STATISTICAL MODELING OF SOCIAL NETWORKS ACTIVITIES

Mohammed A. Aabed and Ghassan AlRegib

School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, GA, 30332 USA
{maabed,alregib}@gatech.edu

ABSTRACT

This paper introduces a new paradigm to characterize and understand the dynamics of a complex social network where we set up a mathematical platform that captures the network dynamics. We propose a novel generic non-parametric model to characterize a general system of social communicators. We divide the network into low-level entities, each of which has some independent features. The different entities are then combined using Bayesian nonparametric statistics, namely Dirichlet processes mixture models (DPMM). This set up was tested using a simulated case study where we show examples of its utility for behavior characterization and predictions.

Index Terms— Social networks, online communities, communities discovery, non-parametric modeling, Dirichlet processes.

1. INTRODUCTION

Since their introduction in the first half of the past decade, social networks have been an appealing format of communication that rapidly gained a massive popularity among different groups and demographics. The actions and activities by the network users and media exchanges are not for entertainment purposes only, they rather touch on every aspect of their lives, be it educational, financial, personal, emotional, etc. Such diversity in the nature of these entities undoubtedly adds to the complexity of these systems. This diversity in applications, among other issues with respect to this domain, calls for a rigorous measurement of the variations and parameters that describe the behavior of entities in these systems. This diversity also adds to the challenge of characterizing the network in a generic way that can be utilized to benefit any of these aspects.

Several works attempted modeling social communities. These studies can be classified in three categories: (i) statistical and graphical network and community modeling [1], (ii) activity based individual and community behavioral analysis [2], and (iii) network optimization and data delivery [3, 4].

Despite their theoretical and applicable novelty, these studies do not stimulate the influences of the events and transaction taking place among users. Furthermore, the broader integration between the media and network activities is merely considered in these works. In this paper, we try to resolve the dilemma of how to characterize the *connections*, *small transactions* and *exchanged media* in nowadays social networks. Furthermore, our work contemplates describing the patterns of these effective and influential movements, inferences about the community, and technical or social consequences. The main challenge in characterizing these entities is to identify the members, activities, connections and media that may have major or broad impacts.

The rest of this paper is organized as follows. Section 2 presents our proposed network setup and paradigm to characterize the network metrics and variables. Section 3 introduces the mathematical details of a nonparametric adaptive machine learning approach based on Dirichlet Process Mixtures. Section 3.2 is a case study where we demonstrate the utility of the proposed model using a specific setup. In this setup, the responses and connections are modeled as Gaussian and Gamma process, respectively. The simulation results of this case study, analysis and applications are discussed in Section 4. Finally, the concluding remarks are given in Section 5.

2. NETWORK GENESIS AND AXIOMS

The proposed model considers a social online community with members space defined by $\mathcal{U} = \{u_0, u_1, \cdots, u_Q\}$. The network is assumed to have N predefined classes (themes) of media, $\mathcal{M}_n \subset \mathcal{M}\ \forall n \in \mathbb{N}_N$[1]. These classes or themes define the different forms, topics or a combination of both for exchanged media in the network. In this context, elements of the members space are assumed to be connected via N-dimensional bidirectional *links* whose *weights* may independently vary in each direction. Every unidirectional link manifests a media propagation channel from one user to another. The weights independency, therefor, accounts for

[1]The notation $\mathbb{N}_N$ is used herein to denote the subset of natural numbers $\{1, 2, \cdots, N\}$

activities variations and different personal penchants between two connected users. The weight of the n^{th} link from u_i to u_j is denoted by $w_{n(i,j)}$ $\forall n \in \mathbb{N}_N$.

To epitomize the community's stance towards a certain class (theme) of media, we introduce a random variate, $s_{n(i)}$, to record and parameterize the response generated by u_i towards a medium of class $\mathcal{M}_n$.

