Fast Sampling for Time-Varying Determinantal Point Processes

MAOYING QIAO, Centre for Quantum Computation & Intelligent Systems
and University of Technology Sydney
RICHARD YI DA XU, University of Technology Sydney
WEI BIAN and DACHENG TAO, Centre for Quantum Computation & Intelligent Systems
and University of Technology Sydney

Determinantal Point Processes (DPPs) are stochastic models which assign each subset of a base dataset with a probability proportional to the subset's degree of diversity. It has been shown that DPPs are particularly appropriate in data subset selection and summarization (e.g., news display, video summarizations). DPPs prefer diverse subsets while other conventional models cannot offer. However, DPPs inference algorithms have a polynomial time complexity which makes it difficult to handle large and time-varying datasets, especially when real-time processing is required. To address this limitation, we developed a fast sampling algorithm for DPPs which takes advantage of the nature of some time-varying data (e.g., news corpora updating, communication network evolving), where the data changes between time stamps are relatively small. The proposed algorithm is built upon the simplification of marginal density functions over successive time stamps and the sequential Monte Carlo (SMC) sampling technique. Evaluations on both a real-world news dataset and the Enron Corpus confirm the efficiency of the proposed algorithm.

CCS Concepts: • **Information systems** → **Information systems applications**; **Data mining**; *Spatial-temporal systems*

Additional Key Words and Phrases: Time-varying determinantal point processes (TV-DPPs), sequential Monte Carlo, fast sampling

ACM Reference Format:
Maoying Qiao, Richard Yi Da Xu, Wei Bian, and Dacheng Tao. 2016. Fast sampling for time-varying determinantal point processes. ACM Trans. Knowl. Discov. Data 11, 1, Article 8 (July 2016), 24 pages.
DOI: http://dx.doi.org/10.1145/2943785

1. INTRODUCTION

Determinantal Point Processes (DPPs) naturally model repulsion interaction where diverse subsets are preferred. It arises as an important tool in random matrix theory [Ginibre 1965], physics (fermions, eigenvalues of random matrices) [Macchi 1975], and Combinatorics (non-intersecting paths, random spanning trees) [Hough et al. 2006]. Recently, it has been introduced to the machine learning field, and shown to be particularly valuable in many data mining applications, such as text summarization [Kulesza and Taskar 2011b], document thread revealing [Gillenwater et al. 2012a], topic model

This work is supported by Australian Research Council Projects: FT-130101457, DP-140102164 and LE-140100061.

Authors' addresses: M. Qiao, W. Bian, and D. Tao, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia; emails: Maoying.Qiao@student.uts.edu.au, brian.weibian@gmail.com, Dacheng.Tao@uts.edu.au; R. Xu, Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia; email: YiDa.Xu@uts.edu.au.

DOI: http://dx.doi.org/10.1145/2943785

[Zou and Adams 2012], information retrieval [Kulesza and Taskar 2011a], pose estimation [Kulesza and Taskar 2012], and neural inhibition [Snoek and Adams 2013].

DPPs have exact inference/sampling algorithms, and they are developed based on eigen-decomposition of DPPs' kernel matrix. Therefore, the time complexity of DPPs sampling is $\mathcal{O}(N^3)$ [Kulesza and Taskar 2012], where N is the total number of items in the dataset. Such time complexity makes DPPs infeasible for real-time processing, especially when N is large. In order to improve the efficiency of DPP inference algorithms, different approaches have been proposed. Dual representation is introduced when the kernel matrix is low-rank (e.g., rank $D \ll N$) based on Gram matrices [Kulesza and Taskar 2012], and the time complexity is reduced to $\mathcal{O}(ND+D^3)$. Nyström approximation and Matrix Ridge Approximation (MRA) [Wang et al. 2014] have also been introduced to enhance the efficiency of eigen-decomposition of a kernel matrix, which is the most time-consuming operation for DPP sampling. Most recently, a scheme based on Markov Chain Monte Carlo (MCMC) technique for DPPs sampling has been proposed by Kang [2013], and the ϵ-mixing time is $\mathcal{O}(N \log(N/\epsilon))$.

The above approaches aim to improve the efficiency of DPPs' computation by focusing on its inference algorithms. However, they do not take into consideration the structure of the data itself, which potentially can be exploited as a valuable source of efficiency improvement [Rakthanmanon et al. 2013]. One such data structure is often found in a setting where there are sequential changes occurring in a dataset, but the changes in each time interval are relatively small when compared to the entries of the entire dataset. This structure has been seen in many online services [Abel et al. 2013]. Taking online news services as an example, the services strive to sequentially display a diverse subset of news sampled from the most recently updated news corpora from many third-party sources. It is easily observed that any real-time updates of the news corpora are relatively small in a short time interval.

In this paper, we propose to derive a fast sampling algorithm to improve the efficiency of time-varying DPPs in handling large-scale datasets. In a time-varying setting, records are collected from many information sources at each time stamp, and the whole dataset is built sequentially. Because the update rates of different information sources vary, only a small proportion of them have new information feeds during a short time interval. Such relatively small changes are utilized to develop a fast sampling scheme for time-varying determinantal point processes (TV-DPPs). We improve efficiency by incorporating a simplified computation of successive marginal density functions into an sequential Monte Carlo (SMC) framework. Our contributions are summarized as follows.

—We propose a novel TV-DPPs setting to accomplish real-time diverse subset sampling task with making use of successive, proportionally small updates of information sources between two successive time stamps, with respect to the overall large-scale dataset.
—We embed a simplification computation over successive marginal density functions into the framework of SMC sampling techniques to obtain a fast DPP sampling algorithm for a sequentially collected large-scale dataset.
—We evaluate the accuracy and efficiency of the proposed algorithm on two real-world scenarios, including news recommendation and Enron event discovery.

The rest of this paper is arranged as follows. Section 2 presents related works. Section 3 introduces the background of two key techniques: DPPs and the SMC method. The proposed TV-DPPs setting and its fast sampling algorithm are detailed in Section 4, and the experimental results on two real-world datasets are reported in Section 5. Section 6 concludes this paper and provides discussions on future work.

2. RELATED WORK

DPPs provide a probability measure over every configuration of subsets on data points, and assign higher probabilities to subsets with dissimilar items. They are widely used to capture the repulsion amongst particles [Amer-Yahia et al. 2014] and to select diverse subsets. One reason for the popularity of DPPs is that their inference can be solved in polynomial time, which is required by many real-time large-scale applications. As for both discrete and continuous variants of DPP [Affandi et al. 2013a], there have emerged many works well developed for both parameter learning and basic inference algorithms, such as Kulesza and Taskar [2012] for marginal and conditional inference as well as sampling algorithm, Gillenwater et al. [2012b] and Affandi et al. [2014] for maximum a posteriori (MAP) estimation, Gillenwater et al. [2014] for learning of DPP kernels, and Biscio and Lavancier [2014] for measuring the repulsion of DPPs.

A diverse subset preferred by a DPP can be seen as covering or summarizing relevant explanations of associated topics, which is useful in a variety of applications. For example, in an information retrieval system, a good search result should be a diverse subset covering a significant interpretation of a query [Carbonell and Goldstein 1998; Raman et al. 2012]. In a text mining field, a good summarization should be a diverse sentence subset which contains unique aspects of the text, avoiding repetitive information [Kulesza and Taskar 2011b; Xu et al. 2014; Xuan et al. 2015b]. As for multiple structure detection task, a good structure subset should contain non-overlapping predictions, rather than share parts of the structure; an example of this is a multiple pose estimation task.

However, directly applying DPP to complex application scenarios has limitations. For example, structural elements are ubiquitous in many application domains, such as chain structure for trajectory and news thread, and pictorial structures for human poses. Based on factorization of structures, both quality and similarity can be directly constructed. However, due to high-dimensional variables in structures, the size of a base set will become exponentially large. Taking advantage of dual representation, Kulesza and Taskar [2010] derive a tractable structure DPP.

Another example is for sequential application scenarios, where we listed two instances. One instance is an online news service system. It tries to provide every user with sequential news subsets, where the diversity not only applies to news articles at any individual point in time, but also needs to be addressed temporally. In order to fulfil such requirements, Markov DPP is developed [Affandi et al. 2013b; Liu et al. 2016]. It models a sequence of random sets which come from a large-scale base set, and maintains two kinds of margin DPPs: one at each single time stamp and the other for pairwise time stamps, and one kind of conditional DPPs for the transitional probability distributions. Therefore, existing inference algorithms for DPPs can be directly applied. However, the inference algorithms could not be effective for a large-scale dataset, since the size of the kernel matrices for each DPP is as large as the cardinality of the base set, which presents itself as a major bottleneck. The other instance of sequential application scenarios is large-scale video summarization. Recently, Gong et al. [2014] developed a sequential DPP model which incorporates the concept of diversity for extracting succinct subsets to summarize large-scale videos. Instead of taking whole frames as a base set as Markov DPPs do, a sequential DPP employs a divide-and-conquer strategy. It first partitions the whole video into disjoint yet consecutive segments, and then maintains margin DPPs for the union of two neighbouring segments. Consequently, the inference for a sequential DPP is more efficient. This setting is different from what we have proposed in this paper, Because of the following two reasons: (1) each segment is sequentially collected from a single source (i.e., the same video) rather than synchronously collected from different information sources, and (2) the divisions do not overlap.

Sequential subset selection techniques based on criteria other than diversity have also been developed in different areas, and we refer interested readers to works of Chen and Hsu [1991], Tollefson et al. [2014], Rao et al. [2003], Liu and Tao [2016], and Xu et al. [2015].

3. BACKGROUND

3.1. Determinantal Point Processes (DPPs)

A point process $\mathcal{P}$ on discrete ground set $\mathcal{S} = \{1, 2, \ldots, N\}$ is called a DPP with a positive semidefinite matrix K indexed by the elements of $\mathcal{S}$ if, when X is a random subset drawn according to $\mathcal{P}$, for every $x \subseteq \mathcal{S}$

$$\mathcal{P}(X \supseteq x) = \det(K_x), \tag{1}$$

for $K \preceq I$. Here, $K_x \equiv [K_{ij}]_{i,j \in x}$ is the restriction of K to the entries indexed by elements of x, and we adopt the convention that $\det(K_\emptyset) = 1$. K is referred as marginal kernel. Clearly, DPP is a probability measure over all $2^{\mathcal{S}}$ subsets of $\mathcal{S}$. The most relevant construction of DPP for our purpose of modelling real data is via L-ensembles [Borodin and Rains 2005]. An L-ensemble defines a DPP through a positive semidefinite matrix L indexed by the elements of $\mathcal{S}$ as

$$\mathcal{P}_L(X = x) = \frac{\det(L_x)}{\det(L + I)}, \tag{2}$$

where I is the $N \times N$ identity matrix. This definition has several practical advantages over the one defined on marginal kernel. First, L-ensemble defines the atomic probabilities over every possible instantiation of X, rather than the marginal probabilities of inclusion as given by marginal kernel K. Second, $\mathcal{P}_L$ has closed-form normalization given by the identity $\sum_{x \subseteq \mathcal{S}} \det(L_x) = \det(L + I)$. Clearly, the summation over exponential-counting subsets is equal to a tractable determinant operator which can be exactly computed in polynomial time complexity. Refer to Kulesza and Taskar [2012] and Affandi et al. [2014] for the other polynomially tractable inferences, e.g., marginalization, conditioning, sampling, and finding the mode. Third, unlike marginal kernel K, the eigenvalues of the positive semidefinite kernel L are not required to be bounded above by one. Thus, it is more feasible for real-world dataset modelling. The relationship between the marginal DPP definition and L-ensemble construction is that a DPP defined with a marginal kernel K has an L-ensemble kernel $L = K(I - K)^{-1}$ (when the inverse exists), and an L-ensemble can be computed from a marginal kernel $K = L(I + L)^{-1}$.

