Regularized Topic-aware Latent Influence Propagation in Dynamic Relational Networks

Shuhui Wang, Liang Li
Inst. of Comp. Tech., CAS
Beijing, China
{wangshuhui, liang.li}
@ict.ac.cn

Chenxue Yang
Institute of Automation, CAS
Beijing, China
chenxue.yang@nlpr.ia.ac.cn

Qingming Huang
School of Computer and
Control Engineering, UCAS
Beijing, China
qmhuang@ucas.ac.cn

ABSTRACT

On social networks, investigating how the influence is propagated helps us understand how the network evolves and the social impact of different topics. In previous study, the influence propagation is modeled based on the static network structure, or the infection between two connected users is recovered from some given event cascades. Unfortunately, existing solutions are incapable of identifying the user susceptibility delivered by user generated content. In this paper, we propose REGINFOIBP, a general regularized learning framework for modeling topic-aware influence propagation in dynamic network structures.

Specifically, the observed time-sequential user topic preference and user adjacency information are factorized by the prior information reflected by a user-influential bipartite relation graph.

The influence propagation is approximated with a nonparametric regularized Bayesian matrix factorization model with polynomial complexity, turning this NP-hard problem into a tractable one. and the influential users are identified by several sampling algorithms with slightly different approximation qualities. To further model dynamic temporal evolution, we construct Markov conditional probabilistic model on the compact latent feature representation. By integrating both topic and structure information into the regularized non-parametric probabilistic learning process, REGIBP is more efficient and accurate in discovering the key factors in the content and influential users in dynamic network structure. Extensive experiments demonstrate that REGINFOIBP better adapts to real data, and achieves better approximation in influence propagation over existing approaches.

PVLDB Reference Format:
Ben Trovato, G. K. M. Tobin, Lars Thørväld, Lawrence P. Leipuner, Sean Fogarty, Charles Palmer, John Smith, Julius P. Kumquat, and Ahmet Sacan. A Sample Proceedings of the VLDB Endowment Paper in LaTeX Format. *PVLDB*, 11 (9): xxxx-yyyy, 2018.
DOI: https://doi.org/TBD

1. INTRODUCTION

The emergence of multifarious social networks in the past decades has initialized the study of information propagation among social networks. An interesting question that received considerable attentions is how to model the phenomenon of influence-driven propagation. For example, to analyze the widely recognized "word-of-mouth" effect, *influence maximization* [22], [17], [7] aims to find the top k influential users with the largest number of follow-ups. As a core research problem in applications ranging from viral marketing to epidemiology, information maximization has been extensively studied. In this paper, instead of dealing with the NP-hard influence maximization problem directly, we study how the influence propagation acts on content with rich social attributes in dynamic relational networks. We aim to explain how content generation and dynamically changing social network structure interact with influence propagation. By understanding the evolution of user preferences and network structure, applications in social networks, e.g., personalized content recommendation, can be better facilitated. We address the research challenges first.

User Generated Content. Content delivery in social networks has become the most important procedure of information propagation. The user generated content (UGC), produced and edited by users and communities, describes events that belongs to certain topics while itself is also born with rich social properties. For example, a tweet or a comment characterizes the associated user, and different authors are connected by their collaborations. However, the impact of different UGC on different users and communities has been ignored in most of previous research on influence maximization. For instance, a user who is an author in sports may not have the same influence on users interested in politics. Users are more likely to influence their follow-ups who share the same interest in a specific topic than those having no common interests with them. Therefore, how to integrate user preferences into the information propagation modeling for more accurate influence estimation needs to be investigated.

Dynamic Networks. Social relations in social networks are evolving time to time, making influence propagation modeling on dynamic networks a challenging problem. For instance, the constantly changing following status on Twitter greatly determines which groups of friends will receive the information at the first place when a user tweets or re-tweets messages. Existing models [17], [9], [8], [23], [10], [4], [26], [2] model the influence diffusion on a given exponential-family random graph or a pre-defined diffusion network structure. However, it is still far from the real situation since the dynamic network properties have not been considered. As another possible solution routine, influence maximization based on combinatorial optimization theories on a time-sensitive diffusion network tends to be intractable. The dynamically changing network topology make it more difficult to obtain a stable solution of influential users.

Proceedings of the VLDB Endowment, Vol. 11, No. 9

DOI: https://doi.org/TBD

This paper aims at developing a general framework for latent influence propagation by considering both topic information and dynamic network structure. We collect two kinds of observable temporal data series, the *user preferences* and the *user adjacency*. The former represents how each user is associated to the topics along the time axis, and the latter reflects the affinity structures among users at different temporal observation points. Given the two observations, we are trying to discover the most influential users during influence propagation process, at the same time we seek to recover the observed data and predict the information propagation panoramic states in the future.

To model the influence propagation more tractably and accurately, we model the influence propagation from the perspectives of diffusion in one time step and propagation between consecutive time steps. In a single time step, the key idea for modeling the diffusion process is taking the influential users as latent features of the observed data where the observations can be represented by a weighted combination of latent features. The combinational weights are embodied by a *feature loading matrix*. We propose a Bayesian regularized nonparametric matrix factorization method to model the mixture of latent features, i.e., the influentials. We use *Indian buffet process* (IBP) [15], [21], [11], [16] to generate the prior of the feature loading matrix. The relation between users and influential users can be modeled by a bipartite graph, based on which the latent influential users can be estimated effectively. Likewise, the user adjacency observations can also be approximated by a combination of the latent influential features. The advantage of factorizing the observations is that the information of the whole network can be approximated using the information delivered by a compact group of influentials, which is similar in spirit with the dimensional reduction techniques in relevant study.

