Fused Gromov-Wasserstein Alignment for Hawkes Processes

Dixin Luo\textsuperscript{1}\textsuperscript{*}, Hongteng Xu\textsuperscript{1,2}\textsuperscript{*}, Lawrence Carin\textsuperscript{1}
\textsuperscript{1}Duke University \textsuperscript{2}Infinia ML, Inc.
\{dixin.luo, hongteng.xu\}@duke.edu

Abstract

We propose a novel fused Gromov-Wasserstein alignment method to jointly learn the Hawkes processes in different event spaces, and align their event types. Given two Hawkes processes, we use fused Gromov-Wasserstein discrepancy to measure their dissimilarity, which considers both the Wasserstein discrepancy based on their base intensities and the Gromov-Wasserstein discrepancy based on their infectivity matrices. Accordingly, the learned optimal transport reflects the correspondence between the event types of these two Hawkes processes. The Hawkes processes and their optimal transport are learned jointly via maximum likelihood estimation, with a fused Gromov-Wasserstein regularizer. Experimental results show that the proposed method works well on synthetic and real-world data.

1 Introduction

There is often a need to align real-world entities in different domains, based on their sequential behavior in continuous time, \textit{e.g.}, linking accounts in different social networks based on behaviors within each network. For each domain, the entities in the domain formulate an event space and their sequential behavior can be represented as event sequences, in which each event is a tuple containing a timestamp and an event type (\textit{i.e.}, the entity involved in the event). When these event sequences yield multi-dimensional point process models, the proposed problem can be reformulated as an alignment problem: learning two point processes and finding the correspondence between their event types.

Focusing on event sequences that are modeled as Hawkes processes, we propose a novel fused Gromov-Wasserstein alignment (FGWA) method. As illustrated in Figure 1(a), the event sequences in each domain are modeled as a Hawkes process parametrized via a base intensity vector and an infectivity matrix. The base intensity captures the intrinsic expected happening rate of each event type, while the infectivity matrix describes the self- and mutually-triggering pattern between different event types. The Wasserstein discrepancy between the two domains is formulated based on their base intensities, and their Gromov-Wasserstein discrepancy is formulated based on their infectivity matrices. We learn an optimal transport to minimize the fusion of these two discrepancies, \textit{i.e.}, the fused Gromov-Wasserstein discrepancy [8]. The learned optimal transport are used to regularize updating of the Hawkes processes. After several iterations, we jointly derive the two Hawkes processes and the optimal transport, indicating the correspondence between their event types.

As shown in Fig. 1(b-e), compared with its competitors our FGWA method learns the optimal transport matrix with the highest certainty — each row just contains one nonzero element.

2 Proposed Alignment Method

A temporal point process with \(C\) event types can be represented as a counting process \(N(t) = \{N_c(t)\}_{c=1}^C\), where each \(N_c(t)\) is the number of type-\(c\) events happening at or before time \(t\). The

\textsuperscript{*}Equal contribution

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.
event sequences of the point process are denoted \( S = \{ s_n = (t^n_i, c^n_i)_{i=1}^N \} \), where \( N \) is the number of sequences, \( I_n \) is the number of events in \( s_n \), with \( t^n_i \in [0, T] \) and \( c^n_i \in \mathbb{C} = \{ 1, ..., C \} \) representing respectively the time-stamp and the event type of the \( i \)-th event. Point processes are characterized by their intensity functions \( \{ \lambda_c(t) \}_{c=1}^C \), where \( \lambda_c(t) = \mathbb{E}[dN_c(t)/dt] \) represents the expected instantaneous happening rate of type-\( c \) events given the history \( \mathcal{H}_c(t) = \{ (t_i, c_i) | t_i < t, c_i \in \mathbb{C} \} \). As a special kind of point process, the Hawkes process has a particular form of intensity [8,11]:

\[
\lambda_c(t) = \mu_c + \sum_{i:t_i < t} \phi_{cc}(t-t_i), \quad \text{for} \quad c \in \mathbb{C}.
\]

Here, \( \mu_c \) is the base intensity, independent of history, capturing the intrinsic happening rate of the type-\( c \) event, and \( \phi_{cc}(t) \) is the impact function measuring the infectivity of the type-\( c' \) event to the type-\( c \) event type, over time. Generally, we can parameterize each impact function by a predefined base function, i.e., \( \phi_{cc'}(t) = \alpha_{cc'} g(t) \), where \( g(t) \) is an exponential function and \( \alpha_{cc'} \) is a learnable coefficient. Therefore, we denote an event sequence yielding to a Hawkes process as \( s \sim \mathcal{HP}(\mu, A) \), with base intensity \( \mu \in \mathbb{R}^C \) and infectivity matrix \( A = [a_{cc'}]_{C \times C} \). Given a set of event sequences \( S \), we can learn a Hawkes processes via maximum likelihood estimation. The likelihood of \( S \) is

\[
L(S; \mu, A) = \prod_n p(s_n; \mu, A) = \prod_n \prod_{i=1}^{I_n} \lambda_{c_i}(t^n_i) \exp\left(-\sum_{c=1}^C \int_0^{T_n} \lambda_c(s)ds \right). \tag{2}
\]

