VigDet: Knowledge Informed Neural Temporal Point Process for Coordination Detection on Social Media

Anonymous Author(s)
Affiliation
Address
email

Abstract

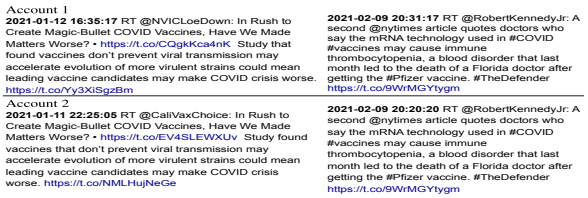
Recent years have witnessed an increasing use of *coordinated* accounts on social media, operated by *misinformation* campaigns to influence public opinion and manipulate social outcomes. Consequently, there is an urgent need to develop an effective methodology for coordinated group detection to combat the misinformation on social media. However, the sparsity of account activities on social media limits the performance of existing deep learning based coordination detectors as they can not exploit useful prior knowledge. Instead, the detectors incorporated with prior knowledge suffer from limited expressive power and poor performance. Therefore, in this paper we propose a coordination detection framework incorporating neural temporal point process with prior knowledge such as temporal logic or pre-defined filtering functions. Specifically, when modeling the observed data from social media with neural temporal point process, we jointly learn a Gibbs-like distribution of group assignment based on how consistent an assignment is to (1) the account embedding space and (2) the prior knowledge. To address the challenge that the distribution is hard to be efficiently computed and sampled from, we design a theoretically guaranteed variational inference approach to learn a mean-field approximation for it. Experimental results on a real-world dataset show the effectiveness of our proposed method compared to SOTA model in both unsupervised and semi-supervised settings. We further apply our model on a COVID-19 Vaccine Tweets dataset. The detection result suggests presence of suspicious coordinated efforts on spreading misinformation about COVID-19 vaccines.

1 Introduction

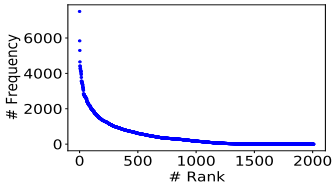
Recent research reveals that the information diffusion on social media is heavily influenced by hidden account groups [1, 26, 27], many of which are *coordinated* accounts operated by *misinformation* campaigns (an example shown in Fig. 1a). This form of abuse to spread misinformation has been seen in different fields, including politics (e.g. the election) [16] and healthcare (e.g. the ongoing COVID-19 pandemic) [27]. This persistent abuse as well as the urgency to combat misinformation prompt us to develop effective methodologies to uncover hidden coordinated groups from the diffusion cascade of information on social media.

On social media, the diffusion cascade of a piece of information (like a tweet) can be considered as a realization of a marked temporal point process where each mark of an event type corresponds to an account. Therefore, we can formulate uncovering coordinated accounts as detecting mark groups from observed point process data, which leads to a natural solution that first acquires account embeddings from the observed data with deep learning (e.g. neural temporal point process) and then conducts group detection in the embedding space [16, 25]. However, the data from social media has a special and important property, which is that the appearance of accounts in the diffusion cascades usually follows a long-tail distribution [14] (an example shown in Fig. 1b). This property brings

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.



(a) Example of collaborated behaviours.



(b) Frequency statistic of accounts.

Figure 1: The figure (a) is an example of coordinated accounts detected by our method on Twitter. They retweet similar anti-vaccine contents about COVID-19 Vaccines from same or different sources. The figure (b) is the frequency statistic of accounts in IRA dataset about the U.S. 2016 Election.

a unique challenge: compared to a few dominant accounts, most accounts appear sparsely in the data, limiting the performance of deep representation learning based models. Some previous works exploiting pre-defined *collective behaviours* [2, 31, 21] can circumvent this challenge. They mainly follow the paradigm that first constructs similarity graphs from the data with some prior knowledge or hypothesis and then conducts graph based clustering. Their expressive power, however, is heavily limited as the complicated interactions are simply represented as edges with scalar weights, and they exhibit strong reliance on predefined signatures of coordination. As a result, their performances are significantly weaker than the SOTA deep representation learning based model [25].

To address above challenges, we propose a knowledge informed neural temporal point process model, named Variational Inference for Group Detection (VigDet). It represents the domain knowledge of collective behaviors of coordinated accounts by defining different signatures of coordination, such as accounts that co-appear, or are synchronized in time, are more likely to be coordinated. Different from previous works that highly rely on assumed prior knowledge and cannot effectively learn from the data [2, 31], VigDet encodes prior knowledge as temporal logic and cubic functions so that it guides the learning of neural point process model and effectively infer coordinated behaviors. In addition, it maintains a distribution over group assignments and defines a potential score function that measures the consistency of group assignments in terms of both embedding space and prior knowledge. As a result, VigDet can make effective inferences over the constructed prior knowledge graph while jointly learning the account embeddings using neural point process.