In our modeled and simulated network, we decree the random variates $w_{n(i,j)}$ and $s_{n(i)}$ to have probability density functions (PDFs) of the exponential family. In Section 3.1, we show how this model is robust for any mixture of densities from this class of probability distributions. *As it will be seen later, this paradigm is generic for any choice of continuous or discrete random measure*. The modeled community in the case study, presented in Section 3.2, builds on $w_{n(i,j)}$ and $s_{n(i)}$ to be drawn from mixtures of Gaussian and Gamma PDFs, respectively. These densities are denoted by $f_{W_n}(x)$ and $f_{S_n}(x)$, respectively.

The PDFs for $w_{n(i,j)}$ and $s_{n(i)}$ characterize their respective random variables (RVs) with respect to the entire community. These densities are also calculated to describe weights and responses of any single user. These individualized densities for u_i are denoted by $f_{W_{n(i)}}(x)$ and $f_{S_{n(i)}}(x)$, respectively.

3. STATISTICAL MODELING

3.1. Infinite Dirichlet Process Mixture Models

In this section, we illustrate our likelihood estimation methodology based on the theory of Dirichlet Processes (DPs) [5]. It should be noted that while this Bayesian nonparametric estimation paradigm is applied to all the introduced probability measures, each function is calculated separately with the proper modifications to the mathematical model.

Let X denote any RV of interest whose distribution is known to follow an infinite mixture model. Consider the following Dirichlet Process Mixture Model (DPMM) as a nonparametric prior in a hierarchical Bayesian setup [5]:

$$G \mid \{\alpha, H\} \sim DP(\alpha, H) \tag{1a}$$

$$\lambda_n \mid G \sim G, n \in \mathbb{N}_N \tag{1b}$$

$$X_n \mid \lambda_n \sim p(x_n|\lambda_n) \tag{1c}$$

where G is a random measure from a DP, H is the base distribution of the DP, α is a scaling parameter, λ_n is an independently drawn RV from G, and X_n is the observable data. In our model, we adopt the stick-breaking representation of the DPs due to Sethuraman [5] which provides a constructive approach to generate a DP.

For the purpose of this work, the observable data is assumed to be samples from an exponential family distribution, in which case the base distribution of the DP is the corresponding conjugate prior.

This modeling approach establishes a clear roadmap between the mixtures (clusters), likelihoods and priors. In addition, it provides a powerful tool to characterize the data in an adaptive and nonparametric fashion maintaining an imperative flexibility on the form of likelihood.

Following this theoretical abstract construction of the clustering problem in a generic social networks, Section 3.2 explains the details of specific network modeling scenario where the responses are drawn from Gaussian masses and the weights are subject to Gamma processes. The results of our simulation and numerical tests are explained in Section 4.

3.2. Case Study: Gaussian-Gamma Community Modeling and Clustering

In this section, we introduce a case study with a specified setup to the ecumenical paradigm introduced in Section 3.1. The modeled case assumes the variates to be of continuous nature for generality and diversity in testing. In particular, the responses, $s_{n(i)} \in \mathbb{R}$, are characterized by a Gaussian Multivaraite Mixture (GMM) with means vector $\boldsymbol{\mu} \in \mathbb{R}^N$ and precision matrix $\boldsymbol{\Sigma} \in \mathbb{R}_+^{N\times N}$, $s_{(i)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The prior of the GMM follows a Wishart-Gaussian model such that $\boldsymbol{\Sigma} \sim W(\mathbf{V}, p)$ with p degrees of freedom and scale matrix $\mathbf{V}$ [6]:

$$f(\boldsymbol{\Sigma}|\mathbf{V}, p) = \frac{|\boldsymbol{\Sigma}|^{(p-N-1)/2} \exp(-\frac{1}{2}\mathrm{Tr}(\mathbf{V}^{-1}\boldsymbol{\Sigma}))}{2^{pN/2}|\mathbf{V}|^{p/2}\Gamma_N(p/2)} \tag{2a}$$