A positive semidefinite kernel matrix L can be expressed as a Gram matrix [Kulesza and Taskar 2010]:

$$L_{ij} = q_i \phi_i^T \phi_j q_j, \tag{3}$$

where $q_i, q_j \in \mathbb{R}^+$ represent the qualities of elements i, j, and $\phi_i, \phi_j \in \mathbb{R}^n$, the unit length feature vectors, represent the similarity between elements i and j with $\phi_i^T \phi_j \in [-1, 1]$. With this decomposition, one can independently model quality and diversity of a subset at the same time with a unified model. It encourages a DPP to choose subsets with elements of high quality as well as dissimilar to each other.

3.2. Sequential Monte Carlo

SMC sampling techniques can be used to sample from a sequence of distributions $\pi_1(x_1), \ldots, \pi_t(x_t)$ for variables $X_{1:t} = \{x_1, x_2, \ldots, x_t\}$, where $\{1, \ldots, t\}$ are time indexes. It improves sequential important sampling with introducing auxiliary variables and artificial backward Markov kernels with density $\{L_k(x_{k+1}, x_k)\}_{k=1}^{t-1}$. It applies importance sampling (IS) technique [Del Moral et al. 2006; Wu et al. 2013; Kantas et al. 2009] between an artificial joint distribution $\widetilde{\pi}_t(X_{1:t}) = \widetilde{\gamma}(X_{1:t})/Z_t$ and a proposed joint importance distribution $\eta_t(X_{1:t}) = \eta_1(x_1)\prod_{k=2}^{t} K_k(x_{k-1}, x_k)$, where Z_t is a normalization constant, and $\widetilde{\gamma}_t(X_{1:t}) = \widetilde{\gamma}_t(x_t)\prod_{k=1}^{t-1} L_k(x_{k+1}, x_k)$, to collect samples. Note that both joint distributions – the artificial one and the proposed joint one – are decomposed in a sequential multiplying manner. The backward Markov kernels $\{L_k(x_{k+1}, x_k)\}_{k=1}^{t-1}$ in the artificial joint distribution play roles as backward conditional distributions $\widetilde{\gamma}(x_{k-1}|x_k)$. Similarly, forward Markov kernels $\{K_k(x_{k-1}, x_k)\}_{k=2}^{t}$ in the proposed joint importance distribution play roles as forward conditional distributions, i.e., $\eta(x_k|x_{k-1})$. The samples from the proposed importance distribution are usually biased from the "true" distribution. IS is applied to correct the discrepancy between them by weighting the samples with the matching degree to the "true" distribution. The likelihood ratios are usually applied as matching criteria, and are formulated as

$$\begin{aligned} w_t(X_{1:t}) &= \frac{\widetilde{\gamma}_t(X_{1:t})}{\eta_t(X_{1:t})} \\ &= w_{t-1}(X_{1:t-1})\widetilde{w}_t(x_{t-1}, x_t), \end{aligned} \tag{4}$$

where

$$\widetilde{w}_t(x_{t-1}, x_t) = \frac{\widetilde{\gamma}_t(x_t)L_{t-1}(x_t, x_{t-1})}{\widetilde{\gamma}_{t-1}(x_{t-1})K_t(x_{t-1}, x_t)}. \tag{5}$$

From Equation (4), it can be easily seen that the importance weights are calculated in a recursive form. At a given time stamp, the joint weight over the time period $1 \sim t$ – $w_t(X_{1:t})$ – is computed by multiplying two components: one is the joint weight accumulated to $t-1$, i.e., $w_{t-1}(X_{1:t-1})$, and the other is the incremental weight from the local pairwise joint distributions with consecutive time stamps, i.e. $\widetilde{w}_t(x_{t-1}, x_t)$. Clearly, at each time stamp, instead of computing the whole joint weight from the beginning, SMC updates it with local incremental weights. Thus, the overall computation is efficient. From the definition in Equation (5), the incremental weight relates to not only the marginal distribution, but also the backward and forward Markov kernels. How to choose these two kernels are a crucial step to achieve high approximation results. The marginal distributions $\{\pi_t(x_t)\}$ can be approximated by the sequentially sampled N particle–weight pairs $\{X_{1:t}^{(i)}, w_t^{(i)}\}_{i=1}^{N}$ [Ohsaka et al. 2014]:

$$\pi_t^N = \sum_{i=1}^{N} w_t^{(i)}\delta\big(x_t^{(i)}\big), \tag{6}$$

where δ is the Dirac Delta function.

4. TIME-VARYING DETERMINANTAL POINT PROCESSES

Sequential data is ubiquitous in real-world scenarios. We represent sequence data with variables $X_{1:t} = \{x_1, x_2, \ldots, x_t\}$, and their corresponding probability measures are denoted as $\pi_1(x_1), \ldots, \pi_t(x_t)$, where t is time indexes. We focus on sequentially sampling from separate distributions $\{\pi_i\}_{i=1}^{t}$, rather than from a joint distribution $\Pi(X_{1:t})$. This is suitable for some applications that require real-time results at each single time stamp. For example, a news provider needs to sample a diverse news subset from all its information sources in order to display what is currently happening to its clients. Since its clients change from time to time, only diverse subsets at single time stamps are essential. No temporal diversity is needed to be considered. Under this case we sequentially sample subsets by starting with π_1, then π_2, and so on.

A time-varying structure assumes the neighbouring ground datasets are largely overlapped. In other words, the difference between them is subtle. Consequently, the probability measures built on the ground datasets are almost the same. Such a property makes it feasible to exploit SMC sampling technique to design a fast sampling algorithm for time-varying samples.

A time-varying determinantal point process (TV-DPP) integrates DPPs into the time-varying setting, whose marginal distribution at each time stamp t obeys DPPs [Kulesza and Taskar 2012]. Formally,

$$\pi(X = x_t) = \frac{\det(L_{t,x_t})}{\det(L_t + I_t)}. \tag{7}$$

Here, I_t is the identity matrix of the same dimension with L_t. $x_t \subseteq \mathcal{S}_t$, where $\mathcal{S}_t$ is the ground dataset at time stamp t. L_{t,x_t} is a submatrix of an L-ensemble kernel matrix L_t, whose indexes are restricted to the elements in x_t. Typically, $L_t = X_t X_t^T$ is constructed from data features $X_t = (x_{t1}, \ldots, x_{tN_t})^T, x_{ti} \in R^D$. Due to the time-varying structure, the ground dataset X_t differs from X_{t+1} by only a few elements. As a result, L_t slightly differs from L_{t-1} by a few rows and columns, which implies $\pi_{t-1} \approx \pi_t$. TV-DPPs are given as an illustrative example in Figure 1(a), and their graphical representation is shown in Figure 1(b).

We apply the SMC framework introduced in Section 3.2 to achieve the sequential sampling task. We make use of a fast DPP sampler [Kang 2013] to collect samples at $t = 1$. We employ MCMC kernels proposed in Kang [2013] as the forward Markov kernels $\{K_t(x_{t-1}, x_t)\}$, where they are invariant to distributions π_t and make use of a standard Metropolis–Hasting algorithm. Regarding the unequal cardinality case, i.e., $|X_{t-1}| \neq |X_t|$, we can dynamically maintain the elements in samples during the Metropolis–Hasting algorithm: (1) when the cardinality increases, the selected elements to be included in a new sample should choose from the new, enlarged ground dataset; and (2) when the cardinality decreases, a new sample transited from the old one should first exclude the elements that do not appear in the new ground dataset. Based on the choice for forward Markov kernel, one good approximation for optimal backward Markov kernel L_{t-1} in Equation (5) is given by [Del Moral et al. 2006]

$$L_{t-1}(x_t, x_{t-1}) = \frac{\pi_t(x_{t-1}) K_t(x_{t-1}, x_t)}{\pi_t(x_t)}. \tag{8}$$

From the artificial joint distribution, one marginal distribution $\pi_t(x_t)$ can be computed as

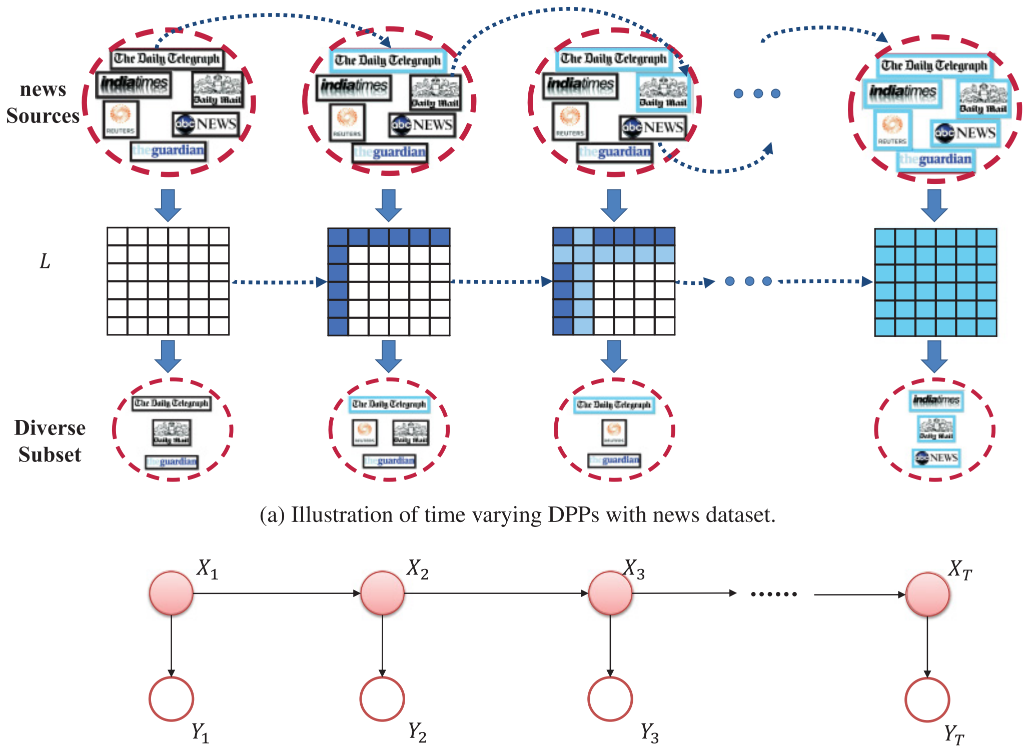
$$\pi_t(x_t) \propto \int \tilde{\gamma}_t(x_t) L_{t-1}(x_t, x_{t-1}) dx_{t-1}. \tag{9}$$