A crucial challenge in our framework is that the group assignment distribution, whose formulation is very similar to the Gibbs distribution in Conditional Random Fields [13], contains a partition function as normalizer [12]. Consequently it is NP-hard to compute or sample, leading to difficulties in both learning and inference [4, 11]. To address this issue, we apply **variational inference** [18]. Specifically, we approximate the Gibbs-like distribution as a mean field distribution [20]. Then we jointly learn the approximation and learnable parameters with EM algorithm to maximize the evidence lower bound (ELBO) [18] of the observed data likelihood. In the E-step, we freeze the learnable parameters and infer the optimal approximation, while in the M-step, we freeze the approximation and update the parameters to maximize an objective function which is a lower bound of the ELBO with theoretical guarantee. Our experiments on real world dataset [16] involving coordination detection validate the effectiveness of our model compared with other baseline models including the current SOTA. We further apply our method on a dataset of tweets about COVID-19 vaccine without ground-truth coordinated group label. The analysis on the detection result suggests the existence of suspicious coordinated efforts to spread misinformation and conspiracies about COVID-19 vaccines.

2 Related Work

2.1 Graph based coordinated group detection

One typical coordinated group detection paradigm is to construct a graph measuring the similarity or interaction between accounts and then conduct clustering on the graph or on the embedding acquired by factorizing the adjacency matrix. There are two typical ways to construct the graph. One way is to measure the similarity or interaction with pre-defined features supported by prior knowledge or assumed signatures of coordinated or collective behaviors, such as co-activity, account clickstream and time sychronization [5, 24, 31]. The other way is to learn an interaction graph by fitting the

data with the temporal point process models considering mutually influence between accounts as scalar scores as in traditional Hawkes Process [35]. A critical drawback of both methods is that the interaction between two accounts is simply represented as an edge with scalar weight, resulting in poor ability to capture complicated interactions. In addition, the performance of prior knowledge based method is unsatisfactory due to reliance on the quality of prior knowledge or hypothesis of collective behaviors, which may vary with time [33].

2.2 Representation learning based coordinated group detection

To address the reliance to the quality of prior knowledge and the limited expressive power of graph based method, recent research trys to directly learn account representations from the observed data. In [16], Inverse Reinforcement Learning (IRL) is applied to learn the reward behind an account's observed behavior and the learnt reward is forwarded into a classifier as features. However, since different accounts' activity traces are modeled independently, it is hard for IRL to model the interactions among different accounts. The current SOTA method in this direction is a neural temporal point process model named AMDN-HAGE [25]. Its backbone (AMDN), which can efficiently capture account interactions from observed activity traces, contains an account embedding layer, a history encoder and an event decoder. The account embedding vectors are optimized under the regularization of a Gaussian Mixture Model (the HAGE part). However, as a data driven deep learning model, the learning process of AMDN-HAGE lacks the guidance of prior knowledge from human. In contrast, we propose VigDet, a framework integrating neural temporal point process together and prior knowledge to address inherent sparsity of account activities.

3 Preliminary and Task Definition

3.1 Marked Temporal Point Process

A marked temporal point process (MTPP) is a stochastic process whose realization is a discrete event sequence $S = [(v_1, t_1), (v_2, t_2), (v_3, t_3), \cdots (v_n, t_n)]$ where $v_i \in \mathcal{V}$ is the type mark of event i and $t_i \in \mathbb{R}^+$ is the timestamp [6]. We denote the historical event collection before time t as $H_t = \{(v_i, t_i) | t_i < t\}$. Given a history H_t, the conditional probability that an event with mark $v \in \mathcal{V}$ happens at time t is formulated as: $p_v(t|H_t) = \lambda_v(t|H_t) \exp\left(-\int_{t_{i-1}}^{t} \lambda_v(s|H_t)ds\right)$, where $\lambda_v(t|H_t)$, also known as intensity function, is defined as $\lambda_v(t|H_t) = \frac{\mathbb{E}[dN_v(t)|H_t]}{dt}$, i.e. the derivative of the total number of events with type mark v happening before or at time t, denoted as $N_v(t)$. In social media data, Hawkes Process (HP) [35] is the commonly applied type of temporal point process. In Hawkes Process, the intensity function is defined as $\lambda_v(t|H_t) = \mu_v + \sum_{(v_i,t_i)\in H_t} \alpha_{v,v_i}\kappa(t - t_i)$ where $\mu_v > 0$ is the self activating intensity and $\alpha_{v,v_i} > 0$ is the mutually triggering intensity modeling mark v_i's influence on v and κ is a decay kernel to model *influence decay* over time.

3.2 Neural Temporal Point Process

In Hawkes Process, only the μ and α are learnable parameters. Such weak expressive power hinders Hawkes Process from modeling complicated interactions between events. Consequently, researchers conduct meaningful trials on modeling the intensity function with neural networks [7, 17, 34, 38, 28, 19, 25]. In above works, the most recent work related to coordinated group detection is AMDN-HAGE [25], whose backbone architecture AMDN is a neural temporal point process model that encodes an event sequence S with masked self-attention:

$$A = \sigma(QK^T/\sqrt{d}), \quad C = F(AV), \quad Q = XW_q,\ K = XW_k,\ V = XW_v \tag{1}$$

