$ and model parameters are fixed.

We model the temporal evolution of the role mixture rate for each group with a series of multivariate Gaussian distributions. At a particular time point, the Gaussian has its mean as the value of the mixture rate. And the mixture rate of the next time point is a normalized sample from this Gaussian distribution. Since we require the mixture rate to be the parameters of a multivariate distribution over features, we apply a soft-max function to normalize the sample drawn from the multivariate Gaussian. The soft-max function is defined as $S(\theta_m) = \frac{\exp\theta_m}{\sum_m \exp\theta_m}$. When the total time length T equals one, d-GLAD reduces to the GLAD model. Figure 2 depicts the probabilistic graphical model of d-GLAD and the meanings of notations used. We summarize the generative process of d-GLAD in Algorithm 3.

In d-GLAD model, since the mixture rate of next time stamp is drawn from a multivariate Gaussian centering around the mixture rate of its previous time stamp, it imposes smoothness on the mixture rates across time, preventing the mixture rate from having drastic changes. The soft-max function maps the samples from the multivariate Gaussian to the parameters for the multinomial distribution. Similar idea can be seen from the generalization of LDA to the dynamic topic model [3]. While it is true that d-GLAD model shares the constraints of GLAD on fixed group/role number and constant self-loop, it has certain intriguing advantages over static models. (i) d-GLAD captures the dynamics of the latent variables G_p and R_p, thus allows an individual to switch groups and roles over time.(ii) The smoothness of the mixture rate over time models the behavior of normal groups, so detecting groups whose mixture rates θ_m^t undergo substantial change becomes easier.

Algorithm 3 Generative Process of DGLAD

for $t = 1 \to T$ **do**
  for $m = 1 \to M$ **do**
    Draw $\theta_m^{(t)} \sim \text{Gaussian}(\theta_m^{(t-1)}, \sigma)$
  end for
  for individual $p = 1 \to N$ **do**
    Draw membership distribution $\pi_p^{(t)} \sim \text{Dir}(\alpha)$
    Draw $G_p^{(t)} \sim \text{Multinomial}(\pi_p)^{(t)}$
    for individual $q = 1 \to p-1$ and $q = p+1 \to N$ **do**
      Sample $Y_{(p,q)}^{(t)} \sim \text{Binomial}\ ((G_p^{(t)})^T B G_q^{(t)})$
    end for
    Draw $R_p^{(t)} \sim \text{Multinomial}(R_p^{(t)}|\mathcal{S}(\theta^{(t)}_{G_p^{(t)}}))$
    Draw $X_p^{(t)} \sim \text{Multinomial}(X_p^{(t)}|\beta_{R_p^{(t)}})$
  end for
end for

5.2 Inference and Learning

The variational inference of d-GLAD is similar to the GLAD model except for the longitudinal factor $\theta_{1:M}^{(1:T)}$. We add a variational distribution $p(\theta_m^{1:T}|\hat{\theta}^{1:T})$ to approximate the original posterior where $\{\hat{\theta}^{1:T}\}$ are variational parameters. Then we apply the variational Kalman Filter technique [3] to infer the sequential latent variables and learn the model parameters. The transition for the mixture rate of each group is Gaussian distributed:

$$\theta^{(t)}|\theta^{(t-1)} \sim \mathcal{N}(\theta^{(t-1)}, \sigma^2 I).$$

We can write the variational distribution for the transition as follows:

$$\hat{\theta}^{(t)}|\theta^{(t-1)} \sim \mathcal{N}(\theta^{(t-1)}, \hat{v}^2 I).$$