The base intensity and the infectivity matrix provide, respectively, the feature of each event type and the relationship among different event types. These two kinds of information can be applied to measure the similarity between different event types in a framework of fused Gromov-Wasserstein discrepancy [9]. In particular, fused Gromov-Wasserstein discrepancy is a combination of traditional Wasserstein discrepancy (WD) [9] and Gromov-Wasserstein discrepancy (GWD) [6]. Focusing on the alignment of Hawkes processes, the proposed fused Gromov-Wasserstein discrepancy can be used as a regularizer when learning the Hawkes process models. Suppose that we have two sets of event sequences corresponding to source and target Hawkes processes, i.e., \( S_s \sim \mathcal{HP}(\mu_s, A_s) \) and \( S_t \sim \mathcal{HP}(\mu_t, A_t) \), where \( \mu_k = [\mu_{kk}] \in \mathbb{R}^{C^2} \) and \( A_k = [a_{kk}] \in \mathbb{R}^{C^2 \times C^2} \) for \( k = s \) and \( t \). We learn these two Hawkes processes and align their event types via maximum likelihood estimation with a fused Gromov-Wasserstein regularizer:

\[
\min_{\{ \mu_k, A_k \geq 0 \}}_{k=s,t} \min_{\{ u_s, u_t \}} \frac{1}{2} \sum_{k=s,t} \log L(S_k; \mu_k, A_k) + \gamma d_{fgw}^2(u_s, u_t; \mu_s, \mu_t, A_s, A_t) \tag{3}
\]

where \( u_s \) and \( u_t \) represent the empirical distribution of the event type in the source and target domain, respectively. These are estimated via the histograms of the counts of events according to \( S_s \) and \( S_t \). The hyperparameter \( \gamma \) controls the significance of the proposed fused Gromov-Wasserstein regularizer. \( d_{fgw}^2(u_s, u_t; \mu_s, \mu_t, A_s, A_t) \) is the discretized version of fused Gromov-Wasserstein discrepancy based on the Hawkes process parameters:

\[
d_{fgw}^2 = \min_{\tau \in \Pi(u_s, u_t)} (1-\alpha) \sum_{i,j} L(\mu_{ij}^s, \mu_{ij}^t) T_{ij} + \alpha \sum_{i,j,i',j'} L(\alpha_{ij}^s, \alpha_{ij}^t) T_{ij} T_{i'j'} \tag{4}
\]

\[
= \min_{\tau \in \Pi(u_s, u_t)} (1-\alpha) L_{\mu} + \alpha L_{A}(T), \tag{4}
\]

where \( L(a, b) = |a-b|^2 \) is a mean-square-error (MSE) loss, and \( \langle \cdot, \cdot \rangle \) represents the matrix inner product. Accordingly, \( L_{\mu} = [L(\mu_{ij}^s, \mu_{ij}^t)] \in \mathbb{R}^{C^2 \times C^2} \) and \( L_{A}(T) = [L_{ij}^T] \), whose element \( L_{ij}^T =

Figure 1: (a) An illustration of our method. (b-e) Comparisons on synthetic data.
$\sum_{i,t} L(a_{ij}^t, a_{ij}^t)T_{ij}$; and $\Pi(u_s, u_t) = \{T \geq 0 | T 1_{C_s} = u_s, T^T 1_{C_t} = u_t\}$, where $1_D$ represents a $D$-dimensional all-one vector. $\alpha$ controls the balance between the Wasserstein term and the Gromov-Wasserstein term. The Wasserstein discrepancy compares the event types of the two Hawkes processes in an absolute way while the Gromov-Wasserstein discrepancy compares their event types in a relational way. Taking them into account, the final optimal transport represents the joint distribution of the event types in different Hawkes processes. As shown in Figure 1(a) the pairs of event types with high probability indicate the correspondence between the event types. The learned optimal transport fills the gap between the source and the target Hawkes processes, and the models can be learned jointly under the guidance of the optimal transport.

### 3 Learning Algorithm

We solve (3) effectively based on an alternating optimization strategy. In each iteration, given the $L$ with high probability indicate the correspondence between the event types. The learned optimal transport balances the influence of Wasserstein discrepancy and that of Gromov-Wasserstein discrepancy. $\alpha$ controls both synthetic and real-world data. In the following experiments, we set $\alpha = 0$. (HP-WD); and 3) aligning Hawkes process purely based on Gromov-Wasserstein discrepancy, i.e., $\alpha = 1$ (HP-GWD). Given the real correspondence $T$ and the optimal transport $\hat{T}$, we evaluate various methods based on the following three measurements: i) Top-$K$ alignment accuracy $\text{Acc} = (T, \text{top}_K(\hat{T}))/C$, where $\text{top}_K(\hat{T})$ converts each row of $\hat{T}$ to binary vector, whose nonzero elements correspond to the maximum $K$ values of the row. ii)
We have proposed an alignment method for Hawkes processes based on fused Gromov-Wasserstein discrepancy, which achieves encouraging results on matching the event types of different Hawkes processes. The proposed method shows the potential of optimal transport techniques to the learning and the alignment of temporal point processes. In the future, we plan to further improve the scalability of the proposed method for large-scale applications.