Between two contiguous time steps, we use hidden Markov chain to model the dynamic temporal evolution in both content and network topology. The hidden state at each time step is the feature loading matrix, which is also referred to as the *influence propagation matrix*. For two observations, two hidden Markov chains for each are constructed, and the chains share their hidden states encoded by the influence propagation matrix for simultaneous modeling of the content and structure information evolution. The joint evolution modeling can be interpreted by two widely accepted assumptions. First, users who share more preferences have higher probabilities to interact with each other and have more common friends. Second, users are more likely to share more preferences when they interact frequently and have more common friends. For model inference, we use MCMC sampling algorithm to learn the posteriors of the dynamic probabilistic model. With the learned model, the mechanisms of influence propagation can be approximated by a Poisson distribution on the learned latent structures. Experiments on two datasets with different content types and social networks, DBLP and Digg, prove that our approach gains more accuracy while ensures the tractability. The key contributions are summarized as follows.

- We propose INFOIBP, which approximates the user preference and user adjacency by using the latent feature model, where each latent feature represents an influential user. The information of the whole network can be approximated by the information delivered by a compact group of influentials.
- We model the dynamic evolution by constructing Markov chain on the compact latent feature representation. By integrating both content and structure information, INFOIBP is a more efficient and accurate influence propagation model in discovering the key factors in massive content and influentials in dynamic network structure.
- Extensive experiments demonstrate that INFOIBP better adapts to real data, and achieves better approximation of influence propagation mechanism over other approaches.

2. APPROACH

In IBP, with an arbitrary order, N customers enter a restaurant one by one, and select their favorite dishes from a limitless supply. The first customer selects a $Pois(\alpha)$ number of dishes. The i-th customer moves along the previously sampled dishes and serves himself according to their popularities within probability $\frac{m_k}{i}$, where m_k is the number of previous customers who have selected the k-th dish. After browsing all the selected dishes, additionally, he tries to select new dishes where the expected number is $Pois(\frac{\alpha}{i})$. Equivalently, IBP defines a prior distribution on feature loading matrix with a finite rows (customers) and an unbounded number of columns (dishes), which is widely applied to latent factor analysis and Bayesian nonparametric matrix factorization. We denote the feature loading matrix as $Z \in \mathbb{R}^{N\times K}$. A K approximation of Z is generated with beta- Bernoulli conjugate pair [15], i.e., $\pi_k|\alpha \sim Beta(\frac{\alpha}{K}, 1)$, $z_{ik}|\pi_k \sim Bernoulli(\pi_k)$, where π_k, sampled from a beta prior, represents the probability that each observation possesses the k-th feature. $z_{ik} \in \{0, 1\}$, is independently sampled from a sequence of *Bernoulli* trials, indicating whether the k-th feature is present in observation i. We can obtain IBP if we integrate π_k, $k \in [1, K]$ when $K \to \infty$. The number of active features K satisfies $K \sim Pois(\alpha H_N)$, where $H_N = \sum_i^N \frac{1}{i}$ is the N-th harmonic number.

The influence propagation can be explained by treating social network users as customers. There is an undetermined number of influentials in the network, so they can be treated as dishes in IBP. **the infections between users and influentials resemble the dish selection process**, i.e., a user has been affected by some influentials, as having selected some favorite dishes. $\pi_k \in [0, 1]$ represents the influence of the k-th influential user. Consequently, learning the influence propagation matrix Z is equivalent to learning a bipartite graph structure between users and the latent influentials. The nonzero entries in each column of Z form the reachable set for the associated influential user.

We propose INFOIBP by modeling the information propagation using the latent feature model IBP. Suppose we obtain observations at T time steps. The user preference matrix, denoted as $\{X^{(t)}\}_{t=1}^T$, where $X^{(t)} \in \mathbb{R}^{N\times D}$ denotes the observation at t-th time step. N is the number of users, and D is the dimension of preferences. Likewise, the user adjacency observation sequence is denote as $\{Y^{(t)}\}_{t=1}^T$, where $Y^{(t)} \in \mathbb{R}^{N\times N}$ is the adjacency matrix (e.g., the co-author relation in citation network, or the friend relation in online social media) of N users. Other key notations used in our model are given in Table 1. The learning procedure of INFOIBP can be described as follows.

Step 1. Build a *latent bipartite graph* between users and inflentials based on IBP, where the Bernoulli distribution are further parameterized with a user activeness degree.

Step 2. Based on the latent bipartite graph, perform nonparametric Bayesian matrix factorization on X and Y.

Step 3. Model the observation sequences $\{X^{(t)}\}_{t=1}^T$ and $\{Y^{(t)}\}_{t=1}^T$ with Hidden Markov Model (HMM).

Step 4. Perform inferencing for posteriors.

2.1 Latent Bipartite Graph Construction

Table 1: Notations in this paper for INFOIBP

Notations	Descriptioin
$Z \in \mathbb{R}^{N\times K}$	The latent bipartite graph with K influentials
$W \in \mathbb{R}^{K\times K}$	The K influentials interaction matrix
$\Phi \in \mathbb{R}^{N\times K}$	Influence weight matrix
$V \in \mathbb{R}^{K\times D}$	Influentials' preferences
α	Average degree of susceptibility to all influentials
π_k	Influence of the kth influential
v_i	Activeness of the ith user
p_{ik}	Probability of user i influenced by influential k
b_k	Persistence probability for following the kth influential
λ	Persistence parameter for evaluating influence spread

The influence propagation modeling is first attributed to constructing the *latent bipartite graph* between users and the latent influentials. Such a bipartite structure serves as the prior of the true influence propagation process. There are a few of previous work exploring the latent structure [29], [6] by using too many extensions based on IBP. In contrast, we recover the potential propagation property of IBP without introducing many extra parameters.