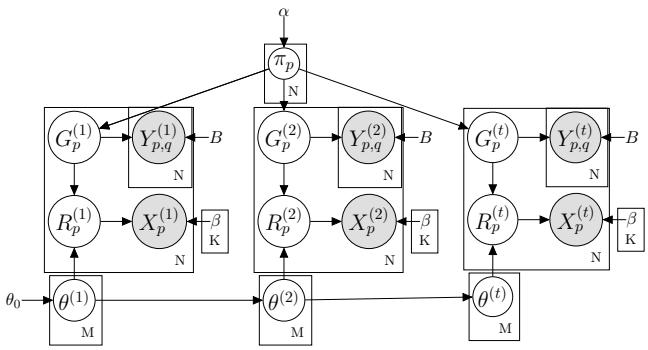
Then we can apply similar variational EM procedure incorporating the transitions to infer the variational parameters. Due to the numerical difficulty of variational Kalman filter method, we also implement a version of the Monte Carlo sampling for d-GLAD model, which is used in our empirical evaluations. The algorithm is elaborated in Algorithm 4. The inference of the transitional part $\{\theta^{1:T}\}$ is based on the Particle Filtering method [22]. The anomaly score of the d-GLAD model is measured by $\|\theta_m^{(t)} - \theta_m^{(t-1)}\|$.

6. EXPERIMENTS

To evaluate the effectiveness of our model, we conduct thorough experiments on synthetic datasets and real world datasets. We study the applications of our approach by analyzing scientific publications and senator voting records.

6.1 Baselines

To our knowledge, all existing algorithms are two-stages approaches: (i) identify groups, (ii) detect group anomalies.



Notation	Description
α	Dirichlet parameter
π_p	membership distribution
$G_p^{(t)}$	group of p at time t
$Y_{p,q}^{(t)}$	pair-wise communication at time t
B	global block probability
$R_p^{(t)}$	role of p at time t
$\beta_{1:K}$	activity mixture rate
$X_p^{(t)}$	activity of p at time t
θ_0	initial Gaussian mean
$\theta_{1:M}^{(t)}$	role mixture rates at time t

Figure 2: Plate notation for the d-GLAD model and the meaning of notations. The subscript p denotes each person in the social network. The superscript t denotes the network snapshot at time stamp t.

Algorithm 4 Monte Carlo Sampling of DGLAD

Initialize $\alpha, \theta_0, \beta_{1:K}, B$
$R_{1:N}^{1:T} = 1/K$, $G_{1:N}^{1:T} = 1/M$, $\pi_{1:N} \sim \text{Dir}(\alpha)$
repeat
  for $p = 1 \to N$ **do**
    for $t = 1 \to T$ **do**
      Update $R_p^{(t)} \sim \text{Mul}(\mathcal{S}(\theta^{(t-1)}_{G_p^{(t-1)}}))\text{Mul}(X_p^{(t)})$
      Update $G_p^{(t)} \sim \text{Mul}(\pi_p^{(t-1)})\text{Mul}(\mathcal{S}(\theta^{(t-1)}_{G_p^{(t-1)}}))$
    end for
    Update $\pi_p \sim \text{Dir}(\alpha)$
  end for
  for $t = 1 \to T$ **do**
    Update $\theta^{(t)}$ using Particle Filtering
  end for
until Convergence

We summarize these algorithms in Table 2. We use following approaches as baseline methods in comparison to GLAD and d-GLAD:

1. **MMSB-LDA**:First use the MMSB model to learn a group membership distribution for each individual node, then assign the node to the group with the highest probability. Finally, for each group, train an LDA model and infer the role identity.

2. **MMSB-MGM**: Group is learned using the same method as MMSB-LDA. For the role inference, train an multi-modal MGM instead of LDA.

3. **Graph-LDA**: Run an off-the-shelf graph clustering algorithm Min-Cut to get group membership and then train a LDA model for each group.

4. **Graph-MGM**: Get group membership with the graph clustering algorithm Min-Cut and then train a MGM model for each group.

6.2 Synthetic Dataset

We experiment on two type of synthetic datasets. One is a synthetic dataset with injected group anomalies. The other is a benchmark dataset generated by a simulator with individual anomaly labels.