Finally, the un-normalized incremental weight is derived as follows via substituting Equations (8) and (9) into Equation (5):

$$\widetilde{w}_t(x_{t-1}, x_t) = \frac{\tilde{\gamma}_t(x_{t-1})}{\tilde{\gamma}_{t-1}(x_{t-1})}. \tag{10}$$

Clearly, the incremental weight relates to two successive marginal distributions π_{t-1} and π_t over the sampled particles $\{x_{t-1}^{(i)}\}_{i=1}^N$ at time stamp $t-1$. In conclusion, under the SMC framework (4), we sequentially move sampling particles forward with transitional kernel $K_t(x_{t-1}, x_t)$, and compute their weights at each time stamp t: $\{w_t^{(i)}\}_{i=1}^N$ with particle weights at time stamp $t-1$: $\{w_t^{(i)}\}_{i=1}^N$ as well as the incremental weights from time stamp $t-1$ to time stamp t: $\{\widetilde{w}_t(x_{t-1}^{(i)}, x_t^{(i)})\}_{i=1}^N$.

There are two time-consuming steps in the above sampling procedure. One is the particle moving-forward step, which actually samples from a marginal DPP



(a) Illustration of time varying DPPs with news dataset.

(b) Graphical representation for TV-DPPs.

Fig. 1. Time-varying DPPs. In the first diagram, the first row represents the news updating process along time stamps. Six different news sources are schematically listed, i.e., 'The Daily Telegraph', 'Daily Mail', 'ABC NEWS', 'The Guardian', 'Reuters', and 'Indiatimes'. From time to time, only a small portion of the news sources are updated. The arrows make which news sources are updating clear: it starts at a news source with old news and points to the same source with new headlines – bordered in cyan – at the next time stamp. The second row shows the evolution of DPP marginal kernel L along with the news updates. The difference between two successive Ls is highlighted with different colours and is apparently tiny. The third row shows explanatory diverse subsets outputted by TV-DPPs. In the second diagram, the solid circles represent the observations, which correspond to the news dataset shown in the first row of the above figure, and the hollow circles represent the variables obeying the DPP distribution, one example of which can be found in the third row of the above figure. One important truth is that given the observations $\{X_1, X_2, \ldots, X_T\}$, the variables $\{Y_1, Y_2, \ldots, Y_T\}$ are independent.

Distribution, and a fast sampling algorithm [Kang 2013] is applied as we have already introduced. The other step is responsible for updating the incremental weights, and it accomplishes this by computing the likelihood ratios between two consecutive DPP distributions. Due to the possibility of large-scale ground datasets, direct computing the likelihoods is impractical. We show how to make use of the proposed time-varying setting (slight changes between two successive distributions) to accelerate this computation. Specifying the general likelihoods with DPP distributions, the incremental particle weights for TV-DPPs are computed with the following determinantal ratio:

$$\begin{aligned}\widetilde{w}_t(x_{t-1}, x_t) &= \frac{\det(L_{t,x_{t-1}})/\det(L_t + I)}{\det(L_{t-1,x_{t-1}})/\det(L_{t-1} + I)} \\ &\propto \det(L_{t,x_{t-1}})/\det(L_{t-1,x_{t-1}}).\end{aligned} \tag{11}$$

Since the term $\det(L_{t-1}+I)/\det(L_t+I)$ is constant over all particles $\{x^{(i)}\}_{i=1}^N$, we simply ignore it and focus on the computation for un-normalized incremental weights.

From time-varying structure, the difference between L_t and L_{t-1} has been small. Therefore, many elements of the set $\{\det(L_{t,x_{t-1}^{(i)}})/\det(L_{t-1,x_{t-1}^{(i)}})\}_{i=1}^N$ are equal to 1, which is a significant speed-up factor in our work. For the rest elements of $\{\det(L_{t,x_{t-1}^{(i)}})/\det(L_{t-1,x_{t-1}^{(i)}})\}$, whose values are not equal to 1, we compute them as follows. We drop particle indexes (i) for clarity. Let L^{cc} denote the shared submatrix between $L_{t,x_{t-1}}$ and $L_{t-1,x_{t-1}}$. Then, $L_{t,x_{t-1}}$ can be decomposed as $[\begin{smallmatrix} L^{cc} & L^{ct} \\ L^{tc} & L^{tt} \end{smallmatrix}]$, where L^{ct}, L^{tc}, and L^{tt} are the rest submatrices for $L_{t,x_{t-1}}$. Similarly, we decompose $L_{t-1,x_{t-1}}$ with symbols L^{cc}, $L^{c,t-1}$, $L^{t-1,c}$, $L^{t-1,t-1}$. Let $k_1 = |L^{cc}|$ and $k_2 = |L^{tt}|$ (or $k_2 = |L^{t-1,t-1}|$) be the cardinalities for shared submatrix and dissimilitude respectively, and $k_1 \gg k_2$, based on the time-varying structure. Then, the determinant ratio in Equation (11) can be computed efficiently by applying the determinant formula of partitioned block matrices which is multiplicative in the Schur complement, without computing the nominator and denominator explicitly. By cancelling the shared $\det(L^{cc})$, the weight can be computed as follows:

$$\widetilde{w}_t(x_{t-1}, x_t) \propto \frac{\det(L^{tt} - L^{tc}(L^{cc})^{-1}L^{ct})}{\det(L^{t-1,t-1} - L^{t-1,c}(L^{cc})^{-1}L^{c,t-1})}, \tag{12}$$

where L^{tt}, L^{cc}, L^{tc}, $L^{t-1,t-1}$, $L^{t-1,c}$ are submatrices of $L_{t,x_{t-1}}$ or $L_{t-1,x_{t-1}}$. It is easily observed that the complexity has reduced from $\mathcal{O}((k_1+k_2)^3)$ to k_2^3. However, when the changes occurring between time intervals $t-1$ and t become large, namely, k_2 growing up, the complexity will increase at an exponential rate. The worst situation for our setting is an extreme case where two contiguous datasets are completely non-overlapping, and then our model degenerates to the case of individually sampling DPP.

As shown in the above formula, the computation of incremental weight relates to matrix addition and multiplication, as well as the inverse for the shared submatrix $(L^{cc})^{-1}$. Based on our assumption – neighbouring shared submatrix (e.g., L^{cc}_{t-1} and L^{cc}_t) is slightly different with several elements – we are able to efficiently update the inverse of shared submatrix by repeatedly applying both matrix block inverse formula and matrix inverse lemma. We elaborate this update step by step. Suppose the shared submatrix of L^{cc}_{t-1} and L^{cc}_t is L^{comm}, and the unshared ones are A, A^T, B. There are two cases for updating the inverse from L^{cc}_{t-1} to L^{cc}_t. One case is when elements are added, i.e., $L^{comm} = L^{cc}_{t-1}$. Therefore, $L^{cc}_t = [\begin{smallmatrix} L^{comm} & A \\ A^T & B \end{smallmatrix}]$. Since $(L^{comm})^{-1}$ is currently known, the inverse of L^{cc}_t can be expanded by first applying matrix block inverse formula [Golub and Van Loan 2012]

$$\begin{aligned}(L^{cc}_t)^{-1} &= \begin{bmatrix} L^{comm} & A \\ A^T & B \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (L^{comm} - AB^{-1}A^T)^{-1} & -(L^{comm})^{-1}A(B - A^T(L^{comm})^{-1}A)^{-1} \\ -B^{-1}A^T(L^{comm} - AB^{-1}A^T)^{-1} & (B - A^T(L^{comm})^{-1}A)^{-1} \end{bmatrix}.\end{aligned} \tag{13}$$

It is easily observed that most inverses are either known (i.e., $(L^{comm})^{-1}$) or easily to be computed due to its small sizes (e.g., $(B - A^T(L^{comm})^{-1}A)^{-1}$), except the inverse of a correction form of L^{comm}. Here, matrix inverse lemma is applied to convert computing the inverse of a correction form to computing the correction of the original matrix, namely

$$\begin{aligned}&(L^{comm} - AB^{-1}A^T)^{-1} \\ &\quad = (L^{comm})^{-1} + (L^{comm})^{-1}A(B - A^T(L^{comm})^{-1}A)^{-1}A^T(L^{comm})^{-1}.\end{aligned} \tag{14}$$

ALGORITHM 1: Fast sampling for TV-DPPs

```
Data: $\{L_i\}_{i=1}^t$
Result: $\{\{x_k^{(i)}, W_k^{(i)}\}_{i=1}^N\}_{k=1}^t$
$\{x_1^{(i)} \sim \det(L_{1,x_1^{(i)}})\}_{i=1}^N$ by Kang [2013] or Kulesza and Taskar [2012] ;
$\{W_1^{(i)} = 1/N\}_{i=1}^N$;
$k = 1$;
repeat
    k = k+1 ;
    $\{\widetilde{w}_k(x_{k-1}^{(i)}, x_k^{(i)}) = \det(L_{k,x_{k-1}^{(i)}})/\det(L_{k-1,x_{k-1}^{(i)}})\}_{i=1}^N$;
    $\{W_k^{(i)} = W_{k-1}^{(i)}\widetilde{w}_k^{(i)} / \sum_{i=1}^N (W_{k-1}^{(i)}\widetilde{w}_k^{(i)})\}_{i=1}^N$ ;
    $\{x_k^{(i)} \sim K(x_{k-1}^{(i)}, x_k^{(i)})\}_{i=1}^N$;
    $N_{k,ESS} = \{\sum_{i=1}^N (W_k^{(i)})^2\}^{-1}$ [Sahlin 2011];
    if $N_{k,ESS} < \alpha \cdot N$ then
        $\{x_k^{(i)} \sim Multi(W_k^{(1)}, \ldots, W_k^{(N)})\}_{i=1}^N$;
        $\{W_k^{(i)} = 1/N\}_{i=1}^N$ ;
        move $\{x_k^{(i)}\}_{i=1}^N$ by $\pi_k$ invariant MCMC kernel $K_{\pi_k}(x_k^{(i)}, \cdot)$ [Kang 2013];
    end
until $k = t$;
```