The generative process of IBP with K features can be directly used to build a bipartite graph with simple parameterization. Considering the equivalence between the influence propagation matrix and the bipartite graph, we use Z to represent the bipartite graph for convenience. Specifically, Z can be represented as:

$$Z = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\delta_{\omega_k}, \;\; z_{ik} \sim Bernoulli(p_{ik}) \tag{1}$$

where $\delta_{\omega_k} = \mathbb{1}(z_{ik} = 1)$ represents the bipartite graph in an additive form. z_{ik} is sampled from a Bernoulli distribution parametrized by p_{ik}, which is interpreted as the probability that the k-th influential affects the i-th user. However, the probability π_k only characterizes the influence from the k-th influential user. To model p_{ik}, the behavior history user i, i.e., the information from observations at the last time step, should also be considered. We introduce an *activeness* parameter v_i for each user i, and p_{ik} can be represented by the following multiplicative form:

$$p_{ik} = v_i\pi_k, \;\; v_i = \frac{1}{1+\exp\left(-\sum_k\left(z'_{ik} + \frac{1}{|S_i|}\sum_{j\in S_i} z'_{jk}\right)\right)} \tag{2}$$

where S_i is the neighbors of user i, and z'_{ik} denotes the influence observation between user i and influential k at the last time step. The activeness v_i is determined by considering the joint influence between the susceptibility of user i and his neighborhood users. Based on p_{ik}, we can sample a prior for the latent bipartite graph.

2.2 Bayesian Regularized Matrix Factorization

Given user preference matrix $X \in \mathbb{R}^{N\times D}$, we aim to learn a low-rank observation matrix, which can be used to approximate the original preference matrix, and identify the latent influentials. Assuming there are K influentials in the network, and they are treated as K latent features, the user preference X can be decomposed into a $N \times K$ binary feature loading matrix (i.e., the bipartite graph) Z, and a matrix $V \in \mathbb{R}^{K\times D}$, representing the preference of the current K influentials, which is the sub-matrix of X where the row indexes denote the IDs of influentials from N users.

During the Bayesian matrix factorization, we generate the feature loading matrix Z using Eqn. 1 as the prior of latent influence propagation. To avoid the overly restrictive unweighted linear combination, we introduce a mixture weight matrix $\Phi \sim \mathcal{N}(0, \sigma_\phi^2)$

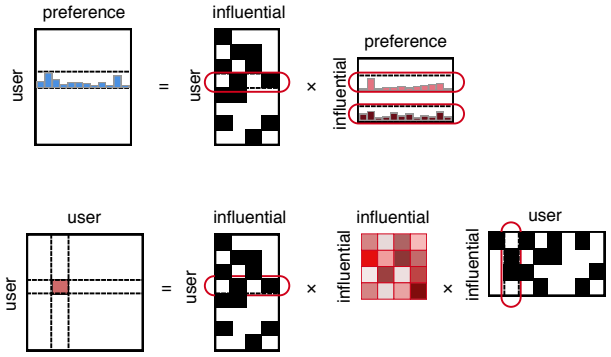


Figure 1: X and Y are jointly factorized by Z.

for model relaxation. In this way, x_i, the preference of i-th user is approximated by a weighted linear mixture of the influentials' preferences. The generative process with the user preferences observations can be formulated as a linear Gaussian model:

$$X = (Z \circ \Phi)V + \epsilon \tag{3}$$

where ϵ is the noise matrix sampled from Gaussian, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, and the operation $\circ$ indicates element-wise multiplication. Equivalently, the observation matrix X is generated from a Gaussian distribution $X \sim \mathcal{N}((Z\circ\Phi)V, \sigma_\epsilon^2 I)$. In fact, the Gaussian distribution assumption on noise can be replaced by any other specified exponential family distributions. For example, another possible choice is Poisson distribution, which has been widely used in recommendation systems. The corresponding Bayesian nonparametric matrix factorization is recognized as the Poisson factorization [13]. In the subsequent sections, we only focus on the linear Gaussian model in Eqn. 3 since the inference is comparatively convenient.

To model the generative process of users adjacencies, we resort to a popular multiplicative models, which has been widely used in link prediction modeling [21], [11], [16]. A principal form of the multiplicative models is logistic-eigendecompositions:

$$Pr(y_{ij} = 1) = \sigma\left(z_{i\cdot} W z_{j\cdot}^T\right) \tag{4}$$

where $z_{i\cdot}$ and $z_{j\cdot}$ are the i-th and j-th rows representing the indicator vectors of influentials for user i and j respectively. $\sigma(\cdot)$ represents the logistic function. $W \in \mathbb{R}^{K\times K}$ is a real-valued symmetric model parameter matrix. In our model, W is sampled from a prior distribution, providing regularization effect on W. The prior distribution can be either a normal distribution with mean 0 and variance σ_n or a Laplacian distribution with parameter σ_l. Based on the multiplicative model in Eqn. 4, similar to collaborative filtering, the more influentials shared by user i and j, the higher probability that the link between them would appear. Here, the logistic function $\sigma(\cdot)$ works as a mapping function from latent influential space to probability space. The user adjacencies are finally sampled from Bernoulli trials.

Note that Z is shared by both X and Y in Eqn. 3 and 4, respectively. Intuitively, when new links are formed, the influentials could influence more users to change their preferences. With the change of user preferences, new influentials with more distinguished preferences are likely to be identified, which consequently drives new links among users to emerge. A demonstration of the Nonparametric matrix factorization of X and Y is showed in Figure 1.

2.3 Temporal Evolution

We extend single time step probabilistic modeling to describe the dynamic temporal evolution. To this end, all the latent bipartite graphs at each time step need to be stringed together. Each entry z_{ik} of the latent bipartite graph is associated with an independent

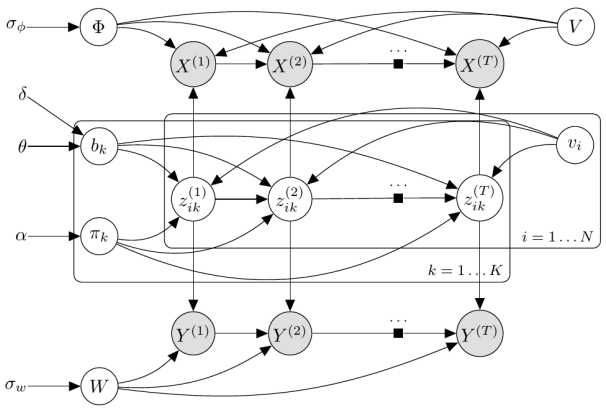


Figure 2: Graphical model of influence propagation in dynamic networks.