Table 2: Two stage models in existing work

Algorithm	Stage-1	Stage-2
Heard 2010 [9]	spectrum	Poisson process
Xiong 2011-a [19]	clustering	Mixture Genre Model
Xiong 2011-b [20]	clustering	Flexible Genre Model
Muandet 2013 [13]	simulator	One class SMM

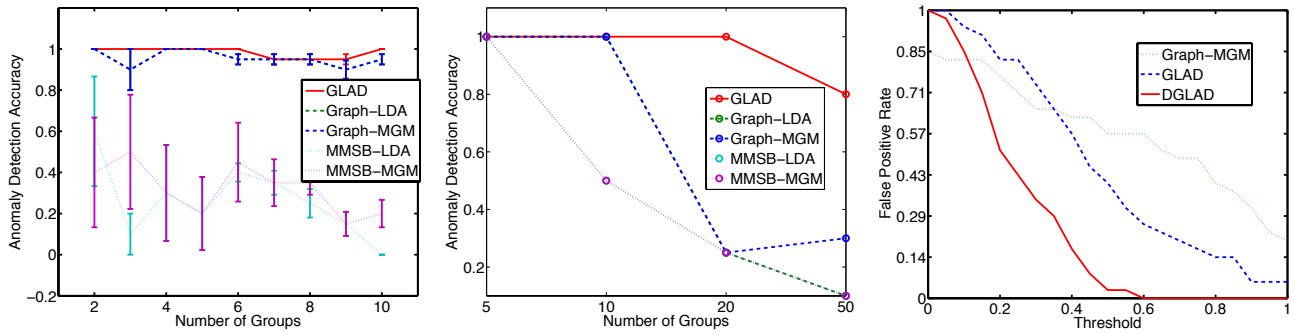
6.2.1 *Synthetic Data with Anomaly Injection*

We generate a network with 500 nodes using GLAD in Algorithm 1. To evaluate the anomaly detection performance, we set the mixture rates of anomalous groups as $[0.9, 0.1]$ and normal groups as $[0.1, 0.9]$. We vary the number of groups from 5 to 50 and inject 20% anomalous groups. The rest 80% groups are normal. Since we know the normal and anomalous mixture rates, we calculate the anomaly score of each group by directly computing the differences between the inferred mixture rate and the ground truth normal mixture rate. During the testing procedure, we rank the groups with respect to their anomaly score and retrieve top 20% groups. For all methods, we set the number of groups and number of roles the same as the ground truth.

We compare the learned groups of three grouping approaches with the ground truth: GLAD, MMSB and Graph, for the case of 5 groups. The inferred group memberships are shown as adjacent matrices in Figure 3. For better visualization, we intentionally put the nodes that belong to the same group together. Ideally, we should observe dense links within groups and sparse links between groups. Therefore, the dark pixels in the plot would aggregate along the principal diagonal of the matrix. We use blue color to highlight the groups learned. The group discovery result of GLAD is the closest to the ground truth. The high connectivity in the graph and the lack of point-wise information could be the reasons for the poor performance of Graph and MMSB.

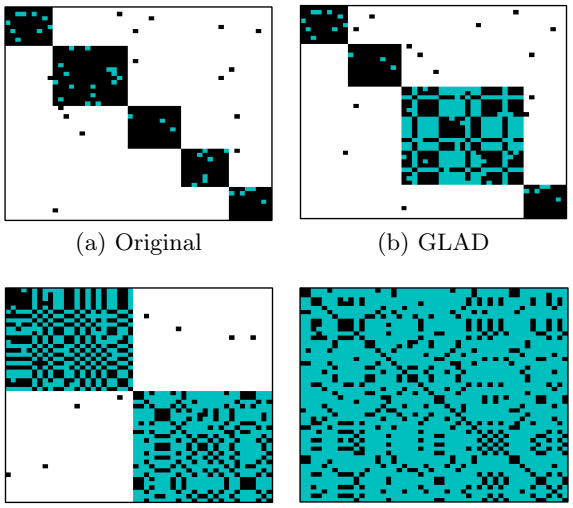
Figure 4 shows the anomaly detection performance with different number of groups for GLAD and four other baselines. GLAD achieves the highest detection accuracy. It is also more robust over 10 random runs. Note that the differences for the first stage of baselines are more obvious than the second stage. This is because the Bernoulli distribution limits the number of samples in the pair-wise data, making the first stage more difficult to learn.