Acknowledgements This research was supported in part by DARPA, DOE, NIH, ONR and NSF.

Table 1: Comparisons for various methods on synthetic and real-world data.

| Method  | Synthetic | Real-world | MIMIC-III | MC3 |
|---------|-----------|------------|-----------|-----|
|         | Acc-1     | Acc-K      | Acc-1     | Acc-1 |
| C=10    | 0.41      | 0.12       | 0.19      | 0.09 |
| C=50    | 0.12      | 0.03       | 0.314     | 0.08 |
| C=100   | 0.03      | 0.03       | 0.081     | 0.06 |

Cosine similarity $\text{Sim} = \langle T, \hat{T} \rangle / (\|T\|_F \|\hat{T}\|_F)$. iii) Entropy $H = -\langle \hat{T}, \log \hat{T} \rangle$. When the real correspondence is bijective, this measurement reflects the uncertainty of the learned correspondence.

**Synthetic data** The synthetic event sequences are generated via the following method: For the source Hawkes process with $C_s$ event types, we generate $\mu_s = [\mu_{s}^{i}]$ with $\mu_{s}^{i} \sim \text{Uniform}[0, 1/C_s]$ and $A_s = [a_{s}^{ij}]$ with $a_{s}^{ij} \sim \text{Uniform}[0,1]$. Given a predefined correspondence matrix $T$, the parameters of the target Hawkes process are $\mu_t = T^\top \mu_s$ and $A_t = T^\top A_s T$. Accordingly, the source and the target event sequences are generated based on Ogata’s thinning algorithm [5]. We keep $C_s = C_t$ and set them from $\{10, 50, 100\}$. For both the source and target Hawkes process, we simulate $\mathcal{C}(C = C_s = C_t)$ event sequences with length $T = C_s^2$, and set the decay function as an exponential function, i.e., $g(t) = \exp(-t)$. We consider 30 trials, and calculate the average results. Because the real correspondence in each trial is a bijective function, we consider the top-1 alignment accuracy in this experiment. Table 1 shows the results of various methods. The proposed FGWA method outperforms its competitors in most situations. Considering the fused Gromov-Wasserstein regularizer is beneficial for our alignment task indeed. The entropy of our optimal transport matrix is often smaller than those of other methods, which means that the correspondence we have learned has high certainty. The optimal transport matrices shown in Figure 1(b-e) further demonstrates that our FGWA method has the highest certainty.

**Real-world data** We further test the proposed method on two real-world datasets: the MIMIC-III dataset [2] and the call-network used in the Mini-Challenge 3 (MC3) of VAST Challenge 2018 [3]. The MIMIC-III records 18,756 patient admission sequences. Each admission is an event in the sequence, containing a pair of diagnose ICD code and procedure ICD code. The dataset contains 56 diagnoses and 25 procedures. According to the coherency of the diagnoses and the procedures in the observed admission sequences, we obtain the correspondence between them. The call-network we used records the phone calls among a company’s employees in continuous time domain, which contains 2,507 callers and 2,481 responders. The pairs of callers and responders appearing in the call-network indicates the correspondence between them. For the MIMIC-III dataset, we consider the sequences of diagnoses and those of procedures, and model them via two Hawkes processes. Applying various alignment methods, we try to estimate the correspondence between diagnoses and procedures. Similarly, for the MC3 dataset, we model the sequences of callers and those of responders via two Hawkes processes and try to estimate the correspondence between them. In both of these two datasets, their correspondences are not bijective. Therefore, we consider top-5 alignment accuracy for the MIMIC-III dataset and top-50 alignment accuracy for the MC3 dataset, respectively. Table 1 shows that our FGWA method outperforms other methods on both datasets.

5 Conclusions and Future Work

We have proposed an alignment method for Hawkes processes based on fused Gromov-Wasserstein discrepancy, which achieves encouraging results on matching the event types of different Hawkes processes. The proposed method shows the potential of optimal transport techniques to the learning and the alignment of temporal point processes. In the future, we plan to further improve the scalability of the proposed method for large-scale applications.

Acknowledgements This research was supported in part by DARPA, DOE, NIH, ONR and NSF.
References

[1] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[2] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[3] D. Luo, H. Xu, Y. Zhen, X. Ning, H. Zha, X. Yang, and W. Zhang. Multi-task multi-dimensional Hawkes processes for modeling event sequences. In *IJCAI*, 2015.

[4] H. Mei and J. M. Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *NIPS*, 2017.

[5] Y. Ogata. On Lewis’ simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[6] G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *ICML*, 2016.

[7] M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[8] T. Vayer, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. Fused Gromov-Wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*, 2018.

[9] C. Villani. *Optimal transport: Old and new*, volume 338. Springer Science & Business Media, 2008.

[10] H. Xu, D. Luo, H. Zha, and L. Carin. Gromov-wasserstein learning for graph matching and node embedding. *arXiv preprint arXiv:1901.06003*, 2019.

[11] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multidimensional Hawkes processes. In *AISTATS*, 2013.