The other case is when elements are deleted, i.e., $L^{comm} = L_t^{cc}$, and $L_{t-1}^{cc} = [\begin{smallmatrix} L^{comm} & A \\ A^T & B \end{smallmatrix}]$. Since $(L_{t-1}^{cc})^{-1}$ is known, we set $(L_{t-1}^{cc})^{-1} \equiv [\begin{smallmatrix} E & F \\ F^T & G \end{smallmatrix}]$. Equally,

$$\begin{bmatrix} L^{comm} & A \\ A^T & B \end{bmatrix} = \begin{bmatrix} E & F \\ F^T & G \end{bmatrix}^{-1}. \tag{15}$$

Again, applying matrix block inverse formula to the left hand of the above equation, we obtain $(E - FG^{-1}F^T)^{-1} = L_t^{cc}$, and therefore,

$$\left(L_t^{cc}\right)^{-1} = (E - FG^{-1}F^T). \tag{16}$$

Note: Usually the variance of the incremental weight has an increasing tendency along the timeline, resulting in a potential degeneracy of the particle approximation. Routinely, the degeneracy is measured by the ESS criterion [Sahlin 2011]. When the ESS is smaller than a predefined threshold $\alpha \cdot N$ (α is a predefined ratio), we re-sample the particles with a multinomial distributions parameterized by the normalized particle weights. To make the resampled particles more diverse, we further randomly move the equally weighted particles with an MCMC kernel of stationary distribution π_t [Gilks and Berzuini 2001]. The whole process of SMC is illustrated in Figure 2.

To sum up, we start the SMC process by initializing particles $\{x_1^{(i)}\}$ of marginal distribution π_1. We sample $\{x_1^{(i)} \sim \pi_1(X = x_1^{(i)})\}_{i=1}^N$ using a fast DPPs sampling algorithm proposed in Kang [2013]. Then, at each time t, we update these samples from $\{x_{t-1}^{(i)}\}_{i=1}^N$ using the above customized SMC sampling scheme. The proposed fast sampling algorithm for time-varying DPPs (TV-DPPs) is shown in Algorithm 1.[1,2]

[1] We use $W_k^{(i)}$ and $\widetilde{w}_k^{(i)}$ to replace $W_k(x_{1:k}^{(i)})$ and $\widetilde{w}_k(x_{k-1}^{(i)}, x_k^{(i)})$ for simplicity.

[2] $Multi(W_k^{(1)}, \ldots, W_k^{(N)})$ stands for the multinomial distribution parameterized by $\{W_k^{(i)}\}_{i=1}^N$.

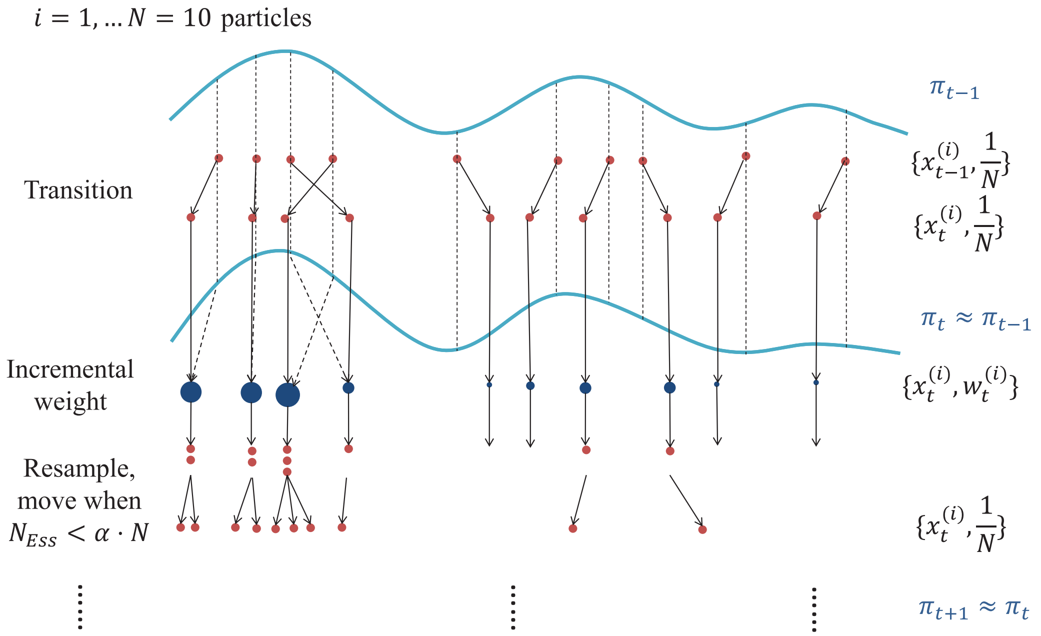

$i = 1, \ldots N = 10$ particles
Transition
Incremental weight
Resample, move when $N_{Ess} < \alpha \cdot N$
π_{t-1}
$\{x_{t-1}^{(i)}, \frac{1}{N}\}$
$\{x_t^{(i)}, \frac{1}{N}\}$
$\pi_t \approx \pi_{t-1}$
$\{x_t^{(i)}, w_t^{(i)}\}$
$\{x_t^{(i)}, \frac{1}{N}\}$
$\pi_{t+1} \approx \pi_t$

Fig. 2. Illustration of Sequential Monte Carlo. At time stamp $t - 1$, 10 particles in red with equal weights are given, i.e., $\{x_{t-1}^{(i)}\}_{i=1}^{10}$. At this stage, two computations will be done – one is computing the incremental weights; the other is computing the particles for the next time stamp. For the incremental weight of each particle at time $t - 1$, according to Equation (10), it is simply the likelihood ratio between time stamps t and $t - 1$. The corresponding relationship is denoted by the dashed line connecting two neighbour distributions and the weight for each particle is illustrated by size of blue solid circle. For the particle's location at next time stamp t, usually, a Markov transition kernel is used to qualify the transition job between two slightly different neighbour distributions. The transition relationship is indicated by solid line with arrow. For the particle's weight at time stamp t, it is gained by multiplying the weight at time stamp $t-1$ by the incremental weight. To alleviate the degeneracy of the algorithm which is measured by effective sample size (ESS), a resampling step is applied when $N_{ESS} < \alpha \cdot N$. High weighted particles will re-birth as several equal weighted particles, while particles with low weights may disappear. To increase the samples' diversity, a move step is followed. Once particle' locations and weights at t are prepared, it will recursively carry out the whole above procedure.

5. EXPERIMENTAL RESULTS

In this section, we report experimental results of our proposed fast sampling for TV-DPPs when applied to a real-world news dataset and the Enron Corpus [Diesner and Carley 2005].

5.1. News Recommendation

We collect the news dataset from six news websites with different topics as resources during the period of $12:16$am 2 Jun, 2014–$8:13$am 5 Jun, 2014. The websites and topics are summarized in Table I.[3] The topic titles are directly named from the categories of the news website.

There are 33 topics in each time stamp. At the beginning $t = 1$, we collect each resource's latest news corpora $\mathcal{S}_{t=1} = \{d_{t=1,k}\}_{k=1}^{33}$. At next time stamp, n topics are updated. We recorded the latest news corpora for these n topics as well as the news for unchanged topics as $\mathcal{S}_{t=2} = \{d_{t=2,k}\}_{k=1}^{33}$. In this manner, we record the sequential

[3] A—ABC, D1—Dailymail, D2—Dailytelegraph, G—Guardian, I—Indiantimes, and R—Reuters.

Table I. Summary of News Categories for Different News Media Sources

	A	D1	D2	G	I	R
sport	√	√	√	√	√	
news	√	√	√		√	
politics	√			√		√
business	√	√	√		√	√
sciencetech		√				
femail		√				
tvshowbiz		√				
technology	√		√		√	√
global				√		
world				√		√
entertainment					√	
markets						√
others	√	√	√		√	

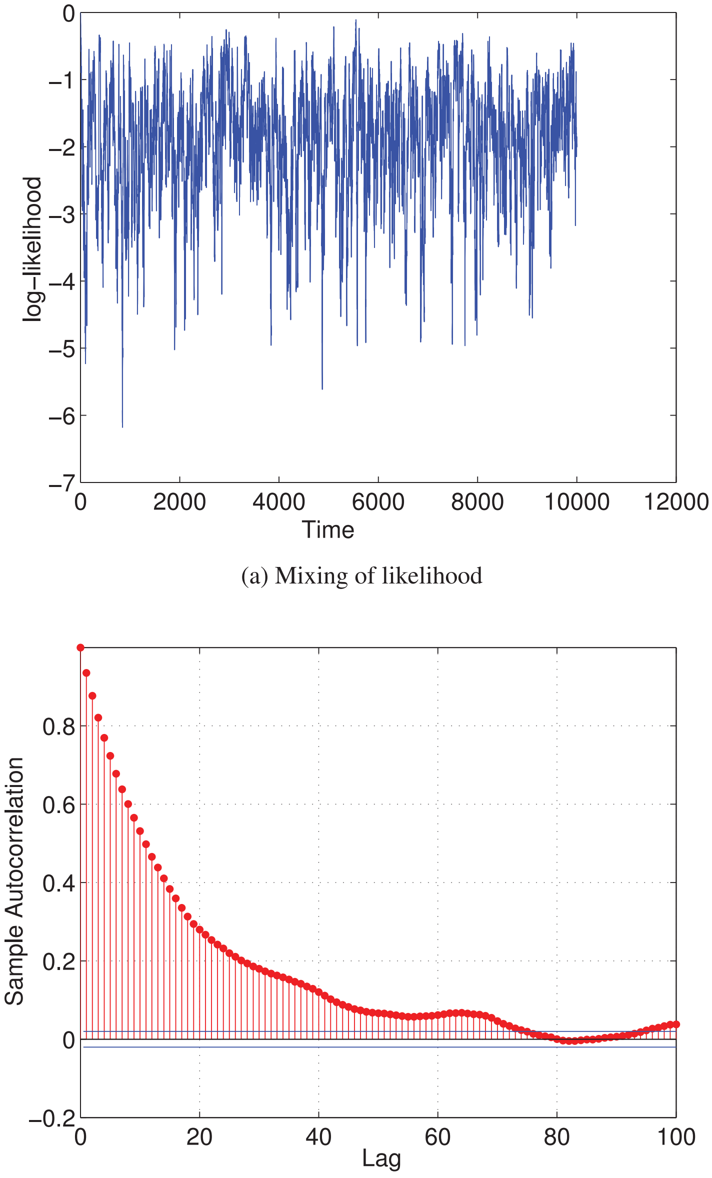
news corpora $\mathcal{S}_{t:T} = \{d_{t,k}\}_{k=1}^{33},_{t=3}^{T}$. Herein, to guarantee the high similarity between two distributions formed by successive datasets, we choose $n = 5$ to verify our algorithm. Regarding the length of the news dataset sequence, $T = 1000$ is selected for experimental demonstration.