We also report the simulation results on group anomaly detection for d-GLAD. The data is generated according to



(a) Robustness (b) Accuracy (c) d-GLAD

Figure 4: Anomaly detection performance of GLAD and baseline methods on synthetic dataset of 500 samples, with 20% anomalous groups. (a): Detection accuracy error bar plot of five algorithms for detection accuracy with group number from 2 to 10. (b): Mean detection accuracy for different number of groups up to 50, averaged over 10 random runs. (c):False positive rate over different thresholds for d-GLAD, MMSB-MGM and GLAD for synthetic data. 10% group anomalies are injected.



(a) Original (b) GLAD

(c) MMSB (d) Graph

Figure 3: The 50×50 adjacent matrix re-arranged by the group membership discovered by three grouping approaches on a subset of synthetic data of 5 groups. Dark pixels denote links and white pixels denote no links. Blue block highlights the learned group membership.

Algorithm 3 with 5 time stamps. We manipulate the mixture rate of 50% of the groups at time point 4 as injected anomalies. Then we raise alarms if the group's mixture rate deviates from the previous time by a certain threshold. In Figure 4(c), we display the false positive rate with different threshold values. For comparison, we train MMSB-MGM and GLAD at each time independently as baselines. It can be seen that d-GLAD achieves the lowest false positive rate, which demonstrates the gain of d-GLAD over static models on the dynamic dataset.

6.2.2 *Benchmark Data with Anomaly Labels*

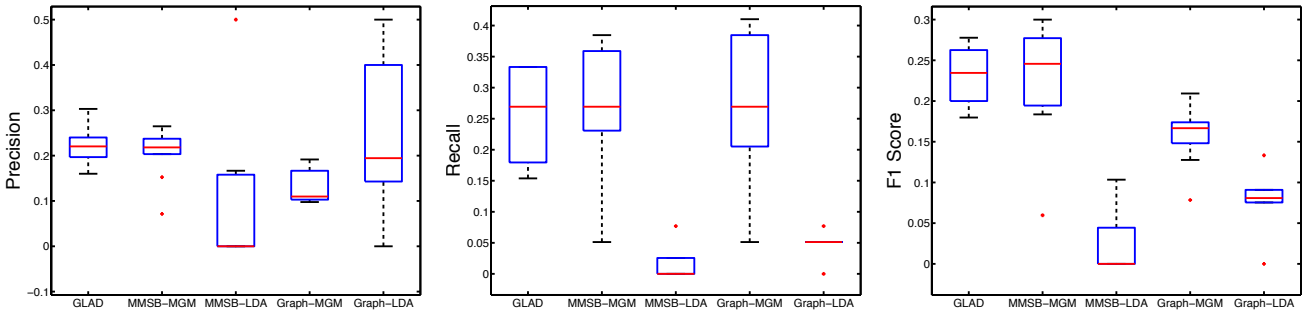
The benchmark data set is generated by a simulator from a federal funded program. It contains email communication records and working activities from 258 company employees. Each employee is featured by 6 types of activities. The labeled dataset contains 39 individual anomalies and 5 of them cannot be detected by any existing algorithms. We set the number of groups as 20 as the optimal setting obtained from cross validation and calculate the anomaly score of each group by MCMC sampling. We treat all members in the most anomalous group as individual anomalies and compare them with the anomaly labels. Though the anomaly labels are point anomalies rather than group anomalies, the anomaly detection result reflects the potential of our approach to tackle other type of difficult anomaly detection problems. The precision, recall and F1 score over 20 runs on the benchmark dataset is shown in Figure 5.

We can see that the GLAD model achieves comparable precision and recall with low variances. In contrast, the detection performances of the two-stage models fluctuates significantly. In terms of the F1 scores in Figure 5(c), both GLAD and MMSB-MGM beat the other algorithms while GLAD has a lower variance than MMSB-MGM. One possible explanation is that the point-wise features prevent the size of the group to become either too large or too small, thus leading to more robust performance.