where σ is a masked activation function avoiding encoding future events into historical vectors, $X \in \mathbb{R}^{L\times d}$ (L is the sequence length and d is the feature dimension) is the event sequence feature, F is a feedforward neural network or a RNN that summarizes historical representation from the attentive layer into context vectors $C \in \mathbb{R}^{L\times d'}$, and W_q, W_k, W_v are learnable weights. Each row X_i in X (the feature of event (v_i, t_i)) is a concatenation of learnable mark (each mark corresponds to an account on social media) embedding E_{v_i}, position embedding $PE_{pos=i}$ with trigonometric integral function [29] and temporal embedding $\phi(t_i - t_{i-1})$ using translation-invariant temporal

kernel function [32]. After acquiring C, the likelihood of a sequence S given mark embeddings E and other parameters in AMDN, denoted as θ_a, can be modeled as:

$$\log p_{\theta_a}(S|E) = \sum_{i=1}^{L} \left[\log p(v_i|C_i) + \log p(t_i|C_i)\right],$$
$$p(v_i|C_i) = \text{softmax}(\text{MLP}(C_i))_{v_i}, \quad p(t_i|C_i) = \sum_{k=1}^{K} w_i^k \frac{1}{s_i^k\sqrt{2\pi}} \exp\left(-\frac{(\log t_i - \mu_i^k)^2}{2(s_i^k)^2}\right) \tag{2}$$
$$w_i = \sigma(V_w C_i + b_w), \quad s_i = \exp(V_s C_i + b_s), \quad \mu_i = V_\mu C_i + b_\mu$$

3.3 Task Definition: Coordinated Group Detection on Social Media

In coordinated group detection, we are given a temporal sequence dataset $\mathcal{S} = \{S_1, ..., S_{|D|}\}$ from social media, where each sequence $S_i = [(v_{i1}, t_{i1}), (v_{i2}, t_{i2}), \cdots]$ corresponds to a piece of information, e.g. a tweet, and each event (v_{ij}, t_{ij}) means that an account $v_{ij} \in \mathcal{V}$ (corresponding to a type mark in MTPP) interacts with the tweet (like comment or retweet) at time t_{ij}. Supposing that the $\mathcal{V}$ consists of M groups, our objective is to learn a group assignment $Y = \{y_v | v \in \mathcal{V}, y_v \in \{1, ..., M\}\}$. This task can be conducted under unsupervised or semi-supervised setting. In unsupervised setting, we do not have the group identity of any account. As for the semi-supervised setting, the groundtruth group identity Y_L of a small account fraction $\mathcal{V}_L \subset \mathcal{V}$ is accessible. Current SOTA model on this task is AMDN-HAGE with KMeans. It first learns the account embeddings E with AMDN-HAGE. Then it obtains group assignment Y using KMeans clustering on learned E.

4 Proposed Method: VigDet

In this section, we introduce our proposed model called **VigDet** (Variational Inference for Group Detection), which bridges neural temporal point process and graph based method based on prior knowledge. Unlike the existing methods, in VigDet we regularize the learning process of the account embeddings with the prior knowledge based graph so that the performance can be improved. Such a method address the heavy reliance of deep learning model on the quality and quantity of data as well as the poor expressive power of existing graph based methods exploiting prior knowledge.

4.1 Prior Knowledge based Graph Construction

For the prior knowledge based graph construction, we apply co-activity [24] to measure the similarity of accounts. This method assumes that the accounts that always appear together in same sequences are more likely to be in the same group. Specifically, we construct a dense graph $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ whose node set is the account set and the weight w_{uv} of an edge (u, v) is the co-occurrence:

$$w_{uv} = \sum_{S \in \mathcal{S}} \mathbb{1}((u \in S) \wedge (v \in S)) \tag{3}$$

However, when integrated with our model, this edge weight is problematic because the coordinated accounts may also appear in the tweets attracting normal accounts. Although the co-occurrence of coordinated account pairs is statistically higher than other account pairs, since coordinated accounts are only a small fraction of the whole account set, our model will tend more to predict an account as normal account. Therefore, we apply one of following two strategies to acquire filtered weight w'_{uv}:

Cubic Function based filtering: the co-occurrence of a coordinated account pair is statistically higher than a coordinated-normal pairs. Thus, we can use a cubic function to enlarge the difference and then conduct normalization:

$$w'_{uv} = (\sum_{S \in \mathcal{S}} \mathbb{1}((u \in S) \wedge (v \in S)))^3 \tag{4}$$

where $u \in S$ and $v \in S$ mean that u and v appear in the sequence respectively. Then the weight with relatively low value will be filtered via normalization (details in next subsection).

Temporal Logic [15] based filtering: We can represent some prior knowledge as a logic expression of temporal relations, denoted as $r(\cdot)$, and then only count those samples satisfying the logic expressions. Here, we assume that the active time of accounts of the same group are more likely to be similar.

Therefore, we only consider the account pairs whose active time overlap is larger than a threshold (we apply half a day, i.e. 12 hours):

$$\begin{aligned} w'_{uv} &= \sum_{S \in \mathcal{S}} \mathbb{1}((u \in S) \wedge (v \in S) \wedge r(u, v, S)), \\ r(u, v, S) &= \mathbb{1}(\min(t_{ul}, t_{vl}) - \max(t_{us}, t_{vs}) > c) \end{aligned} \tag{5}$$

where t_{ul}, t_{vl} are the last time that u and v appears in the sequence and t_{us}, t_{vs} are the first (starting) time that u and v appears in the sequence.