We extract normalized Term Frequency-Inverse Document Frequency (TF-IDF) feature vectors [Wu and Luk. 2008; Xuan et al. 2015a] to represent news articles. All news corpora $\mathcal{S}_{1:T}$ are employed to compute the IDF. We apply cosine similarity [Gillenwater et al. 2012a] to construct kernel matrix L. $L(d_i, d_j) = \text{cos-sim}(d_i, d_j)$ and $\text{cos-sim}(d_i, d_j)$ is defined as

$$\frac{\sum_{w \in W} \text{tf}_{d_i}(w)\,\text{tf}_{d_j}(w)\,\text{idf}^2(w))}{\sqrt[2]{\sum_{w \in W} \text{tf}^2_{d_i}(w)\,\text{idf}^2(w)}\sqrt[2]{\sum_{w \in W} \text{tf}^2_{d_j}(w)\,\text{idf}^2(w)}}, \tag{17}$$

where W is a subset of the words found in the two documents. The TF-IDF is usually sparse; thus, the distances amongst TF-IDFs are quite small and will lead to poor diversity. We re-feature each news article referred as $\tilde{d}_i$ with binary vector where the j-th coordinate of d_i is 1 if news article j is amongst the N_{nei} nearest neighbours of news article i according to the cosine similarities, and 0 otherwise. The L-ensemble kernel matrix is computed by $L(i, j) = e^{-\alpha \cdot \text{cos-sim}(\tilde{d}_i, \tilde{d}_j)}$. We fix $\alpha = 1$ throughout the experiment. Regarding N_{nei}, we tried a wide rough range of values [100, 200, . . . , 1000, and 700 is chosen as it achieved the best results.

5.1.1. Sampling Analysis at $t = 1$. We apply the MCMC method – a fast DPPs sampling algorithm [Kang 2013] – to initialize our particles. (The sampling algorithm in Kulesza and Taskar [2012] can also be applied, in which the low-rank assumption of the kernel matrix at time stamp 1 will also improve the initialization efficiency.) The mixing of likelihood of the sampled subsets is shown in Figure 3(a). In our experiment, the burn-in period is used to wait for the mixing of the Markov chain. At the beginning, the starting subset is smaller (such as containing one single element) than the subset when it is mixed. In this case, the likelihood of the starting subset is usually high. In the following several iterations, it is highly possible that more elements will be added into the starting subset, which will lead to lower likelihood. This phenomenon leads to decreasing tendencies of sequential likelihoods, which matches the beginning curve in Figure 3(a). Afterwards, when it approaches to mixing, the Markov chain will collect subsets by alternative operations of adding and deleting elements and the sequential likelihood is expected in oscillation tendency, as shown in Figure 3(a). We set the burn-in period as 10,000 in our experiment.



(a) Mixing of likelihood

(b) ACF

Fig. 3. Analysis for fast DPPs sampling algorithm at $t = 1$.

The consecutive samples are not independent due to the Markov property. We compute the AutoCorrelation Function (ACF) parameterized by lag k to determine the interval between two independent samples x_t, x_{t+k}. We apply indirect measure Q to define the autocorrelation coefficient [Sandvik 2013] at lag k

$$r_Q(k) = (< Q_t Q_{t+k} > - < Q_t >^2)/\big(< Q_t^2 > - < Q_t >^2\big). \tag{18}$$

And, in our experiment, we set the measurement Q_t for sample x_t (sampled at time stamp t) as its diversity likelihood, namely

$$Q_t = \det(x_t). \tag{19}$$

The plot of the ACF is shown in Figure 3(b). As expected, the sample autocorrelation coefficient is decreasing with the increasing of lag k. When the $r_Q(k_{ACF})$ is below τ, we claim that the sampled subset x_t is independent with subset $x_{t+k_{ACF}}$. Here, we set $\tau = 0.02$ and $k_{ACF} = 80$. We collect $N = 100$ independent samples by extracting one sample subset for every k_{ACF} iteration as initial particles for TV-DPPs sampling.

In the next section, we analyse the performance of our proposed fast TV-DPPs sampling with both qualitative and quantitative demonstrations.

5.1.2. Quantitative Analysis. We analyse the diversity and time complexity of our algorithm (SMC-DPPs) by comparing to the baseline algorithm, namely, separate fast DPPs (sep-DPPs) sampling algorithm at each time stamp.

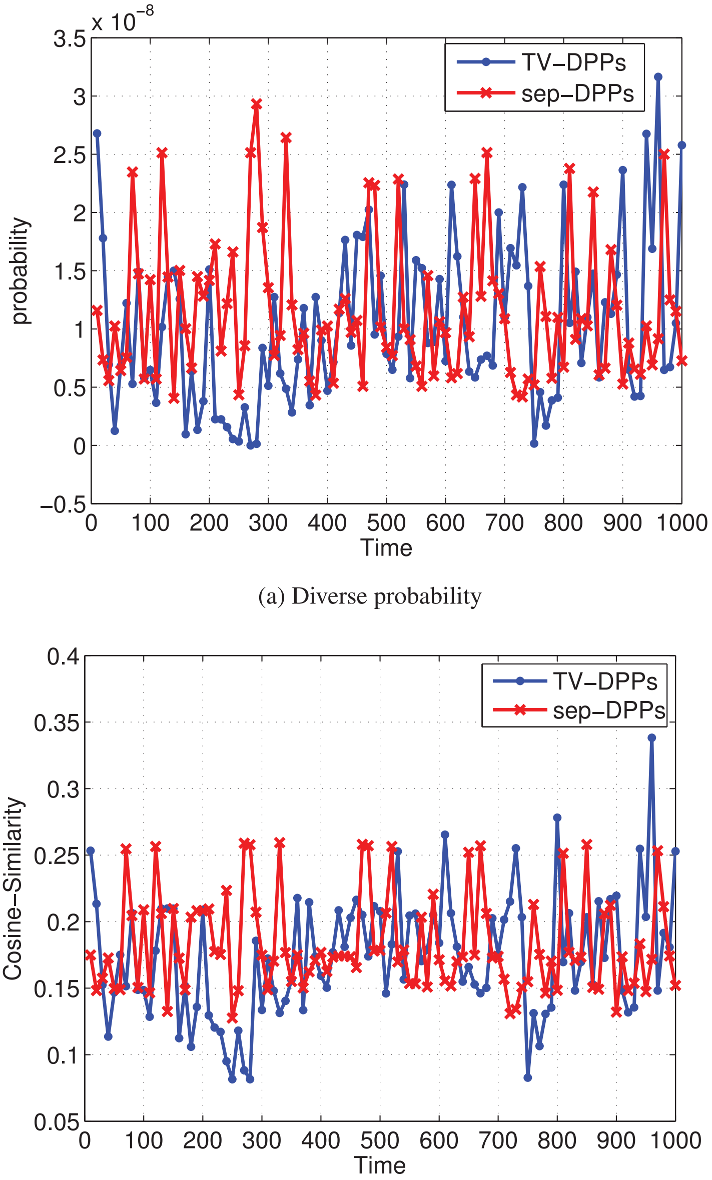
Accuracy Analysis: We compare TV-DPPs and the baseline sep-DPPs with regards to diverse probabilities and cosine similarities of subsets that approximate the maximal mode of DPP distributions. At each given time stamp, for TV-DPPs, we measure each particle with ground-truth DPP distribution, and find the particle with the maximal diverse probabilities. We do the same selection for sep-DPPs' samples. The comparison result is shown in Figure 4(a). For clear visualization, here we illustrate 100 time stamps by taking one every 10 time stamps. The curves declare that our speedup TV-DPPs sampling algorithm is comparable to sep-DPPs in terms of diverse accuracy.

To further demonstrate the accuracy of our algorithm, we also compute the cosine similarities of news articles in the subsets owning the maximal diverse probabilities at each given time stamp. The cosine similarities of pairwise news articles are computed with TF-IDF feature vectors. The results for both TV-DPPs and sep-DPPs are plotted in Figure 4(b). Clearly, our TV-DPPs algorithm performs comparably. These results illustrate the effectiveness of our algorithm to some extent.

Time Complexity Analysis: We also compute the time complexity of TV-DPPs and sep-DPPs. We directly apply the mixing time in Kang [2013] for burn-in time of every fast DPPs sampling. For sep-DPPs, N independent samples of diverse subsets are sampled by collecting one sample per k_{ACF} steps at one given time stamp. Then, the total time cost by the T-length sequence is: $T \times [|S| \log(|S|/\epsilon) + N \times k_{ACF}]$, where $S = 33$ in our case. For TV-DPPs, at the beginning $t = 1$, one sep-DPPs procedure is applied and N independent particles are prepared. After that, each particle costs constant time to move to the next time stamp. The total time complexity is summarized as $|S| \log(|S|/\epsilon) + N \times k_{ACF} + N \times (T - 1)$. Whether the algorithm is less time-consuming depends on $|S|$, N, k_{ACF}, as well as T. The time costs in log-space for comparison of sep-DPPs and TV-DPPs in our experiment are plotted in Figure 5.

At the beginning point, TV-DPPs (in blue) costs the same time as sep-DPPs (in red). With the t increasing, TV-DPPs algorithm performs much more efficiently than the sep-DPPs scheme.

5.1.3. Qualitative Analysis. To give an intuitive sense of diversity of sampled news article subset from the TV-DPPs samples, one news article subset randomly selected at $t = 12$ is visualized in Figure 6. Since news articles with high-dimensional TF-IDF features are difficult to visualize, here PCA was applied to reduce dimensions. There are 16 articles in total; thus, the maximal dimension after PCA is 15. A parallel coordinates plot with 15 coordinates is displayed in Figure 6(a). Different coloured curves represent different news articles, and their corresponding news titles are listed in Figure 6(b). There are two strong pieces of information which strongly suggest the diversity property of the selected news article subset. First, any two different coloured curves are not



(a) Diverse probability

(b) Cosine similarity

Fig. 4. Diversity comparison of news subsets selected by sep-DPPs and TV-DPPs with regard to both diverse probability and cosine similarity.

overlapped, as clearly seen in Figure 6(a). Second, there are no two news titles listed in Figure 6(b) textually the same. Obviously, this result qualitatively demonstrates the effectiveness of TV-DPPs.