6.3 Real World Datasets

6.3.1 *Scientific Publications*

Researchers study the topics of papers seeking for concise representations of scientific publications, which contain both pair-wise data like co-authorship and point-wise data such as bag of words features. Detecting anomalous topic distributions in scientific publications can sharpen our understanding of the structure of research communities and possibly reveal unusual research trends. In order to quantify our method, we resort to anomaly injection and construct a dataset with group anomaly labels. One way to construct group anomalies is the scenario that a conference paper corpus is contaminated by group of papers from conferences in other domains.



(a) Precision (b) Recall (c) F1 Score

Figure 5: (a): Precision (b): Recall (c): F1 score on the benchmark dataset of GLAD and four baseline methods over 20 runs. All members in the anomalous groups are treated as individual anomalies and compared with 39 true anomalies.

Table 3: Group Anomaly Accuracy of GLAD and four baselines on DBLP publications. With KDD papers treated as normal groups and other conferences are treated as group anomalies respectively.

Methods	GLAD	Graph-LDA	Graph-MGM	MMSB-LDA	MMSB-MGM
DBLP:KDD/CVPR	**0.4167**	0.3333	0.3333	0.2500	0.2500
DBLP:KDD/ICML	**0.2500**	0.0833	0.0833	0.1667	0.1667
DBLP:KDD/SIGMOD	**0.2875**	0.0750	0.0500	0.1625	0.1625
DBLP:KDD/CIKM	**0.4500**	0.4000	0.3625	0.2625	0.2625
DBLP:KDD/EDBT	**0.2625**	0.0500	0.0875	0.2000	0.2000

We create a dataset from a pre-processed Digital Bibliography and Library Project (DBLP) dataset from [6]. The dataset consists of conference papers from 20 conferences of four major area: database (DB), data mining (DM), information retrieval (IR) and artificial intelligence (AI). Each paper has a bag-of words feature vector with a vocabulary size of 11,771 and associated 28,702 authors information. The detailed statistics of the dataset are shown in the top half of Table 4. We set up the group anomaly detection scenario as follows: we randomly sample groups of papers from KDD and treat them as normal groups. Then we sample groups of papers from the other conferences (e.g, CVPR, ICML , SIGMOD) and inject them into KDD papers as group anomalies. If the two papers have at least one common author, we add a link between them.

Accordingly, all conferences share four topics. But different conferences might have difference point of emphasis, resulting in different mixture rates of topics. Our goal is to pick out the "anomalous" papers from the corpus. We sample 50 groups of papers and inject 20% group anomalies. We apply different models with 50 groups and 4 roles to the data for inference of the membership and role distributions. Then we rank 50 groups with respect to their anomaly scores. We treat the top 20% groups as the detected anomalies. Table 3 shows the anomaly detection accuracy by GLAD and four other baselines. GLAD is superior to all four baselines models for different combination of normal/abnormal settings. We also display the topics learned by the GLAD model. In Table 5, we show the top ten most representative words for the four topics, which well reproduce the topic results reported in [6].

Since the DBLP dataset does not contain time-specific information which is not suitable for the d-GLAD model, we process another ACM dataset downloaded from ArnetMiner

Table 4: Key statistics of the DBLP and ACM publication datasets

DBLP			
# of docs	28,569	# of authors	28,702
# of conf	20	# of words	11,771
# of links	104,962	# of area	4
ACM			
# of docs	31,574	# of authors	4,474
# of year	10	# of words	8,024

[16]. The dataset contains the publications from year 2000 to 2009 by 4,474 authors, mainly from the data mining community. In order to study the topic evolution for academic scholars, we extract the abstracts of all publications and group them by authors and publishing years. For each author, we construct a bag of words feature vector out of all the papers he/she has written in one year. And the communication networks we generate are based on the co-authorship of the papers. Whenever two authors have collaborations in a certain year, we create a link between them for the network snapshot in that year.