4.2 Integrate Prior Knowledge and Neural Temporal Point Process

To integrate prior knowledge and neural temporal point process, while maximizing the likelihood of the observed sequences $\log p(\mathcal{S}|E)$ given account embeddings, VigDet simultaneously learns a distribution over group assignments Y defined by the following potential score function given the account embeddings E and the prior knowledge based graph $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$:

$$\Phi(Y; E, \mathcal{G}) = \sum_{u \in \mathcal{V}} \varphi_\theta(y_u, E_u) + \sum_{(u,v) \in \mathcal{E}} \phi_{\mathcal{G}}(y_u, y_v, u, v) \tag{6}$$

where $\varphi_\theta(y_u, E_u)$ is a learnable function measuring how an account's group identity y_u is consistent to the learnt embedding, e.g. a feedforward neural network. And $\phi_{\mathcal{G}}(y_u, y_v, u, v)$ is pre-defined as:

$$\phi_{\mathcal{G}}(y_u, y_v, u, v) = \frac{w_{uv}}{\sqrt{d_u d_v}} \mathbb{1}(y_u = y_v) \tag{7}$$

where $d_u, d_v = \sum_k w_{uk}, \sum_k w_{vk}$ are the degrees of u, v and $\mathbb{1}(y_u = y_v)$ is an indicator function that equals 1 when its input is true and 0 otherwise. By encouraging co-appearing accounts to be assigned in to the same group, $\phi_{\mathcal{G}}(y_u, y_v, u, v)$ regularizes E and φ_θ with prior knowledge. With the above potential score function, we can define the conditional distribution of group assignment Y given embedding E and the graph $\mathcal{G}$:

$$P(Y|E, \mathcal{G}) = \frac{1}{Z} \exp(\Phi(Y; E, \mathcal{G})) \tag{8}$$

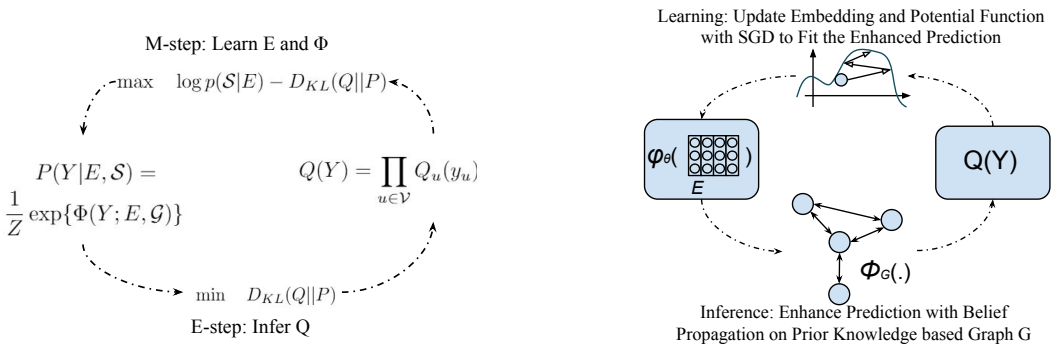
where $Z = \sum_Y \exp(\Phi(Y; E, \mathcal{G}))$ is the normalizer keeping $P(Y|E, \mathcal{G})$ a distribution, also known as partition function [12, 10]. It sums up $\exp(\Phi(Y; E, \mathcal{G}))$ for all possible assignment Y. As a result, calculating $P(Y|E, \mathcal{G})$ accurately and finding the assignment maximizing $\Phi(Y; E, \mathcal{G})$ are both NP-hard [4, 11]. Consequently, we approximate $P(Y|E, \mathcal{G})$ with a mean field distribution $Q(Y) = \prod_{u \in \mathcal{V}} Q_u(y_u)$. To inform the learning of E and φ_θ with the prior knowledge behind $\mathcal{G}$ we propose to jointly learn Q, E and φ_θ by maximizing following objective function, which is the Evidence Lower Bound (ELBO) of the observed data likelihood $\log p(\mathcal{S}|E)$ given embedding E:

$$O(Q, E, \varphi_\theta; \mathcal{S}, G) = \log p(\mathcal{S}|E) - D_{KL}(Q||P) \tag{9}$$

In this objective function, the first term is the likelihood of the obeserved data given account embeddings, which can be modeled as $\sum_{S \in \mathcal{S}} \log p_{\theta_a}(S|E)$ with a neural temporal point process model like AMDN. The second term regularize the model to learn E and φ_θ such that $P(Y|E, \mathcal{G})$ can be approximated by its mean field approximation as precisely as possible. Intuitively, this can be achieved when the two terms in the potential score function, i.e. $\sum_{u \in \mathcal{V}} \varphi_\theta(y_u, E_u)$ and $\sum_{(u,v) \in \mathcal{E}} \phi_{\mathcal{G}}(y_u, y_v, u, v)$ agree with each other on every possible Y.The above lower bound can be optimized via variational EM algorithm [18]:

E-step: Inference Procedure. In E-step, we aim at inferring the optimal $Q(Y)$ that minimizes $D_{KL}(Q||P)$. Note that the formulation of $\Phi(Y; E, \mathcal{G})$ is same as Conditional Random Fields (CRF) [13] model although their learnable parameters are different. In E-step such difference is not important as all parameters in $\Phi(Y; E, \mathcal{G})$ are frozen. As existing works about CRF [12, 10] have theoretically proven, following iterative updating function of belief propagation converges at an optimal solution:

$$Q_u(y_u = m) = \frac{\hat{Q}_u(y_u = m)}{Z_u} = \frac{1}{Z_u} \exp\{\varphi_\theta(m, E_u) + \sum_{v \in \mathcal{V}} \sum_{1 \le m' \le M} \phi_{\mathcal{G}}(m, m', u, v) Q_v(y_v = m')\} \tag{10}$$



(a) The overview of the learning algorithm.

(b) The intuition behind the framework

Figure 2: Figure (a) presents VigDet's EM loop, while figure (b) illustrates cooresponding intuition.

where $Q_u(y_u = m)$ is the probability that account u is assigned into group m and $Z_u = \sum_{1 \le m \le M} \hat{Q}_u(y_u = m)$ is the normalizer keeping Q_u as a valid distribution.

M-step: Learning Procedure. In M-step, given fixed inference of Q we aim at maximizing O_M:

$$O_M = \log p(\mathcal{S}|E) - D_{KL}(Q||P) = \log p(\mathcal{S}|E) + \mathbb{E}_{Y \sim Q} \log P(Y|E, \mathcal{G}) + \text{const} \tag{11}$$

The key challenge in M-step is that calculating $\mathbb{E}_{Y \sim Q} \log P(Y|E, \mathcal{G})$ is NP-hard [4, 11]. To address this challenge, we propose to alternatively optimize following theoretically justified lower bound:

Theorem 1 *Given a fixed inference of Q and a pre-defined $\phi_{\mathcal{G}}$, we have following inequality:*

$$\begin{aligned}\mathbb{E}_{Y \sim Q} \log P(Y|E, \mathcal{G}) &\ge \mathbb{E}_{Y \sim Q} \sum_{u \in \mathcal{V}} \log \frac{\exp\{\varphi_\theta(y_u, E_u)\}}{\sum_{1 \le m' \le M} \exp\{\varphi_\theta(m', E_u)\}} + \textit{const} \\ &= \sum_{u \in \mathcal{V}} \sum_{1 \le m \le M} Q_u(y_u = m) \log \frac{\exp\{\varphi_\theta(m, E_u)\}}{\sum_{1 \le m' \le M} \exp\{\varphi_\theta(m', E_u)\}} + \textit{const}\end{aligned} \tag{12}$$

The proof of this theorem is provided in the Appendix. Intuitively, above objective function treats the Q as a group assignment enhanced via label propagation on the prior knowledge based graph and encourages E and φ_θ to correct themselves by fitting the enhanced prediction. Compared with pseudolikelihood [3] which is applied to address similar challenges in recent works [23], the proposed lower bound has a computable closed-form solution. Thus, we do not really need to sample Y from Q so that the noise is reduced. Also, this lower bound does not contain $\phi_{\mathcal{G}}$ explicitly in the non-constant term. Therefore, we can encourage the model to encode graph information into the embedding.

Joint Training: The E-step and M-step form a closed loop. To create a starting point, we initialize E with the embedding layer of a pre-trained neural temporal process model (in this paper we apply AMDN-HAGE) and initialize φ_θ via clustering learnt on E (like fitting the φ_θ to the prediction of KMeans). After that we repeat E-step and M-step to optimize the model. The pseudo code of the training algorithm is presented in Alg. 1.

Algorithm 1 Training Algorithm of VigDet.

Require: Dataset $\mathcal{S}$ and pre-defined $\mathcal{G}$ and $\phi_{\mathcal{G}}$
Ensure: Well trained Q, E and φ_θ
1: Initialize E with the embedding layer of AMDN-HAGE pre-trained on S.
2: Initialize φ_θ on the initialized E.
3: **while** not converged **do**
4: Acquire Q by repeating Eq. 10 with E, φ_θ and $\phi_{\mathcal{G}}$ until convergence.{E-step}
5: $\varphi_\theta, E \leftarrow \text{argmax}_{\varphi_\theta, E} \log p(\mathcal{S}|E) + \mathbb{E}_{Y \sim Q} \sum_{u \in \mathcal{V}} \log \frac{\exp\{\varphi_\theta(y_u, E_u)\}}{\sum_{1 \le m' \le M} \exp\{\varphi_\theta(m', E_u)\}}$. {M-step}
6: **end while**

4.3 Semi-supervised extension

Above framework does not make use of the ground-truth label in the training procedure. In semi-supervised setting, we actually have the group identity Y_L of a small account fraction $\mathcal{V}_L \subset \mathcal{V}$. Under this setting, we can naturally extend the framework via following modification to Alg. 1: For account $u \in \mathcal{V}_L$, we set Q_u as a one-hot distribution, where $Q_u(y_u = y'_u) = 1$ for the groundtruth identity y'_u and $Q_u(y_u = m) = 0$ for other $m \in \{1, ..., M\}$.