We also plot the sequential sampling results from both TV-DPPs and sep-DPPs to give a holistic analysis in Figure 7. The X-axis represents time stamps, while the Y-axis denotes the indexes of news articles. As a particle filtering alike algorithm, TV-DPPs algorithm gives smoother samples of news articles along the timeline (shown

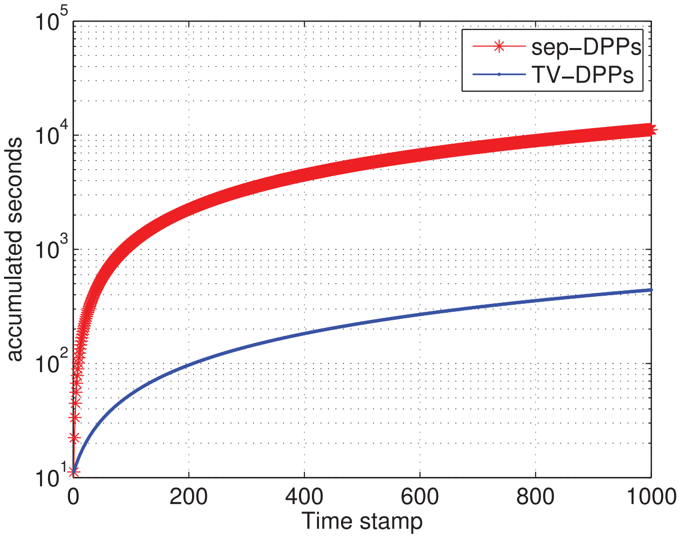


Fig. 5. Time cost comparison between sep-DPPs and TV-DPPs for news recommendation. The X-axis indicates the time stamps, while the Y-axis shows the accumulated seconds over time.

in Figure 7(a)) than samples of the independent selection of sep-DPPs scheme (shown in Figure 7(b)). Thus, it can be concluded that the TV-DPPs algorithm gives more consistent subsets along time stamps.

5.2. Enron Corpus

The Enron email corpus[4] contains a large set of email messages amongst the employees of the Enron corporation, and appeals to many researchers (The titles and indexes of its employees are also publicly accessible[5,6]). Various aspects of the Enron corpus have been investigated, such as natural language processing[7] [Klimt and Yang 2004], email classification,[8] quantitative analysis of social networks generated from the corpus [Zhong et al. 2014; Mcauley and Leskovec 2014; Iwata et al. 2012; Shetty and Adibi 2004], and so on. Here, we focus on discovering the dynamic event developments from its communication network.

The communication network is extracted as nodes and edges, where the nodes represent the employees in Enron and the edges represent undirected correspondences from senders to recipients. There are 150 nodes in total. Note that to avoid asymmetric distance between two correspondents, we simply drop the direction of each edge, and we also drop the communication frequencies between the correspondents. Email correspondences exclusively to or from addresses outside the Enron or Andersen domains are removed from our network. As introduced in Diesner et al. [2005], the number of emails and people involved in email communication between May 1999 and March 2002 was relatively high; therefore, we selectively focus on the period from August 1, 2001 to December 1, 2001. At each day, a snapshot of the communication network with fixed 150 nodes and edges is extracted from the correspondences between the

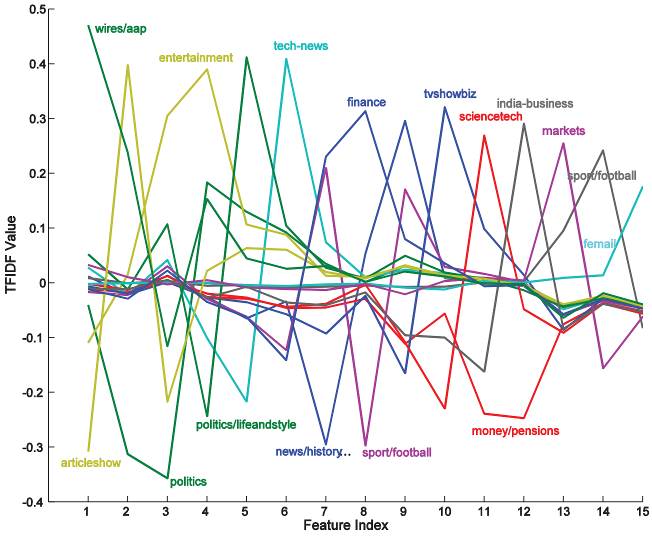
[4] http://www.enron-mail.com/email/.

[5] http://enrondata.org/assets/edo_enron-custodians-data.html.

[6] http://enrondata.org/assets/edo_enron-custodians.txt.

[7] http:www.ceas.cc/papers-2004/168.pdf.

[8] http://www.cs.umass.edu/∼ronb/.

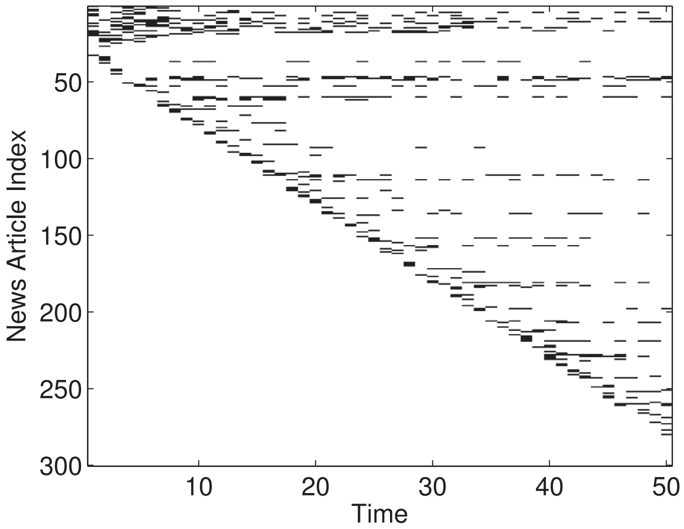


(a) A parallel coordinates plot of a set containing sixteen news articles is displayed, with X-axis representing the feature indices and Y-axis the coordinate values. Each curve represents one news article, and the non-overlapping phenomenon validates the diversity of the news article set.

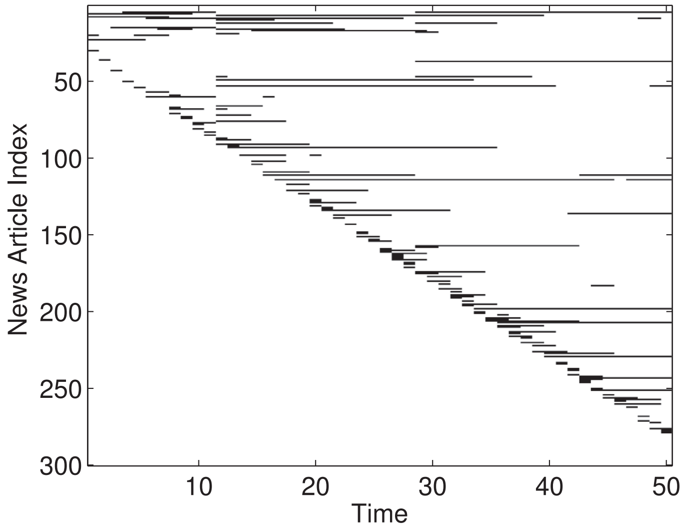
1. **news/history**: Tiananmen Square anniversary: Questions and Answers
2. **politics**: Senator plans bill to offer military veterans option on medical care
3. **money/pensions**: General Boss: Ban insurers as pensions advisers...
4. **tech-news**: Infosys investors lose faith in Narayana Murthy 2.0
5. **sport/football**: Rickie Lambert says his mum and dad cried at his return ...
6. **entertainment**: Shah Rukh's kids watch us: JusReign and Rupan Bal
7. **sport/football**: Graham Wallace to be grilled by North American-based ...
8. **tvshowbiz**: No wonder she's smiling! Michael Clarke's wife Kyly beams ...
9. **politics/lifeandstyle**: We should act against fast-food producers, not fat people
10. **sciencetech**: Look out gardeners! It's boom time for bugs that want to eat ...
11. **femail**: Girl who grew up with giants: Stalin, Chaplin, Lawrence of Arabia...
12. **markets**: JGBs slip as Japan capex data helps boost stocks
13. **articleshow**: Johnson's bouncer most pivotal of three turning points
14. **india-business**: Govt wants banks to fund local M&As
15. **finance**: Government should boost SMEs with NI breaks, says Lord Bilimoria
16. **wires/aap**: Crows expect injured Jacobs to face Freo

(b) Sixteen news titles with topic tags are listed. Each news title corresponds to one curve in the above plot with the same colour. This textual information further confirms the diversity of the news article set.

Fig. 6. Demonstration of diverse subset of news articles sampled by TV-DPPs.



(a) Sequential diverse subsets of news dataset sampled by sep-DPPs.



(b) Sequential diverse subsets of news dataset sampled by the proposed TV-DPPs.

Fig. 7. Demonstration of news topic drift. Comparing sequential subsets from TV-DPPs with the subsets from sep-DPPs, it extracts a more smooth news evolvement.

150 employees. Rather than from one single day's correspondence, here the edges are extracted from the past accumulating 30 days. It will not only artificially create overlaps between subsequent time stamps, but also stabilize the communication network at each time stamp. This will facilitate the detection of dynamic events happening in the Enron incorporation along the time line.

One example of the extracted communication network is shown in Figure 8. It is drawn by the NetDraw software provided by Borgatti [2002]. The two columns of

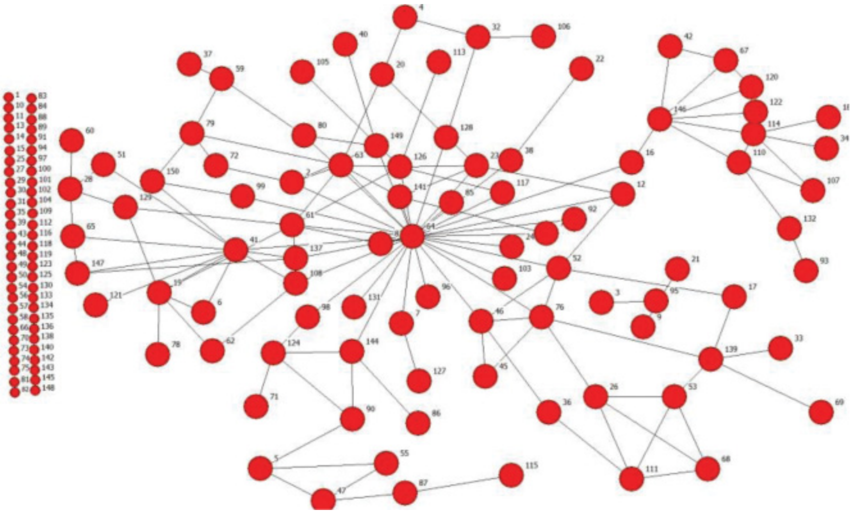


Fig. 8. One example of Enron communication network.

separate points demonstrate the employees who did not send or receive any email messages to or from other Enron employees during a period of around one month. The connected points show a hierarchical structure, i.e., the peripheral points tend to communicate with second-peripheral points, while the points at the centre of the communication network mutually connects each other and tend to have higher communication degrees.