Markov chain $z_{ik}^{(1)}, z_{ik}^{(2)}, \dots, z_{ik}^{(T)}$. For any two contiguous time slices, we define a probability transition matrix [11], [12]:

$$Q_{ik} = \begin{pmatrix} 1 - p_{ik} & p_{ik} \\ 1 - b_k & b_k \end{pmatrix} \tag{5}$$

where p_{ik} is the probability of being influenced at the next time step, which is similarly defined in Eqn. 1. For the influence part of p_{ik}, i.e., π_k, is sampled from the beta prior $Beta(\frac{\alpha}{K}, 1)$. b_k denotes the user independent persistence parameter, which measures the probability that the users are kept influenced at the next time step. b_k is sampled from another beta prior $Beta(\theta, \delta)$. Note that in both Eqn. 3 and 4, all the other latent variables are assumed to be temporarily constant except for Z, X and Y.

A schematic illustration of the full graphical model for the influence propagation in the dynamic networks is given by Figure 2. The model can be formulated as:

$$\begin{aligned} \pi_k &\sim Beta(\frac{\alpha}{K}, 1),\ b_k \sim Beta(\theta, \delta) \\ z_{ik}^{(t)} &\sim Bernoulli\left((\pi_k v_i)^{1 - z_{ik}^{(t-1)}} b_k^{z_{ik}^{(t-1)}} \right) \\ w_{kk'} &\sim \mathcal{N}(0, {\sigma_w}^2),\ \phi_{ij} \sim \mathcal{N}(0, \sigma_\phi^2) \\ y_{ij}^{(t)} &\sim Bernoulli\left(\sigma\left(z_{i\cdot}^{(t)} W {z_{j\cdot}^{(t)}}^T \right) \right) \\ x_{ij}^{(t)} &\sim \mathcal{N}\left((z_{i\cdot}^{(t)} \circ \phi_{i\cdot}) \cdot V_{\cdot j}^{(t)}, {\sigma_\epsilon}^2 \right) \end{aligned} \tag{6}$$

Unlike the canonical latent feature model, where the latent feature space is unknown, the latent influential space of our proposed model is equivalent to the user space, and the influentials are the subset of the whole group of users. Therefore, the problem of influence propagation is equivalent to a set selection problem, which is originally an NP-hard problem. By incorporating the content modeling on user preferences, we obtain the model solution on dynamic social content with dynamic affinities, which identifies the true influentials (i.e., the user indexes) along the whole observed temporal range.

3. MODEL INFERENCE

To obtain the posterior of Z, we use the forward-backward algorithm [24] to approximate the full posterior distribution of $z_{ik}^{(t)}$ for the whole temporal ranges. In the forward pass, to collect the evidences of all the possible state transitions of Z, all the possible states are traversed and the transition probabilities are calculated. In order to enhance the computational efficiency, a dynamic programming cache for recording the transition probability from Z^t to Z^{t+1}, $t = 1, ..., T-1$ is created during the learning procedure. In the backward pass, each state variable $z_{ik}^{(t)}$ is sampled from the programming cache. The posterior states of Z in the reversed order $T, T-1, \dots, 1$ can be obtained.

Specifically, in the forward pass, all the transition probabilities are cached in a $T-1$ sequence with 2×2 transition matrices $\{P_2, P_3, \dots, P_T\}$, whose entry is denoted by $p_{rs}^{(t)}$. The forward pass can be described as:

$$\begin{aligned} p_{rs}^{(t)} &= Pr(z_{ik}^{(t-1)} = r, z_{ik}^{(t)} = s | \{X^{(l)}\}_{l=1}^{t}, \{Y^{(l)}\}_{l=1}^{t}, \Omega) \\ &\propto \Big[Pr(z_{ik}^{(t-1)} = r | \{X^{(l)}\}_{l=1}^{t-1}, \{Y^{(l)}\}_{l=1}^{t-1}, \Omega) \\ &\quad \cdot Q_{ik}(r, s) \cdot Pr(Y^{(t)}, X^{(t)} | z_{ik}^{(t)} = s, \Omega) \Big] \end{aligned} \tag{7}$$

$$\sum_r p_{rs}^{(t-1)} = Pr(z_{ik}^{(t-1)} = r | \{X^{(l)}\}_{l=1}^{t-1}, \{Y^{(l)}\}_{l=1}^{t-1}, \Omega) \tag{8}$$

where Ω represents all the other latent variables and parameters, $Q_{ik}(r, s)$ is the transition prior probability given by Eqn. 5. In backward pass, a simple sampling is carried out as:

$$Pr(z_{ik}^{(t)} = r | z_{ik}^{(t)}, \{Y^{(l)}\}_{l=1}^{t+1}, \{X^{(l)}\}_{l=1}^{t+1}, \Omega) \propto p_{r, z_{ik}^{(t+1)}}^{(t+1)} \tag{9}$$

As mentioned in Section 3.3, calculating the posterior of V is essentially a set selection problem. Considering the non-conjugate combinatorial property of identifying the indices for V, the Metropolis-Hastings algorithm is applied.

$$Pr(\text{accept } I_V^*) = \min\left\{ 1, \frac{Pr(\{X^{(t)}\}_{t=1}^{T}, \{Y^{(t)}\}_{t=1}^{T} | I_V^*) Pr(I_V^*)}{Pr(\{X^{(t)}\}_{t=1}^{T}, \{Y^{(t)}\}_{t=1}^{T} | I_V) Pr(I_V)} \right\} \tag{10}$$

where the index set of candidate influentials is denoted as I_V^*, and the index set of the influentials at the last sampling step is denoted as I_V. The likelihood $Pr(\cdot)$ can be calculated independently from X and Y. To generate a proposal of influential users index set I_V^*, three different sampling algorithms are used to be applied.