5 Experiments

5.1 Coordination Detection on IRA Dataset

We utilize Twitter dataset containing coordinated accounts from Russia's Internet Research Agency (IRA dataset [16, 25]) attempting to manipulate the U.S. 2016 Election. The dataset contains tweet sequences (i.e., tweet with account interactions like comments, replies or retweets) constructed from the tweets related to the U.S. 2016 Election.

This dataset contains activities involving 2025 Twitter accounts. Among the 2025 accounts, 312 are identified through U.S. Congress investigations[1] as coordinated accounts and other 1713 accounts are normal accounts joining in discussion about the Election during during the period of activity those coordinated accounts. This dataset is applied for evaluation of coordination detection models in recent works [16, 25]. In this paper, we apply two settings: unsupervised setting and semi-supervised setting. For unsupervised setting, the model does not use any ground-truth account labels in training (but for hyperparameter selection, we hold out 100 randomly sampled accounts as validation set, and evaluate with reported metrics on the remaining 1925 accounts as test set). For the semi-supervised setting, we similarly hold out 100 accounts for hyperparameter selection as validation set, and another 100 accounts with labels revealed in training set for semi-supervised training). The evaluation is reported on the remaining test set of 1825 accounts. The hyper parameters of the backbone of VigDet (AMDN) follow the original paper [25]. Other implementation details are in the Appendix.

5.1.1 Evaluation Metrics and Baselines

In this experiment, we mainly evaluate the performance of two version of VigDet: VigDet (CF) and VigDet (TL). VigDet (CF) applies Cubic Function based filtering and VigDet (TL) applies Temporal Logic based filtering. We compare them against existing approaches that utilize account activities to identify coordinated accounts.

Unsupervised Baselines: Co-activity clustering [24] and Clickstream clustering [31] are based on pre-defined similarity graphs. HP (Hawkes Process) [35] is a learnt graph based method. IRL[16] and AMDN-HAGE[25] are two recent representation learning method.

Semi-Supervised Baselines: Semi-NN is semi-supervised feedforward neural network without requiring additional graph structure information. It is trained with self-training algorithm [37, 22]. Label Propagation Algorithm (LPA) [36] and Graph Neural Network (GNN) (we use the GCN [9], the most representative GNN) [9, 30, 8] are two baselines incorporated with graph structure. In LPA and GNN, for the graph structures (edge features), we use the CF and TL based prior knowledge graphs (similarly used in VigDet), as well as the graph learned by HP model as edge features. For the node features in GNN, we provide the account embeddings learned with AMDN-HAGE.

Ablation Variants: To verify the importance of the EM-based variational inference framework and our proposed objective function in M-step, we compare our models with two variants: **VigDet-E** and **VigDet-PL** (PL for Pseudo Likelihood). In VigDet-E, we only conduct E-step once to acquire group assignments (inferred distribution over labels) enhanced with prior knowledge, but without alternating updates using the EM loop. In VigDet-PL, we replace our proposed objective function with pseudo likelihood function from existing works.

Metrics: We compare two kinds of metrics. One kind is threshold-free: Average Precision (AP), area under the ROC curve (AUC), and maxF1 at threshold that maximizes F1 score. The other kind need threshold: F1, Precision, Recall, and MacroF1. For this kind, we apply 0.5 as threshold for the binary (coordinated/normal account) labels..

[1] https://www.recode.net/2017/11/2/16598312/russia-twittertrump-twitter-deactivated-handle-list

Table 1: Results on unsupervised coordination detection (IRA) on Twitter in 2016 U.S. Election

Method (Unsupervised)	AP	AUC	F1	Prec	Rec	MaxF1	MacroF1
Co-activity	16.9	52.5	24.6	17.8	40.7	27.1	49.5
Clickstream	16.9	53.5	21.5	20.5	22.8	21.5	53.2
IRL	20.0	61.0	26.5	21.9	33.6	34.0	54.3
HP	33.7	69.4	37.6	38.7	36.5	54.5	63.3
AMDN-HAGE	80.5	89.9	69.6	94.3	55.5	75.8	82.7
AMDN-HAGE + Kmeans	82.0	93.3	73.0	90.9	61.2	77.0	84.5
VigDet-PL(TL)	83.3	94.0	70.7	89.6	59.0	77.8	83.2
VigDet-E(TL)	85.5	94.6	73.1	**95.3**	59.4	79.5	84.6
VigDet(TL)	86.1	94.6	73.4	95.1	59.9	**79.6**	84.8
VigDet-PL(CF)	84.5	95.0	71.9	91.4	59.6	79.3	83.9
VigDet-E(CF)	85.1	94.3	73.6	92.7	61.2	78.8	84.9
VigDet(CF)	**87.2**	**95.0**	**75.2**	91.7	**63.9**	79.3	**85.7**

5.1.2 Results

Table 1 and 2 provide results of model evaluation against the baselines averaged in the IRA dataset over five random seeds. As we can see, VigDet, as well as its variants, outperforms other methods on both unsupervised and semi-supervised settings, due to their ability to integrate neural temporal point process, which is the current SOTA method, and prior knowledges, which are robust to data quality and quantity. It is noticeable that although GNN based methods can also integrate prior knowledge based graphs and representation learning from SOTA model, our model still outperforms it by modeling and inferring the distribution over group assignments jointly guided by consistency in the embedding and prior knowledge space.