The means of the likelihood is a multivariate Gaussian conditioned on the likelihood of the precision matrix, $\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim N(r\boldsymbol{\Sigma}, \mathbf{m})$, where $\mathbf{m}$ is the mean of $\boldsymbol{\mu}$ and r is a calibration factor of prior measurements on measurement scale:

$$f(\boldsymbol{\mu}|\boldsymbol{\Sigma}) = (2\pi)^{-N/2}|r\boldsymbol{\Sigma}|^{1/2} \exp(-\tfrac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^\top r\boldsymbol{\Sigma}(\boldsymbol{\mu}-\mathbf{m})) \tag{2b}$$

The marginal posterior density of $\boldsymbol{\mu}$ was calculated and the posterior hyperparameters are updated as follows [6]:

$$r' = r + n \tag{2c}$$

$$p' = p + n \tag{2d}$$

$$\mathbf{m}' = \frac{1}{r+n}(r\mathbf{m} + \boldsymbol{\chi}_i^\top \mathbf{u}_n) \tag{2e}$$

$$p\mathbf{V}^{-1\prime} = p\mathbf{V}^{-1} + \Sigma_{i=1}^n \boldsymbol{\chi}_i\boldsymbol{\chi}_i^\top + r\mathbf{m}\mathbf{m}^\top - r'\mathbf{m}'\mathbf{m}'^\top \tag{2f}$$

Furthermore, the weights follow a mixture of Gamma likelihoods, $w_{n(i,j)} \sim \Gamma(\kappa, \theta)$, with shape and scale parameters κ and θ, respectively. In this proposed setup, we select a univariate prior where the inverse of the scale parameter is Gamma distributed conditioned on the scale parameter of the likelihood, $\theta^{-1}|\kappa \sim \Gamma(\alpha, \beta)$ where α and β are the shape and scale hyperparameters, respectively. The shape parameter of the likelihood is calibrated according to the communication patterns between the users. That is, the value of κ is decremented after a certain number of responses from u_j

to a medium shared by u_i. The rate of this communication still, however, follows a Gamma DPMM. Reducing the shape parameter of the Gamma likelihood drives the weights towards a lower value which indicates a stonger association in this constructed setup. This mechanism allows for discovering undeclared associations between the users and potential ties between users through common affiliations. Hence, the hyperparameters update model is as follows [7]:

$$\alpha' = \kappa n + \alpha \tag{3a}$$

$$\beta' = \frac{\beta}{1 + b\boldsymbol{\chi}_i^\top \mathbf{u}_n} \tag{3b}$$

The marginal (compound) distribution of the data is given by [7]:

$$p_{w_{n(i,j)}}(x) = \frac{\Gamma(\kappa+\alpha)}{\Gamma(\kappa)\Gamma(\alpha)} x^{\kappa-1}(x+1)^{-(\kappa+\alpha)} \tag{3c}$$

4. ANALYSIS AND APPLICATIONS

In this section, we introduce the results and analysis of a simulated case study that we call "John's Network", as shown in Fig. 1. In this simulated case, we consider a single media theme for demonstration purposes, $N = 1$, without any loss of generality. The initialization of the shape and scale parameters of the weight RVs are as shown in Fig. 1. The numbers indicate the values of the shape and scale parameters of the likelihood, (κ, θ). It should be noted here that we abstract the actual meaning of these numerical values. In real social systems, the natures and inter-arrival times of the responses may be incorporated to affect the weight values. In this scenario, John generates and shares his media with all the users to whom he is directly connected. Prior to generating any responses, the system does not assume any knowledge about the interests (penchants) of these users. The network is then updated based on the embracement and communication patterns of the users. In Section 4.1, we discuss the procedure and results for processing the responses to identify the associatively of the users, including the implicit ones. In Section 4.2, the responses are used to identify the users interests and characterize the probability functions for this modality.