Uniform sampling. Given the user index set $\{1, 2, \dots, N\}$, we uniformly sample K candidates as the influential users. Recall the equivalence between the influential user space and common user space, the prior for the index of any influential user being sampled is equal to $\frac{1}{N-K}$. Therefore, the prior can be canceled from the formulation of the acceptance ratio in Eqn. 10. Consequently, the acceptance ratio reduces to a likelihood ratio as:

$$Pr(\text{accept } I_V^*) = \min\left\{ 1, \frac{Pr(\{X^{(t)}\}_{t=1}^{T}, \{Y^{(t)}\}_{t=1}^{T} | I_V^*)}{Pr(\{X^{(t)}\}_{t=1}^{T}, \{Y^{(t)}\}_{t=1}^{T} | I_V)} \right\} \tag{11}$$

Degree reduction sampling. Users who have large degrees to their neighbors are more likely to be influentials. To sample K influentials from all users, we pre-compute the user degrees and rearrange their indices with degree decreasing. We can fit this degree reduction curve with an exponential distribution $\frac{1}{\beta} e^{-\frac{x}{\beta}}$. The proposed influential users set can be sampled from this distribution until K different indices are got. Note that $K \ll N$, we can assume that the degree reduction curve will not changed between any two iterations. The prior items both in numerator and denominator can also be canceled when calculating the acceptance ratio.

Determinantal point process. From content analysis perspective, we expect the influential users work as a compact representation for all the users. In other words, the selected influential users need to cover the whole user set in some feature space as much as possible. Determinantal point process (DPP) [18] has been widely

used to select diverse sample points from a specific feature space for the whole dataset. In this paper, we take the space expanded by user preference as the feature space, so that the user preference observations can be directly taken as the sample points. In DPP, the probability of the selected samples $\mathcal{Y}$ equals to the determinant of the marginal kernel matrix of samples $\mathcal{Y}$, i.e. $Pr(\mathcal{Y}) = \det(K_{\mathcal{Y}})$. We use the algorithm 3 in [18] to sample the influential users. In this situation, the prior ratio defined in Eqn. 10 turns to the ratio of each determinant, $\frac{\det(K_{I_V^*})}{\det(K_{I_V})}$.

3.1 Time Complexity of Model Inference

There are two time-consuming steps for generating a sample. One is the forward-backward samplings for hidden Markov chain, and the other is sampling the indexes of influentials. In the forward-backward pass, for each sample $z_{ik}^{(t)}$, the computational complexity of calculating likelihoods defined by Eqn. 3 and 4 are $\mathcal{O}(K^2NDT)$ and $\mathcal{O}(K^2N^2T)$, respectively. Due to the sparsity of observation considering the large number of users in a long time range, the number of non-zero observations is denoted as $L \sim \max\{N, D\}$, where $\sim$ means equivalent by order of magnitude. Additionally, recall that K follows the distribution $K \sim Pois(\alpha H_N)$, the complexities of both Eqn. 3 and 4 can be represented by $\mathcal{O}(\alpha^2 LT\log^2 N)$. Likewise, for sampling indices of influentials in Eqn. 10, it also requires $\mathcal{O}(\alpha^2 LT\log^2 N)$ computations. But the number of samples reduces to $K(N-K)$. In other words, the complexity of our inference algorithm for INFOIBP is **equivalent to** most of the existing link prediction approaches based on latent feature models.

4. EVALUATING THE INFLUENCE

The active number K of features in IBP follows the distribution $Pois(\alpha H_N)$. In this section, we prove that to evaluate the influence of K influentials (i.e., calculate the number of rows with at least on non-zero element in Z) can inferred by a Poisson distribution.

Each column of Z is an influential user that has influenced at least one individual user, and is a sequence of N independent Bernoulli trials with the probability p_{ik} given by Eqn.1. Since p_{ik} is characterized by user i and influential k simultaneously, these Bernoulli trials are not identically distributed. To this end, a specific distribution called *Poisson binomial distribution* is introduced to calculate the expectation of influence, which is a discrete probability distribution of a sum of independent Bernoulli trials.

Consider the k-th column of $Z^{(t)}$ at t-th time step, $Z_{\cdot k}^{(t)}$, it is a collection of N independent Bernoulli distributed variables, which is known as Poisson binomial distribution. The expectation is the mean of the expectation of N Bernoulli distributions, i.e., $\mathbb{E}[Z_{\cdot k}^{(t)}] = \sum_{i=1}^{N} \tilde{p}_{ik}^{(t)}$, where

$$\tilde{p}_{ik}^{(t)} = \lambda \mathbb{1}(z_{ik}^{(t-1)} = 1) + (1-\lambda) v_i^{(t)} \pi_k \mathbb{1}(z_{ik}^{(t)} = 1) \qquad (12)$$

characterizes the influence probability at t-th step. It has two components weighed by a global persistence parameter λ, the first part is the probability of keeping the state as last time step, and the second part represents the impact of current state. For any user i, i.e., the i-th row of $Z^{(t)}$, the probability that he has been influenced at t-th step is:

$$Pr(Z_{i\cdot}^{(t)} \cdot \mathbf{1} > 0) = 1 - \prod_{k=1}^{K} (1 - \tilde{p}_{ik}^{(t)}) \qquad (13)$$

The expectation of the number of all the influenced users is:

$$\mathbb{E}\left[\sum_i \mathbb{1}(Z_{i\cdot}^{(t)} \cdot \mathbf{1} > 0)\right] = \sum_{i=1}^{N} 1 - \prod_{k=1}^{K} (1 - \tilde{p}_{ik}^{(t)}) \qquad (14)$$

where $\mathbb{1}(\cdot)$ is an indicator function. The Poisson binomial distribution can be approximated by a Poisson distribution given by *Le Cam's Theorem* [19]. Consequently, according to *Le Cam's Theorem*, the expected number of influenced users can be estimated by Poisson distribution with the rate mass in Eqn. 14.

5. EXPERIMENTS

The effectiveness of INFOIBP needs to be validated from two aspects. First, how can the model accurately predict the how the users are influenced in the influenced propagation process. Second, by integrating content modeling, how the effectiveness of our method can be enhanced for influence maximization.