Ablation Test: Besides baselines, we also compare VigDet with its variants VigDet-E and VigDet-PL. As we can see, for Cubic Filtering strategy, compared with VigDet-E, VigDet achieves significantly better result on most of the metrics in both settings, indicating that leveraging the EM loop and proposed M-step optimization can guide the model to learn better representations for E and φ_θ. As for Temporal Logic Filtering strategy, VigDet also brings boosts, although relatively marginal. Such phenomenon suggests that the performance our M-step objective function may vary with the prior knowledge we applied. Meanwhile, the VigDet-PL performs not only worse than VigDet, but also worse than VigDet-E. This phenomenon shows that the pseudolikelihood is noisy for VigDet and verifies the importance of our objective function.

5.2 Analysis on COVID-19 Vaccines Twitter Data

We collect tweets related to COVID-19 Vaccines using Twitter public API, which provides a 1% random sample of Tweets. The dataset contains 62k activity sequences of 31k accounts, after filtering accounts collected less than 5 times in the collected tweets, and sequences shorter than length 10. Although the data of tweets about COVID-19 Vaccine does not have ground-truth labels, we can apply VigDet to detect suspicious groups and then analyze the collective behavior of the group.

Detection: VigDet detects 7k suspicious accounts from the 31k accounts. We inspect tweets and account features of the identified suspicious group of coordinated accounts.

Representative tweets: We use topic mining on tweets of detected coordinated accounts and show the top representative tweets in Table 3.

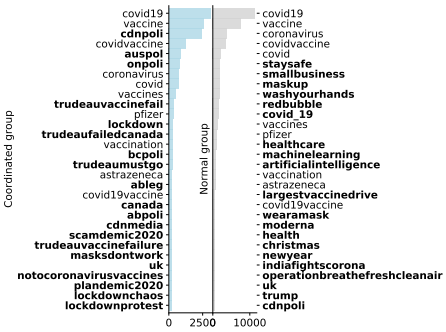


Figure 3: Top-30 hashtags in tweets of identified coordinated and normal accounts.

Table 2: Results on semi-supervised coordination detection (IRA) on Twitter in 2016 U.S. Election

Method (Semi-Supervised)	AP	AUC	F1	Prec	Rec	MaxF1	MacroF1
LPA(HP)	63.3	76.8	68.1	76.2	61.8	71.6	81.5
LPA(TL)	69.7	85.9	62.3	88.5	48.6	66.1	78.6
LPA(CF)	71.1	85.3	60.8	66.5	56.4	68.3	77.2
AMDN-HAGE + Semi-NN	77.1	87.8	70.5	76.6	65.5	72.3	82.8
AMDN-HAGE + GNN (HP)	75.5	84.0	72.0	83.0	65.1	76.6	83.7
AMDN-HAGE + GNN (CF)	80.6	89.5	73.0	86.3	63.7	76.4	84.5
AMDN-HAGE + GNN (TL)	81.3	90.2	73.6	78.2	**70.2**	77.2	84.6
VigDet-PL(TL)	87.7	95.5	73.9	94.2	61.4	80.0	85.1
VigDet-E(TL)	**88.1**	**95.7**	73.4	**94.6**	60.4	**80.8**	84.8
VigDet(TL)	88.0	**95.7**	73.6	94.2	60.9	**80.8**	84.9
VigDet-PL(CF)	85.1	95.3	69.7	93.4	55.9	79.0	82.8
VigDet-E(CF)	87.1	95.2	74.4	92.8	62.4	79.7	85.3
VigDet(CF)	87.6	95.6	**76.1**	87.2	68.1	79.8	**86.2**

Table 3: Representative tweets from topic clusters in tweets of identified coordinated accounts.

If mRNA vaccines can cause autoimmune problems and more severe reactions to coronavirus' maybe that's why Gates is so confident he's onto a winner when he predicts a more lethal pandemic coming down the track. The common cold could now kill millions but it will be called CV21/22?
This Pfizer vax doesn't stop transmission,prevent infection or kill the virus, merely reduces symptoms. So why are they pushing it when self-isolation/Lockdowns /masks will still be required. Rather sinister especially when the completion date for trials, was/is 2023
It is = You dońt own anything, including your body. - Full and absolute ownership of your biological being. - Disruption of your immune system. - Maximizing gains for #BillGatesBioTerrorist. - #Transhumanism - #Dehumanization'
It also may be time for that "boring" O'Toole (as you label him) to get a little louder and tougher. To speak up more. To contradict Trudeau on this vaccine rollout and supply mess. O'Toole has no "fire". He can't do "blood sport". He's sidelined by far right diversions.

Account features: The two groups (identified coordinated and normal accounts) are clearly distinguished in the comparison of top-30 hashtags in tweets posted by the accounts in each group (presented in Fig. 3). In bold are the non-overlapping hashtags. The coordinated accounts seem to promote that the pandemic is a hoax (#scamdemic2020, #plandemic2020), as well as anti-mask, anti-vaccine and anti-lockdown (#notcoronavirusvaccines, #masksdontwork, #livingnotlockdown) narratives, and political agendas (#trudeaumustgo). The normal accounts narratives are more general and show more positive attitudes towards vaccine, mask and prevention protocols.