4.1. Gamma Modeling of Weights

We discuss in this part the users associativity, i.e. linkages, in John's network. In this setup, the weights are generated from a random Gamma process with the corresponding shape value and a random scale value between 0 and 20, $\theta \sim U(0, 20)$. The hyperparameters are initialized to the worst value, which in most of the cases is the unity, $\alpha = 1$ and $\beta = 1$. The value of the shape likelihood for any user is decremented by 2 after the user generated a fixed number of responses (20 in our setup). For instance, after 60 embracements by Pat of John's media, her likelihood shape drops from an initial 11 to 5. This

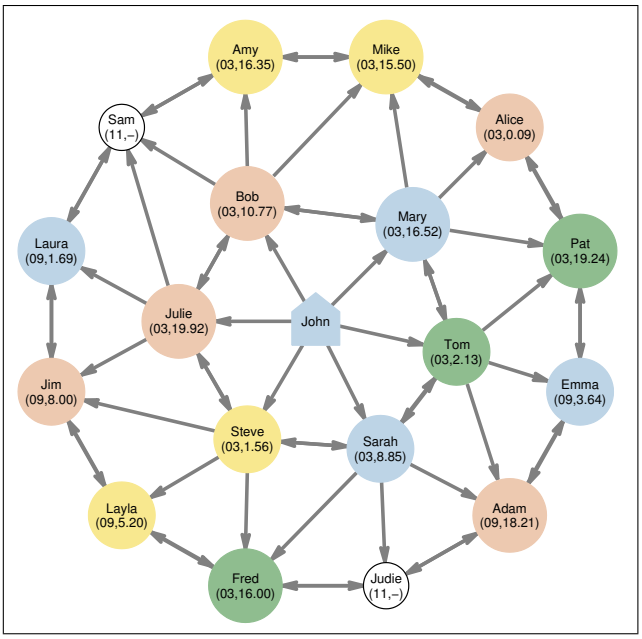

Fig. 1. The clustering of John's social network based on the users' responses to a single media theme. The nodes are color-coded to reflect the interests biases where every color represents one of the Gaussian components in Fig. 2.

systematic mechanism of controlling one parameter, κ, and allowing the other, θ, to be randomly adaptive has two advantages. Firstly, calibrating the shape parameter provides a controlled metric of the communication pattern between John and his network without loosing the generality of model. Secondly, the rates of these communications is estimated based on the samples and history of responses to give an accurate prediction of these patterns. This latter step leads to an accurate characterization of the likelihood. We opted for this approach for its generality and practical sense. The described Gamma conjugate-prior introduced in Section 3.2 assumes a known shape parameter and randomizes the scale parameter. In the context of this setup, the users' weights are generated from Gamma processes with varying shapes. For instance, Alice's weights are generated from 3 Gamma processes with shapes 11, 9 and 7. Estimating the global weights likelihood for John, which includes all the users, is done by clustering the samples based on their users shape parameters. That is, the global Gamma DPMM for John will have a mixture of 5 different Gamma shapes: 11, 9, 7 5 and 3. Despite that the results are shows for specific values of the parameters and hyperparameters, various tests have been conducted for several combinations of the parameters values. In all the cases, we run the algorithm for 100 iterations where an optimal likelihood is reached within 10 Gibbs iterations with the proper initialization of the hyperparameters without any knowledge about the randomly estimated parameters.