Evaluation Criteria. We measure the performances of *link prediction* and *preference prediction* for a set of users, which have been widely adopted in previous study [2], [21], [11]. To evaluate the performance of influence maximization, we measure the *degree of influence spread*, i.e., the number of influenced users given a specific number of influentials. The more users that have been influenced, the more powerful the model will be.

Datasets.We conduct experiment on two different types of real-world social content data. **DBLP**[1] is a co-authorship network with $1.27M$ authors and $8.3M$ co-authorships. We extract a subset with 2213 active authors through 20 years ranging from 1993 to 2012. The observation interval is set to be 1 year on this data. For user preference modeling, we trained an LDA model with 50 topics to get the topic distribution of each authors. **Digg**[2] is a story voting social network with $139K$ users and $1.7M$ friend relationships. We extract a subset including 393 active users, whose friend relationships and voting histories on 3553 stories are collected over 30 days. The observation interval is set to be 1 day. We use the 3553-dim voting vector to represent the preference of each user.

5.1 Predictive Accuracy

Different social networks have different influence propagation behaviors. For example, in DBLP, the information propagation process covers a long range of 20 years. the influence is established by co-authorship among authors. Therefore, Predicting the co-authorship can be directly reduced to link prediction based on the latent influence propagation model. For Digg dataset, the temporal range is much shorter, and the change in user preference is drastic. Users are described by a high dimensional vectors, which reflects more information on the user preferences.

Predicting the links and preferences is equivalent to calculating $Pr(Y^{(t)}|Y^{(t-1)})$ and $Pr(X^{(t)}|X^{(t-1)})$ defined in Eqn. 7. In order to approximate the conditional distributions, we generate multiple samples of the influence propagation matrices $\{Z^{(l)}\}_{l=1}^{t-1}$. After 300 burn-in samples, we obtain 50 samples for each of 10 MCMC chains. Then 10 independent samples of $Z^{(t)}$ are drawn from the samples. Finally, we choose the average likelihood to approximate the predicting conditional distribution.

Link Prediction. On both DBLP and Digg, we perform link prediction and use the test log-likelihood as the performance measurement. We compare the following approaches:

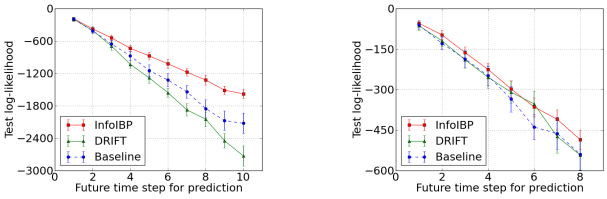
DRIFT [11]: a nonparametric model which undertakes the link prediction task without considering content information.

Baseline: the truncated version of INFOIBP with $\alpha = 1.0$, which stops extending the number of influentials when the truncation level $K = 10$ is reached.

INFOIBP-US, using uniform sampling to propose influential user indices, the full version with $\alpha = 1.0$, where the average value of K is around 35 in the experiment.

[1] http://arnetminer.org/DBLP_Citation

[2] http://www.isi.edu/ lerman/downloads/digg2009.html



(a) Prediction on DBLP. (b) Prediction on Digg.

Figure 3: Predicting network structures.

Table 3: AUC values of preference prediction on Digg.

Test	Link Estab. Test	Sel. Probs.
INFOIBP (avg)	**0.875**	**0.873**
AIR	0.859	0.812
TIC	0.619	0.771
IC	0.619	0.792

INFOIBP-DRS, using degree reduction sampling to propose influential user indices, the full version with $\alpha = 1.0$, where the average value of K is around 40 in the experiment.
INFOIBP-DPP, using determinantal point process to sample influential user indices, the full version with $\alpha = 1.0$, where the average value of K is around 37 in the experiment.

Figure 3 shows the comparison of different models. The result reflects that our method achieves 42.2% and 17.1% performance gain on DBLP and Digg dataset,respectively. Table 2 shows the AUC values for predicting the link establishment over 8 future time steps. Results show that there is little difference when using different sampling algorithm to generate influential user indices. Our method outperforms DRIFT by 41.9% on average. We also explore how the accuracy is improved by increasing the truncation level by comparing the three versions of baseline. The accuracy gain decreases as the truncation level grows from 10 to 30. Figure 4 shows the ROC curves of the first 8 time-step predictions on DBLP.

Preference Prediction. Similar to [2], we test the performance on link establishment test and selection probabilities respectively. The former is carried out as a binary prediction task, where we take every user-item pair as a Bernoulli trail parameterized by the predicted probability. For the latter case, it can be directly performed with the predicted probability. On Digg dataset, we focus on the voting behavior and evaluate the performance again with AUC values. We evaluate the predictive accuracy in selection probabilities test with the predicted likelihoods. By varying the link establishment thresholds, we can obtain the AUC values. The link establishment test requires an extra adaptive binarizing step. The results are shown in table 3. Compared to AIR, TIC and IC model reported by [2], our method outperforms existing methods on both tests.

5.2 Influence Maximization

The problem of influence maximization is to maximize the number of influenced users with a fixed number of influentials. On dynamically evolved social networks, the influence should be maximized as much as possible during the whole evolution process, so that the influence can be globally maximized or approximately globally maximized. Moreover, we need to study whether integrating the content information can maximize the influence.

For INFOIBP, to get different values of K, we run our model with different truncation levels. For evaluating the influence spread, we make use of the Poisson approximation guaranteed by *Le Cam's Theorem* [19]. We compare the expected influence spread to the actual influence spread from the trained model over 10 time steps

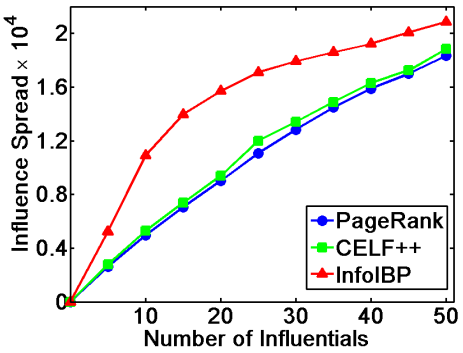


Figure 6: Influence spread compared to CELF++ and PageRank at a single time step.

in Figure 5 (a)-(c). The curves in Figure 5(a), 5(b) and 5(c) correspond to different persistence parameters $\lambda = 0.5, 0.7, 0.9$. The results illustrate that the Poisson approximations accurately estimate the trend of influence spread over the temporal evolution. By increasing the value of persistence λ, more accurate estimation can be obtained due to the inertia of temporal evolution.