We prepare the L-ensemble kernel matrix for DPP at each time stamp by a Gram matrix construction. First, we calculate shortest path distances for every pairwise nodes based on the network structure. Then, the distance measurement is transformed into similarity through an exponential operator. Finally, we introduce degree of each node, which is the number of edges incident to the node, as a qualification term. The degree exhibits how many employees this node is communicating with, which naturally qualifies the importance of the representing employee in Enron incorporation. Formally, the entry of L_{ij} in L-ensemble kernel matrix is computed by

$$L_{ij} = d_i \times \exp\{-\lambda \cdot dist(i, j)\} \times d_j, \tag{20}$$

where d_i represents the degree of node i, $dist(i, j)$ is the shortest path distance between the two nodes i and j, and λ is the parameter to adjust the relative similarity.

5.2.1. Quantitative Analysis. Because the sampling analysis at $t = 1$ is quite similar to the analysis for news recommendation and is not the main concern of this paper, we simply skip this step and directly proceed to analyse the efficiency of the proposed TV-DPPs. Like the experimental setting for news recommendation, we compare the proposed model with baseline sep-DPPs, which samples subset at each time stamp by performing separate DPP sampling algorithms. We sequentially sample subsets for a period of one year starting from January 1, 2011 and the demonstrating chart is shown in Figure 9.

At the beginning $t = 1$, the two different schemes take exactly the same time to do diverse subset sampling, since sep-DPPs apply the fast MCMC sampling algorithm to do sampling for each time stamp while TV-DPPs apply the same fast MCMC sampling algorithm to initialize particles for the subsequent SMC sampling scheme. After that, sep-DPPs scheme repeats the same sampling algorithm which always cost around 57s for a 150 nodes communication network. But, for the proposed TV-DPPs scheme, it

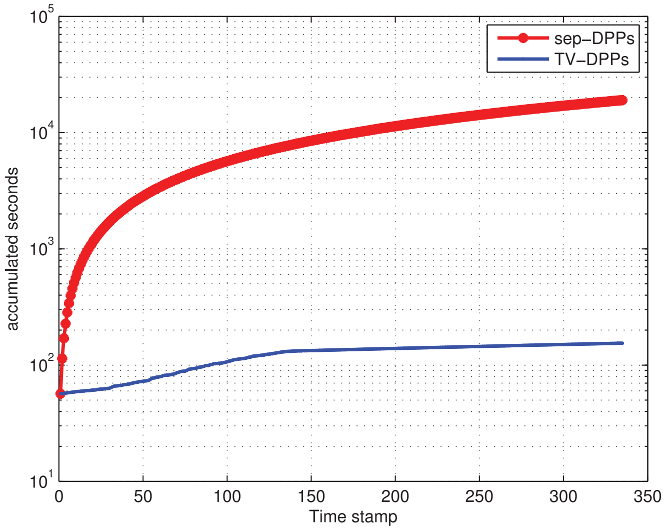


Fig. 9. Time cost comparison between sep-DPPs and TV-DPPs for Enron communication network. The X-axis represents the day stamps, while the Y-axis shows the accumulated seconds.

costs less than 0.2s to get a subsequent diverse subset. In a long-time run, our proposed TV-DPPs scheme significantly improves the efficiency of time-varying diverse subset sampling task. The sampling results are explained in the next section.

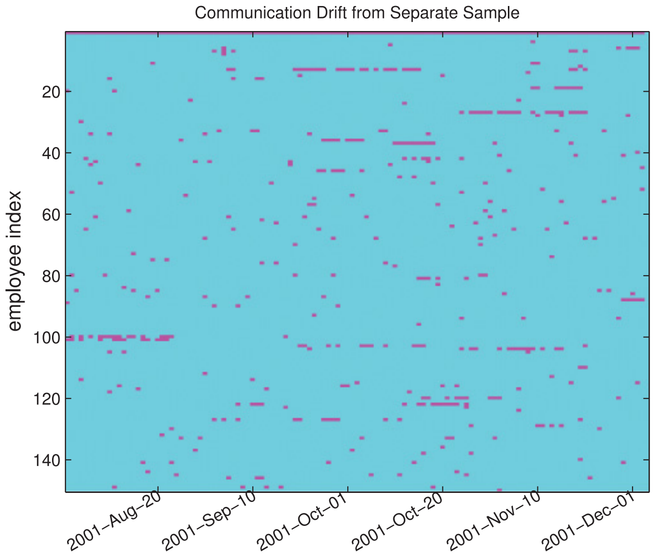
5.2.2. Qualitative Analysis. The sequential sampling results of both sep-DPPs and TV-DPPs along the timeline from August 1, 2001 to December 1, 2001 are plotted in Figure 10.

From Figure 10(a), at each day, a diverse subset out of a 150 employees is coloured in purple. For the holistic time period, it can be seen that points are randomly scattered and it is hard to discover any events happening within the company. Comparatively, from Figure 10(b), it can be easily seen that three clearly separated stages are obtained with smooth diverse subsets in each stage. Two separating points – one around 01 October, 2001 and the other around 10 November, 2001 – are easily noticeable. Actually, these three stages coincide with different important events happening in Enron incorporation, and they are Diesner et al. [2005]: (1) Before October 2001, the communication topic is not indicated in related references. (2) In October 2001, Andersen criminally instructed Enron to destroy any documentation related to the circumstance. (3) In December 2001, Enron became insolvent and was forced to file for bankruptcy.

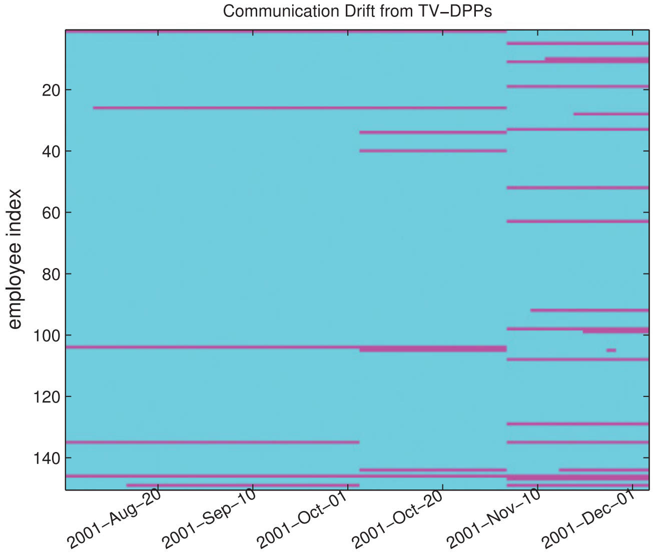
Although the TV-DPPs scheme is supposed to smoothly sample diverse subsets along time stamps, the resampling step in the time-varying DPP sampling procedure is designed to suddenly change the samples of diverse subset when these samples no longer correctly represent the present DPP distribution. This is the reason why three clearly different stages appear in the sampling results of the proposed TV-DPPs.

We detail each stage with a concrete time stamp of diverse sample shown in Figure 11(a)–(d), i.e., September 10, 2001 for the first stage, October 20, 2001 for the second stage, and November 20, 2001 for the third stage. In each sub-figure, there are 150 nodes in total, and the subset highlighted by bigger cyan circles is one diverse subset sample outputted from the proposed algorithm.

From the three examples of diverse subset, we claim that the DPP model at each time stamp prefers to select 'leader' points of peripheral branch, rather than points owning highest degrees. This is due to the combination of points' degree measurement



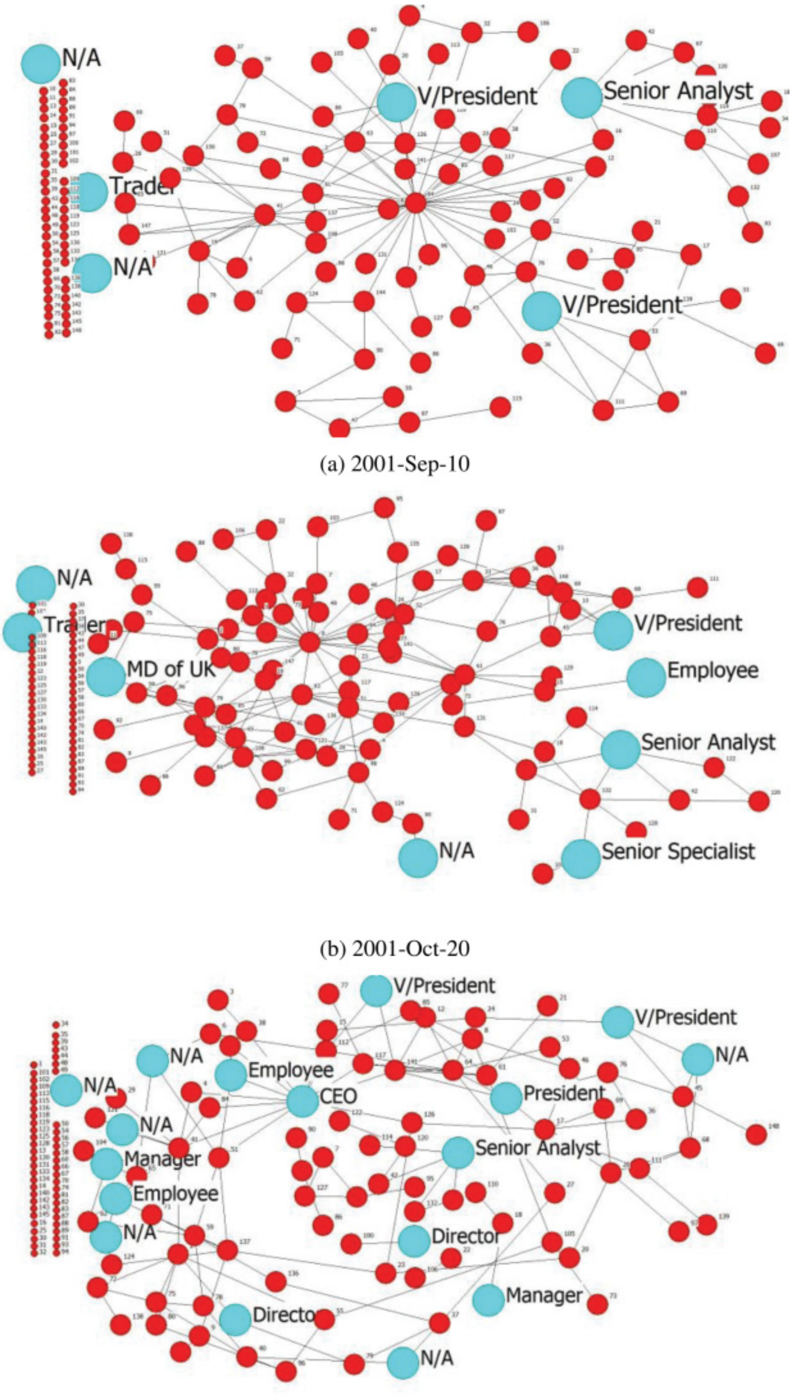
(a) Sequential diverse subsets of Enrom employees sampled by sep-DPPs.