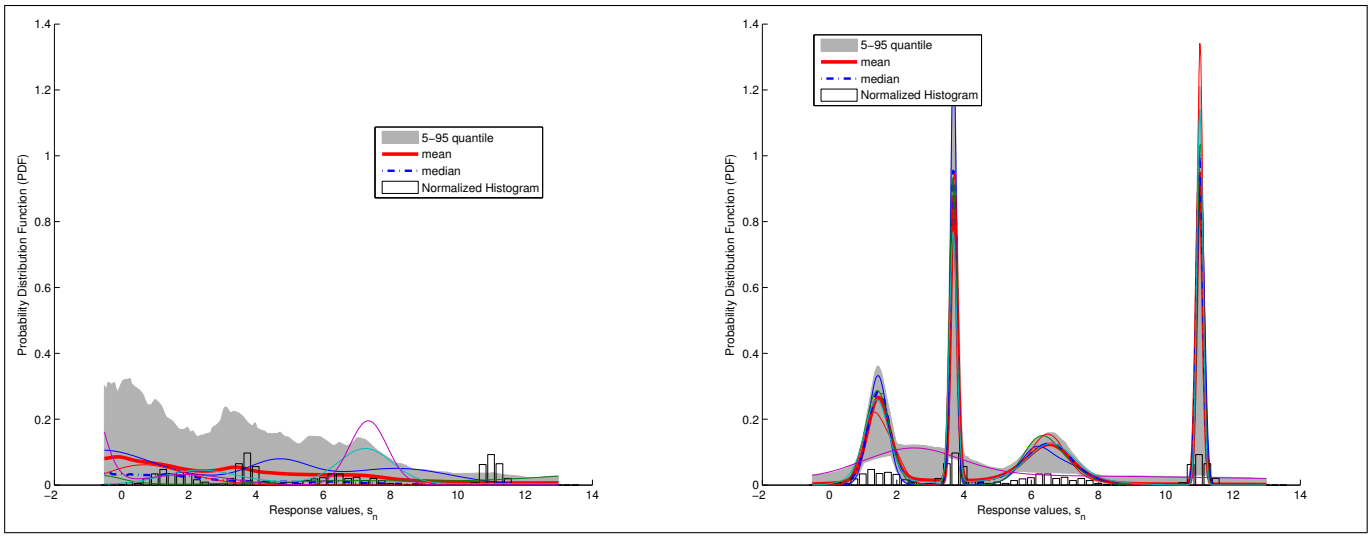

Fig. 2. The clustering of the users' responses after applying Gaussian DPMM: priors (left) and posteriors (right) over probability.

4.2. Gaussian Modeling of Responses

In our view of the network, we use the term "theme" to describe a general class of media, which could be sports, politics, music, etc. For each theme, there are different classification "criteria". For example, in music, these criteria could be different artists, genres, etc. These definitions and assignments are specified by the network operator or social scientists. Fig. 1 shows the clustering of the users' responses based on four major criteria for a single theme. In our simulations, the response samples were generated from four mixtures of random means and standard deviations. The hyperparamters were naturally initialized without any estimation or conjecture about the randomly generated means and variances. Samples of the estimated Gaussian DPMMs for this community is shown in Fig. 2.

5. CONCLUSION

In this paper, we investigated the the problem of understanding the behavior and communication patterns of the users of social networks. We proposed a novel outlook to probabilistically formulate and characterize these responses, transactions, and communication patterns in social networks. The proposed scheme makes use of the generality of Dirichlet Process to satisfy the multimodality and multidimensionality requirements in these systems while maintaining an accurate and flexible Bayesian estimator. This is of great utility to the service operators and network engineers of social networks. Furthermore, we have shown, through an simulated case study, the potential applications and outcomes of the proposed framework.

References

[1] Haizheng Zhang, Ke Ke, Wei Li, and Xuerui Wang, "Graphical models based hierarchical probabilistic community discovery in large-scale social networks," *International Journal of Data Mining, Modelling and Management*, vol. 2, no. 2, pp. 95 – 116, 2010.

[2] Y. Zhou, X. Guan, Q. Zheng, Q. Sun, and J. Zhao, "Group dynamics in discussing incidental topics over online social networks," *IEEE Network*, vol. 24, no. 6, pp. 42 –47, Nov.-Dec. 2010.

[3] W.S. Lin, H.V. Zhao, and K. Liu, "Cooperation stimulation strategies for peer-to-peer wireless live video-sharing social networks," *IEEE Instrumentation and Measurement Magazine*, vol. 19, no. 7, pp. 1768 –1784, July 2010.

[4] J. Park and M. van der Schaar, "A game theoretic analysis of incentives in content production and sharing over peer-to-peer networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 704 –717, Aug. 2010.

[5] J Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[6] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin, *Bayesian Data Analysis*, Chapman and Hall/CRC, 2 edition, July 2003.

[7] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1, Wiley-Interscience, 2 edition, Oct. 1994.