We run the influence spread at a single time step and compare to other two approaches, which only consider the network structure information.
CELF++ [14], a state-of-the-art model derived from *Greedy*.
PageRank [5], a popular algorithm to rank the importance of web pages, which is also used for picking the k highest-ranked nodes as the most k influential users.

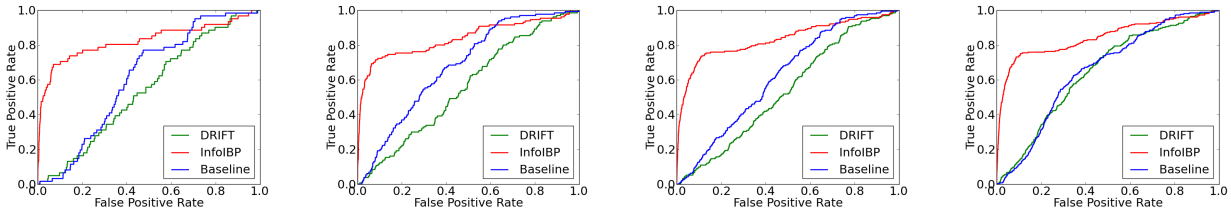
Figure 6 reveals that by integrating the user preferences, our model outperforms CELF++ and PageRank with a more extensive influence spread. Any two users who share some preferences may influence each other with a certain probability, although there is no direct path between them. In summary, the experiment validates that by integrating the content information, our model INFOIBP improves the propagation spread over other state-of-the-art approaches.

5.3 Running Time

We compare our method with the following influence propagation models, which both explicitly consider the influence propagation at a given time.
INFLUMAX [23]: A continuous time Markov process to model the influence propagation, where each state is the dominate set of the current node.
CONTINEST [10]: A graph sampling approach which effectively estimates the neighborhood size.
We sample the transmission rate from e^{-t} for every edge in these two models and explore the influence propagation for 10 time steps on a subgraph with 230 nodes and 790 edges on DBLP. Figure 7(a) shows that our model has a comparable running time with CONTINEST. Because there is a exponential complexity for finding the dominate sets, INFLUMAX turns to be intractable when the target number of influentials is large. A similar comparison in Figure 7(b) indicate that the running time increases with the number of users N by setting $K = 10$.

6. RELATED WORK

Researchers have embarked on content related influence analysis [25], [20], [28], [2], [3]. However, the models focusing on user-to-user influence on topics [25], [20], [28], have not worked on the influence maximization task yet. Two topic-aware extensions to the canonical Independent Cascade Model (Topic-aware



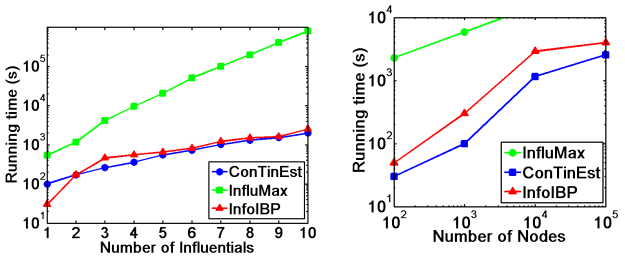
(a) Future time step 1 (b) Future time step 2 (c) Future time step 3 (d) Future time step 4

(e) Future time step 5 (f) Future time step 6 (g) Future time step 7 (h) Future time step 8

Figure 4: ROC curves for predicting 8 future time steps on DBLP dataset.

Table 2: AUC values of link establishment prediction on DBLP.

Time step	1	2	3	4	5	6	7	8
Baseline-10	0.612	0.679	0.622	0.653	0.689	0.523	0.604	0.587
Baseline-20	0.731	0.749	0.720	0.744	0.761	0.707	0.711	0.699
Baseline-30	0.812	0.823	0.817	0.834	0.829	0.815	0.819	0.808
DRIFT	0.544	0.558	0.534	0.646	0.584	0.589	0.624	0.467
INFOIBP-US	**0.824**	**0.830**	**0.831**	**0.834**	**0.839**	**0.830**	**0.831**	**0.836**
INFOIBP-DRS	**0.832**	**0.838**	**0.833**	**0.842**	**0.842**	**0.837**	**0.839**	**0.844**
INFOIBP-DPP	**0.837**	**0.835**	**0.843**	**0.832**	**0.838**	**0.834**	**0.836**	**0.839**
INFOIBP (avg)	**0.831**	**0.834**	**0.836**	**0.836**	**0.840**	**0.834**	**0.835**	**0.840**



(a) Running time vs. K (b) Running time vs. N

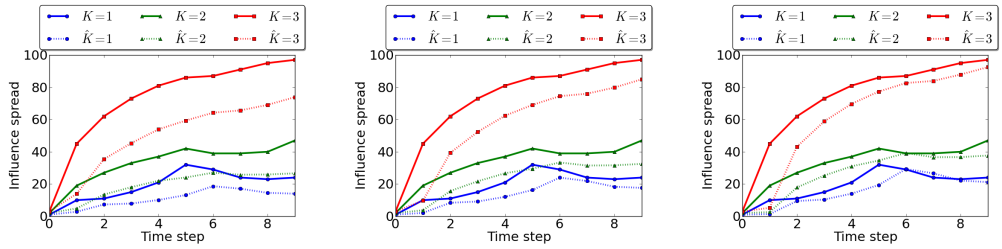
Figure 7: Running time comparison.