(b) Sequential diverse subsets of Enron employees sampled by the proposed TV-DPPs.

Fig. 10. Demonstration of Enron communication drift. No communication patterns or events can be detected from the above figure. However, three different stages are clearly obtained with smoothness at each stage. Two separating points – one around 01 October, 2001 and the other around 10 November, 2001 – coincide with two important turn points for Enron incorporation.

and pairwise shortest distance measurement for the construction of the kernel matrix for sequential DPPs. In Figure 11(a), the communication amongst Enron employees is not that active, and a diverse subset containing only six nodes is sampled and the related titles are N/A, N/A, Trader, Vice President, Vice President, Senior Analyst



(a) 2001-Sep-10

(b) 2001-Oct-20

(c) 2001-Nov-20

Fig. 11. Three Enron communication networks from different stages detected by TV-DPPs.

respectively. In Figure 11(b), maybe due to Andersen's criminal command, the email communication amongst Enron employees was becoming more active during October in 2001, and a larger diverse subset is sampled by the proposed scheme. The position titles for the sampled nodes are N/A, Trader, Manager Director of UK, Vice President, Employee, Senior Analyst, and Senior Specialist, respectively. It can be seen that more employees in important positions are involved. Finally, in Figure 11(c), along with the event of filing for bankruptcy, more and more employees from different branches of Enron incorporation are involved and selected, and their titles are N/A, Manager, Employee, Director, CEO, Vice President, President and Senior Analyst.

To conclude, using these reasonable explanation, we claim the effectiveness of the proposed scheme. And we infer that the proposed TV-DPPs can be broadly applied to real-world applications with similar settings to the Enron communication network.

6. CONCLUSION

We proposed a fast sampling algorithm for DPPs for subset selection from a big dataset with time-varying structures. The algorithm uses the simplification of marginal density functions over successive time stamps, and also utilizes the SMC sampling technique. The proposed algorithm provides us with a real-time diverse sampling scheme by utilizing the phenomenon in which typically only proportionally small changes occur at each time stamp with respect to the entire dataset. The most prominent application of our work is the online news service, in which news corpora are updated from the multiple sources continuously, before the most diverse subset is selected to be shown to the viewers. Another application mentioned in this paper is Enron corpus which is based on online network structures. There are many other potential applications of our work to benefit the data mining community. One potential application is the modelling of latent attributes associated with a person-of-interest (POI) in a social network, such as the POI's community membership over time. In this scenario, the POIs friends in social network sites, such as Facebook or Instagram can be treated as information sources (similarly to the online news setting) where their interests are constantly updated using text and/or images. At any given time stamp, we can select a diverse subset of all interests from his social circle as an additional important cue to infer the POIs attributes.

REFERENCES

Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. 2013. Twitter-based user modeling for news recommendations. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2962–2966.

Raja Hafiz Affandi, Emily B. Fox, Ryan P. Adams, and Ben Taskar. 2014. Learning the parameters of determinantal point process kernels. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*.

Raja Hafiz Affandi, Emily B. Fox, and Ben Taskar. 2013a. Approximation inference in continuous determinantal point processes. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Raja Hafiz Affandi, Alex Kulesza, and Emily B. Fox. 2013b. Markov determinantal point processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Sihem Amer-Yahia, Francesco Bonchi, Carlos Castillo, Esteban Feuerstein, Isabel Mendez-Diaz, and Paula Zabala. 2014. Composite retrieval of diverse and complementary bundles. *IEEE Trans. Knowl. Data Eng.* 26, 11 (2014), 2662–2675.

Steve P. Borgatti. 2002. NetDraw software for network visualization. *Lexington, KY: Analytic Technologies* (2002), 95.

Alexei Borodin and Eric M. Rains. 2005. Eynard–Mehta theorem, Schur process, and their Pfaffian analogs. *J. Stat. Phys.* 121, 3–4 (2005), 291–317.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–336.

Pinyucen Chen and Lifang Hsu. 1991. On a sequential subset selection procedure for the least probable multinomial cell. *Stoch. Anal. Appl.* 20, 9 (1991), 2845–2862.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. 2006. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B: Stat. Method.* 68, 3 (2006), 411–436.

Jana Diesner and Kathleen M. Carley. 2005. Exploration of communication networks from the Enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*. Citeseer.

Jana Diesner, Terrill L. Frantz, and Kathleen M. Carley. 2005. Communication networks from the Enron email corpus it's always about the people. Enron is no different. *Comput. Math. Organ. Theory* 11, 3 (2005), 201–228.

Walter R. Gilks and Carlo Berzuini. 2001. Following a moving target – Monte Carlo inference for dynamic Bayesian models. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* 63, 1 (Jan. 2001), 127–146.

J. Gillenwater, A. Kulesza, E. Fox, and B. Taskar. 2014. Expectation-maximization for learning determinantal point processes. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 3149–3157.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012a. Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012b. Near-optimal MAP inference for determinantal point processes. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Jean Ginibre. 1965. Statistical ensembles of complex, quaternion, and real matrices. *J. Math. Phys.* 6, 3 (1965), 440–449.

Gene H. Golub and Charles F. Van Loan. 2012. *Matrix Computations*. Vol. 3. JHU Press.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2069–2077.

J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Balint Virag. 2006. Determinantal processes and independence. *Probab. Surv.* 3 (2006), 206–229.

Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda. 2012. Sequential modeling of topic dynamics with multiple timescales. *ACM Trans. Knowl. Discov. Data* 5, 4 (Feb. 2012), 19:1–19:27.

Byungkon Kang. 2013. Fast determinantal point process sampling with application to clustering. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2319–2327.

N. Kantas, A. Doucet, S. S. Singh, and J. M. Maciejowski. 2009. An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In *Proceedings of the 15th IFAC Symposium on System Identification (SYSID)*.

Bryan Klimt and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Proceedings of Euroean Conference on Machine Learning (ECML)*. Springer, 217–226.

Alex Kulesza and Ben Taskar. 2010. Structured determinantal point processes. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 1171–1179.

Alex Kulesza and Ben Taskar. 2011a. K-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*.

Alex Kulesza and Ben Taskar. 2011b. Learning determinantal point processes. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*.

Alex Kulesza and Ben Taskar. 2012. *Determinantal Point Processes for Machine Learning*.

Tongliang Liu and Dacheng Tao. 2016. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 3 (2016), 447–461.

Tongliang Liu, Dacheng Tao, Mingli Song, and Stephen Maybank. 2016. Algorithm-dependent generalization bounds for multi-task learning. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016). DOI:http://dx.doi.org/10.1109/TPAMI.2016.2544314

Odile Macchi. 1975. The coincidence approach to stochastic point processes. *Adv. Appl. Probab.* 7, 1 (1975), 83–122.

Julian Mcauley and Jure Leskovec. 2014. Discovering social circles in ego networks. *ACM Trans. Knowl. Discov. Data* 8, 1 (Feb. 2014), 4:1–4:28.

Christophe Ange Napoléon Biscio and Frédéric Lavancier. 2014. About repulsiveness of determinantal point processes.

Naoto Ohsaka, Takuya Akiba, Yuichi Yoshida, and Ken-ichi Kawarabayashi. 2014. Fast and accurate influence maximization on large networks with pruned Monte-Carlo simulations. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. 2013. Data mining a trillion time series subsequences under dynamic time warping. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 3047–3051.

Karthik Raman, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Learning to diversify from implicit feedback. In *Proceedings of WSDM Workshop on Diversity in Document Retrieval*.

Bhaskar D. Rao, Kjersti Engan, Shane F. Cotter, Jason Palmer, and Kenneth Kreutz-Delgado. 2003. Subset selection in noise based on diversity measure minimization. *IEEE Trans. Signal Process.* 51, 3 (2003), 760–770.

Kristoffer Sahlin. 2011. *Estimating Convergence of Markov Chain Monte Carlo Simulations*. Master's thesis. Stockholm University, Stockholm, Sweden.

Anders W. Sandvik. 2013. *Monte Carlo Simulations in Classical Statistical Physics*. Technical Report. Department of Physics, Boston University.

Jitesh Shetty and Jafar Adibi. 2004. *The Enron Email Dataset Database Schema and Brief Statistical Report*. Information Sciences Institute Technical Report, University of Southern California 4 (2004).

Jasper Snoek and Ryan P. Adams. 2013. A determinantal point process latent variable model for inhibition in neural spiking data. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Eric Tollefson, David Goldsman, Anton Kleywegt, and Craig Tovey. 2014. Optimal selection of the most probable multinomial alternative. *Seq. Anal.* 33, 4 (2014), 491–508.

Shusen Wang, Chao Zhang, Hui Qian, and Zhihua Zhang. 2014. Using the matrix ridge approximation to speedup determinantal point processes sampling algorithms. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

Feng Wu, Shlomo Zilberstein, and Nicholas R. Jennings. 2013. Monte-Carlo expectation maximization for decentralized POMDPs. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 397–403.

Ho Chuang Wu and Robert Wing Pong Luk. 2008. Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inf. Syst.* 26, 3 (Jun. 2008), 13:1–13:37.

Chang Xu, Dacheng Tao, and Chao Xu. 2014. Large-margin multi-view information bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 8 (2014), 1559–1572.

Chang Xu, Dacheng Tao, and Chao Xu. 2015. Multi-view intact space learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 12 (2015), 2531–2544.

Junyu Xuan, Jie Lu, Guangquan Zhang, and Xiangfeng Luo. 2015a. Topic model for graph mining. *IEEE Trans. Cybern.* 45, 12 (2015), 2792–2803.

Junyu Xuan, Jie Lu, Guangquan Zhang, Richard Xu, and Xiangfeng Luo. 2015b. Infinite author topic model based on mixed gamma-negative binomial process. In *Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM)*, 489–498.

Erheng Zhong, Wei Fan, and Qiang Yang. 2014. User behavior learning and rransfer in composite social networks. *ACM Trans. Knowl. Discov. Data* 8, 1 (Feb. 2014), 6:1–6:32.

James Y. Zou and Ryan P. Adams. 2012. Priors for diversity in generative latent variable models. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*.

Received June 2015; revised Febrary 2016; accepted May 2016