Due to the lack of labels, it is difficult to directly evaluate our model on anomaly detection task. As an alternative, we design a prediction task to compare the modeling performance of GLAD and d-GLAD on ACM publications. Specifically, we separate the papers into training and testing sets and measure the predictive model log-likelihood on the testing data. For d-GLAD, we train our model using a series of publications from previous years, and test on the year immediately after. For the GLAD model, as it is a static model, time independence assumption applies. We train the model using previous year and test on the next year. The model fitting results are shown in Table 6. Out

Table 6: Prediction negative log likelihood for GLAD and d-GLAD on ACM dataset over 9 years.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
GLAD	28421.63	28023.68	**30184.66**	**32039.92**	28317.67	**30539.66**	26105.21	34340.53	25967.75
DGLAD	**34411.28**	**33411.14**	29935.87	31958.92	**30082.65**	29696.12	**30042.77**	**34395.68**	**31683.49**

Table 5: The most representative words learned by GLAD on DBLP dataset of four topics: database, data mining, information retrieval and artificial intelligence.

DB	DM	IR	AI
databases	data	web	query
object	mining	information	system
access	efficient	learning	management
database	query	search	processing
oriented	algorithm	retrieval	web
security	queries	clustering	efficient
based	clustering	query	performance
indexing	databases	text	infomation
systems	algorithms	model	distributed
privacy	large	classification	optimization

of 9 training-testing experiments, d-GLAD model achieves higher log-likelihood than GLAD model for 6 times, indicating d-GLAD as a better fit for the evolving publication modeling.

6.3.2 US Senate Voting

We collect the voting records from the government website of United States 109th Congress [1] using the New York Time Congress API [2]. The records of 109th Congress contain 100 senators' voting spanning two sessions from Jan 1st 2005 to Dec 31st 2006. We divide the 24 months records into 8 time slots, where each slot denotes a 3-month interval. Then we apply the method of [11] to construct a network from original yay/nay votes. For the nodes features, we collect the statistics of votes in six dimensions, namely House Joint Resolution(hjres), House of Representatives(hr), Presidential Nomination(pn), Simple Resolution(s), sconres(Senate Concurrent Resolution) and Senate Joint Resolution(sjres). We evaluate GLAD on single aggregated network and d-GLAD on the 8 time slots time-varying data.

We set the number of groups as 2 and number of roles as 3 as the Senate consists of two major parties and maintains three types of committees. Figure 6 shows the groups inferred by GLAD. The blue nodes denote Democratic party members and the red ones are Republican. Compared with known facts, the model correctly reveals the party affiliation except for two outliers: Ben Nelson (Democratic) and James Jeffords (Independent). The underlying reason is that the votes of these two senators are often at odds with the leadership of his party, leading to false grouping. We conduct an anecdotal investigation and find that the congressional vote rating from the National Journal placed Ben Nelson to the right of five Senate Republicans in 2006. For James Jeffords, he served as a Republican until 2001, when he left the party to become an Independent and began caucusing with the Democrats.

[1]http://www.senate.gov/
[2]http://developer.nytimes.com/docs/read/congress_api

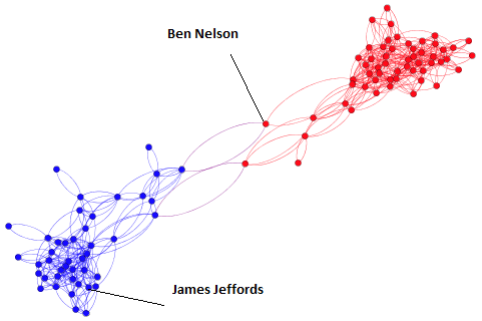


Figure 6: Common votes graph with party labels inferred by GLAD for 100 senators on the aggregated network. Compared with ground truth, two outliers are highlighted due to their anomalous voting behavior.

Since there are merely two groups, it is impetuous to say one party is more anomalous than the other. Instead, we use d-GLAD to detect time points when the role mixture rates change dramatically. In fact, d-GLAD raises an alarm at the 7th time-step for Democratic. A well known political event happened during this time is that Democratic senator Joseph Lieberman lost the Democratic Party primary election and became a independent Democratic in September 2006. Though it may be over-optimistic to draw the conclusion that this event causes the sudden change of role mixture rates, it serves as an evidence that the dynamics of the voting behavior is closely related to the party affiliation of members.