Independent Cascade Model, TIC) and Linear Threshold Model (Authoritativeness-Interest-Relevance, AIR) have been proposed in [2], where user-to-user influence is parameterized by a topic distribution substantially learned from the network structure, which is not really involving the content of topics. A novel model Followship-LDA [3] tries to perform topic discovery and social influence analysis simultaneously. Nevertheless, the generative process is overly dependent on LDA topic modeling and the work is biased in favor of scaling the computing from a system perspective. All the works above do not consider influence propagation in dynamic networks.

On the contrary, there are a few works analysing influence on dynamic networks. Wang, et. al. [27] extends the work [25] to dynamic networks which still focuses on user-to-user influence dynamic. Aggarwal, et. al. [1] proposes a forward-backward approach to trace the information flow by discovering interaction patterns between any pair of users. A novel model MaxG proposed by [33] tries to uncover the influence diffusion in dynamic networks by probing a subset of nodes. As a more challenging problem, Zhan *et al.* [30] consider influence maximization across heterogenous social networks with partially aligned users. Besides, influence maximization is closely related to the problem of link prediction in social networks [32], where a comprehensive literature review can be found in [31]. However, this category of research works do not take content information into account for influence estimation.

7. CONCLUSION

We propose INFOIBP, a new influence propagation model from nonparametric Bayesian perspective. INFOIBP integrates both content and structure information into Markov Chain based temporal evolution modeling of influence propagation Under the proposed latent influence propagation framework, learning the influence propogation mechanism is equivalent to solving tractable models of matrix factorization and Maximum-a-posteriori-estimation in polynomial complexity. We experimentally find the proposed method is more accurate in modeling the influence propagation mechanism and predicting the user activities. In future work, we will extend our method to deal with continuous-time social content and network structure observations using the functional theory. We also intend to devise a more efficient algorithm for scalable inference.



(a) $\lambda = 0.5$ (b) $\lambda = 0.7$ (c) $\lambda = 0.9$

Figure 5: Influence spread evaluation on DBLP. (a)-(c) give the approximated expected influence spread (dotted) vs. actual influence spread counted from trained model (solid).

8. REFERENCES

[1] C. C. Aggarwal, S. Lin, and P. S. Yu. On influential node discovery in dynamic social networks. SDM '12, pages 636–647, 2012.

[2] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. ICDM '12, pages 81–90, 2012.

[3] B. Bi, Y. Tian, Y. Sismanis, A. Balmin, and J. Cho. Scalable topic-specific influence analysis on microblogs. WSDM '14, pages 513–522, 2014.

[4] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. SODA '14, pages 946–957, 2014.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30:107–117, 1998.

[6] F. Caron. Bayesian nonparametric models for bipartite graphs. NIPS '12, pages 2051–2059. 2012.

[7] S. Chen, J. Fan, G. Li, J. Feng, K. lee Tan, and J. Tang. Online topic aware influence maximization. In *Proceedings of VLDB Endowment*, volume 8, 2015.

[8] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. KDD '10, pages 1029–1038, 2010.

[9] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. KDD '09, pages 199–208, 2009.

[10] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha. Scalable influence estimation in continuous-time diffusion networks. NIPS '13, pages 3147–3155. 2013.

[11] J. R. Foulds, C. Dubois, A. U. Asuncion, C. T. Butts, and P. Smyth. A dynamic relational infinite feature model for longitudinal social networks. AISTATS '11, pages 287–295, 2011.

[12] J. V. Gael, Y. W. Teh, and Z. Ghahramani. The infinite factorial hidden markov model. NIPS '08, pages 1697–1704. 2008.

[13] P. Gopalan, F. J. Ruiz, R. Ranganath, and D. Blei. Bayesian nonparametric poisson factorization for recommendation systems. AISTATS '14, pages 275–283, 2014.

[14] A. Goyal, W. Lu, and L. V. Lakshmanan. Celf++: Optimizing the greedy algorithm for influence maximization in social networks. WWW '11, pages 47–48, 2011.

[15] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the indian buffet process. In *NIPS*, pages 475–482, 2005.

[16] C. Heaukulani and Z. Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. ICML '13, pages 275–283, 2013.

[17] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD '03, pages 137–146, 2003.

[18] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. arXiv preprint arXiv:1207.6083, 2012.

[19] L. Le Cam. An approximation theorem for the poisson binomial distribution. *Pacific Journal of Mathematics*, 10:1181–1197, 1960.

[20] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. CIKM '10, pages 199–208, 2010.

[21] K. Miller, M. I. Jordan, and T. L. Griffiths. Nonparametric latent feature models for link prediction. NIPS '09, pages 1276–1284. 2009.

[22] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. KDD '02, pages 61–70, 2002.

[23] M. G. Rodriguez and B. Schölkopf. Influence maximization in continuous time diffusion networks. ICML '12, pages 313–320, 2012.

[24] S. L. Scott. Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97:337–351, 2002.

[25] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. KDD '09, pages 807–816, 2009.

[26] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. SIGMOD '14, pages 75–86, 2014.

[27] C. Wang, J. Tang, J. Sun, and J. Han. Dynamic social influence analysis through time-dependent factor graphs. ASONAM '11, pages 239–246, 2011.

[28] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. WSDM '10, pages 261–270, 2010.

[29] F. Wood, T. L. Griffiths, and Z. Ghahramani. A non-parametric Bayesian method for inferring hidden causes. UAI '06, pages 536–543, 2006.

[30] Q. Zhan, J. Zhang, S. Wang, P. Yu, and J. Xie. Influence maximization across partially aligned heterogenous social networks. In *PAKDD*, pages 58–69, 2015.

[31] J. Zhang and P. S. Yu. Link prediction across heterogeneous social networks: A survey. 2014.

[32] J. Zhang and P. S. Yu. Integrated anchor and social link predictions across partially aligned social networks. In *IJCAI*, 2015.

[33] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun. Influence maximization in dynamic social networks. ICDM '13, pages

1313–1318, 2013.