Also, we measure percentage of unreliable and conspiracy news sources shared in the tweets of the detected coordinated accounts, which is 52.56%, compared to 25.6% in the normal account group. Even Twitter suspended accounts is higher in the coordinated group (4.26%) compared to 3.45% otherwise. Percentage of recent accounts (created in 2020-21) is higher in coordinated group (19%) compared to 15% otherwise. Disinformation and suspensions are not exclusive to coordinated activities, and suspensions are based on Twitter manual process and get continually updated over time, also accounts created earlier can include recently compromised accounts; therefore, these measures cannot be considered as absolute ground-truth.

6 Conclusion

In this work, we proposed a variational inference framework for prior knowledge guided neural temporal point process to detect coordinated groups on social media. Comparison experiments and ablation test on IRA dataset verify the effectiveness of our model and inference. Furthermore, we apply our model to uncover suspicious misinformation campaign in COVID-19 vaccine related tweet dataset. Behaviour analysis of the detected coordinated group suggests efforts to promote anti-vaccine misinformation and conspiracies on Twitter.

References

[1] Adam Badawy, Aseel Addawood, Kristina Lerman, and Emilio Ferrara. Characterizing the 2016 russian ira influence campaign. *SNAM*, 9(1):31, 2019.

[2] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Cascade-based community detection. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 33–42, 2013.

[3] Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

[4] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

[5] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, pages 477–488, 2014.

[6] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer, 2007.

[7] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *ACM SIGKDD*, pages 1555–1564, 2016.

[8] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216*, 2017.

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[10] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[11] Vladimir Kolmogorov and Ramin Zabin. What energy functions can be minimized via graph cuts? *IEEE transactions on pattern analysis and machine intelligence*, 26(2):147–159, 2004.

[12] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[13] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[14] Kristina Lerman, Rumi Ghosh, and Tawan Surachawala. Social contagion: An empirical study of information spread on digg and twitter follower graphs. *arXiv preprint arXiv:1202.3162*, 2012.

[15] Shuang Li, Lu Wang, Ruizhi Zhang, Xiaofu Chang, Xuqin Liu, Yao Xie, Yuan Qi, and Le Song. Temporal logic point processes. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5990–6000. PMLR, 2020.

[16] Luca Luceri, Silvia Giordano, and Emilio Ferrara. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *ICWSM*, volume 14, pages 417–427, 2020.

[17] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *NIPS*, pages 6754–6764, 2017.

[18] Radford M. Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models*, page 355–368, USA, 1998.

[19] Takahiro Omi, Kazuyuki Aihara, et al. Fully neural network based model for general temporal point processes. In *NIPS*, pages 2122–2132, 2019.

[20] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[21] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering coordinated networks on social media. 2021.

[22] V Jothi Prakash and Dr LM Nithya. A survey on semi-supervised learning techniques. *arXiv preprint arXiv:1402.4645*, 2014.

[23] Meng Qu, Yoshua Bengio, and Jian Tang. Gmnn: Graph markov neural networks. *arXiv: Learning*, 2019.

[24] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *CIKM*, pages 797–806. ACM, 2017.

[25] Karishma Sharma, Emilio Ferrara, and Yan Liu. Identifying coordinated accounts in disinformation campaigns. *arXiv preprint arXiv:2008.11308*, 2020.

[26] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM TIST*, 2019.

[27] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. Coronavirus on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309*, 2020.

[28] Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. 2020.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[30] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[31] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. Unsupervised clickstream clustering for user behavior analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 225–236, 2016.

[32] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Self-attention with functional time representation learning. In *NeurIPS*, 2019.

[33] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *ACM WebSci*, pages 353–362, 2019.

[34] Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-attentive hawkes process. 2020.

[35] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, 2013.

[36] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

[37] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

[38] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss about some limitations in the Ablation Test of Section 5.1.2. Specifically, the performance of our model may vary with the quality of the prior knowledge. Also, the M-step tends to raise the general performance but can add false positives.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] None
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs are included in the Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The IRA data is a restricted access dataset from Twitter. Therefore, we cannot directly share it. We requested it from the Authors of this work [16], and must redirect any requests to the last listed Author.
   We do provide our source code. We also provide the COVID-19 Vaccine Tweets data (collected by us using Twitter public APIs) in supplementary material, in accordance with Twitter policies. However, it does not have ground truth labels, but the detection and analysis results can be replicated and we provide details for it in the code and Appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes, we specify the data splits in the paper. Other details are included in the Appendix due to the page limits.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Due to the table size limitation, we put them in the Appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Due to the page limits, we put them in the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] In the source code
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the COVID-19 dataset (collected by us) in accordance with Twitter policies (https://developer.twitter.com/en/developer-terms/agreement-and-policy).
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The datasets used are in accordance with Twitter APIs and policies and no personally identifiable information is utilized in the model, or presented in the paper.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

Appendix is included in supplementary materials.