7. CONCLUSION

In this paper, we study the problem of group anomaly detection in social media analysis. We identify the group membership and the role for each individual and define the group anomaly with respect to the *role mixture rate*. We develop a hierarchical Bayes model: the GLAD model, for detecting the group anomaly. The GLAD model utilizes both the pair-wise and point-wise data and automatically infers the group membership and the role at the same time. To further account for the dynamic nature of social media, we generalize GLAD to the d-GLAD model as an extension for handling time series. We derive the variational Bayesian method as well as the Monte Carlo sampling for the model inference.

The superior performances of GLAD and d-GLAD are demonstrated on synthetic datasets and real world social media datasets. For example, GLAD successfully detects "anomalous" papers from the scientific publication corpus with injected anomalies. d-GLAD uncovers the party affiliation changes in the Senate which correspond to political events. The type of group anomaly that we define here has also been investigated in the applications of astronomy [19],

particle physics [13] and fluid mechanics [20]. But our work is the first study in group anomaly detection that aims to identify groups and detect anomalies simultaneously. During this study, we mainly focus on the applications in social media analysis, but we believe that our model is capable of tackling similar problems in other domains.

Our current approach has a number of limitations. First, the lack of labeled data for group anomaly increases the difficulty in evaluation. Here we create a benchmark dataset using anomaly injection. But the injected anomaly cannot fully represent the naturally occurred anomaly in practice. Secondly, we do not fully address the problem of model selection for our approach. For the experiments on senator and publication datasets, we either manually choose the model hyper-parameters based on our prior knowledge or use cross-validation. Finally, given the complexity of the model and the particle filtering approach, the d-GLAD model is computational expensive. Thus it is not yet scalable to data with long time span.

For future directions, we will apply the variational Bayesian method for d-GLAD or parallel sampling procedure to achieve further speed-up for large-scale data. Furthermore, it would be interesting to build non-parametric models to automatically learn the number of groups and roles. Future research should also address better procedures and criteria for evaluating the group anomaly detection methods in real world social media applications.

8. ACKNOWLEDGMENTS

The research was sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) under the Anomaly Detection at Multiple Scales (ADAMS) program, Agreement Number W911NF-11-C-0200 and NSF research grants IIS-1134990. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency, or the U.S. Government.

9. REFERENCES

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(1981-2014):3, 2008.

[2] L. Akoglu, M. McGlohon, and C. Faloutsos. Anomaly detection in large graphs. In *In CMU-CS-09-173 Technical Report*. Citeseer, 2009.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[5] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895, 2009.

[6] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1271–1279. ACM, 2011.

[7] S. Fortunato. Community detection in graphs. June 2009. Physics Reports 486, 75-174 (2010).

[8] D. M. Hawkins. *Identification of outliers*. Chapman and Hall, 1980.

[9] N. A. Heard, D. J. Weston, K. Platanioti, and D. J. Hand. Bayesian anomaly detection methods for social networks. 2010.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[11] M. Kolar, L. Song, A. Ahmed, E. P. Xing, et al. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010.

[12] G. Kossinets. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, Jan. 2006.

[13] K. Muandet and B. SchÃűlkopf. One-class support measure machines for group anomaly detection. Mar. 2013.

[14] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 653–658. ACM, 2004.

[15] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan statistics on enron graphs. *Computational & Mathematical Organization Theory*, 11(3):229–247, 2005.

[16] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.

[17] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. 2006.

[18] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[19] L. Xiong, B. Poczos, J. Schneider, A. Connolly, and J. Vanderplas. Hierarchical probabilistic models for group anomaly detection. pages 789–797, 2011.

[20] L. Xiong, B. Póczos, and J. G. Schneider. Group anomaly detection using flexible genre models. In *NIPS*, pages 1071–1079, 2011.

[21] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[22] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later.

[23] D. Koller and N. Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.