

Hidden Buyer Identification in Darknet Markets via Dirichlet Hawkes Process

Panpan Zheng
Amazon
Seattle WA, USA
panpz@amazon.com

Shuhan Yuan
Utah State University
Logan UT, USA
shuhan.yuan@usu.edu

Xintao Wu
University of Arkansas
Fayetteville AR, USA
xintaowu@uark.edu

Yubao Wu
Georgia State University
Atlanta GA, USA
ywu28@gsu.edu

Abstract—Darknet markets are underground markets for various illicit transactions, including selling or brokering drugs, weapons, and stolen credit cards. To combat these illicit activities in cyberspace, it is critical to understand the activity behaviors of participants in the darknet markets. Currently, many studies focus on studying the activities of vendors. However, there is no much work on analyzing buyers. The key challenge is that the buyers are anonymized in darknet markets. To ensure the anonymity of transactions, we only observe the first and last digits of a buyer's ID, such as “a**b”, on most of the darknet markets. To tackle this challenge, we propose a hidden buyer identification model, called UNMIX, which can group transactions from one hidden buyer into one cluster given a transaction sequence from an anonymized ID. UNMIX is able to model the temporal dynamics information as well as the product, comment, and vendor information associated with each transaction. Then, the transactions with similar patterns in terms of time and content are grouped as a subsequence from one hidden buyer. Experiments on the data collected from three real-world darknet markets and one DBLP publication dataset demonstrate the effectiveness of our approach measured by various clustering metrics. Case studies on real transaction sequences explicitly show that our approach can group transactions with similar patterns into the same clusters.

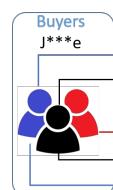
Index Terms—User identification, Dirichlet Hawkes process, Darknet market

I. INTRODUCTION

Darknet markets are online commercial websites that strongly provide privacy guarantees to both vendors and buyers. The markets are hosted in the darknet based on TOR service to hide IP addresses and adopt cryptocurrencies, such as Bitcoin, as payment methods. Due to its anonymity, most of the transactions on darknet markets are related with trading illicit goods, such as illicit drugs, stolen credit cards, or even weapons.

To combat illicit transactions in the cyberspace, it is important to analyze the behavior of participants in darknet markets. Many studies focus on studying the behavior of vendors, such as linking multiple accounts from a same vendor (Sybil accounts) [1], [2], [3]. However, there is no much work on analyzing buyers. One of the key challenges is that the buyers are anonymized in darknet markets.

In order to protect the buyers' privacy and encourage the buyers to publish their comments, most of the darknet markets only reveal the first and last digits of a buyer's ID, such as “a**b”, on the comment page. As a result, one observed anonymized ID can link to many different real-world buyers. Figure 1 shows an illustrative example of one mixed transaction sequence from the anonymized ID $J^{**}e$. Each transaction contains information about four attributes: product, date, vendor, and comment in addition to the anonymized buyer name information. Our goal is to group those mixed transactions into clusters based on both content and temporal dynamics such that each cluster contains all transactions from one particular real-world user. Such disambiguation will allow us to learn transaction patterns of darknet markets and predict future transactions. Meanwhile, the law enforcement could use the proposed techniques in this work to uncover criminals based on darknet transactions.



Buyers	Product	Date	Vendor	Comment
$J^{**}e$	40 GRAM PURE MDMA 84%	10/14/2018 06:58 pm	JOOS	Very happy!
	MDMA Crystals 90% pure - 5g	10/12/2018 08:52 pm	bionik	Good stuff A++ ...

	3.5G Peruvian Cocaine A-GRADE	10/09/2018 01:06 pm	JOSS	Trusted Vendor ...
	Dutch MDMA 87% pure - 10g	10/09/2018 09:20 pm	bionik	Straight fire ...
	2.5 Gram HIGH CLASS COCAINE	10/07/2018 04:50 pm	Narco	Shipping fast ...

Figure 1: Illustrative example of transaction sequence from anonymized ID “ $J^{**}e$ ”, where transactions are from various real-world buyers.

In this work, we propose UNMIX, a hidden buyer identification model, based on the Dirichlet-Hawkes Process (DHP) [4], which is able to group continuous-time transactions for each hidden buyer by modeling temporal dynamics, product, title, and comment of each transaction. Temporal dynamics here refers to the time patterns of purchase behavior in darknet market. Each buyer has its own temporal dynamics. For instance, buyer “Jaae” often purchases one kind of heroines on Monday night once a week while buyer “Jbbe” buys the same drug on Tuesday and Thursday (twice a week). The output from our model are clusters, each of which contains transactions from one specific buyer. The idea of our proposed model is to have the Hawkes process model the intensity rate of transactions, while the Dirichlet process captures the buyer-transaction cluster relationships (i.e., each cluster con-

tains transactions from one buyer). In darknet markets, each transaction sequence from an anonymized ID (e.g., “J***e”) consists of a few real buyers, where different real buyers tend to have different transaction patterns in terms of transaction comments, time and vendors. As a result, UNMIX, which is able to group transactions based on similar patterns, can identify hidden buyers.

UNMIX is a novel approach to achieve hidden buyer identification by integrating all the information associated with transactions, including temporal dynamics, products, comments, and vendors. The prior of each transaction belonging to one hidden buyer is determined by its temporal dynamics as different hidden buyers often exhibit different temporal dynamics. Specifically, the Hawkes process, one type of temporal point processes, is adopted to model the self-excitation phenomenon among transactions over continuous-time (e.g., buying illicit drugs in the past can raise the probability of buying them again in the future). The temporal dynamics of each identified hidden buyer is then characterized by one Hawkes process. Besides the temporal dynamics, texts in product titles and comments and vendors involved in transactions are also incorporated into our model, which are characterized by a multinomial distribution and a categorical distribution, respectively. Meanwhile, by leveraging the Dirichlet process, the proposed model complexity grows as more transactions are collected over time, so our approach allows that the number of hidden users from a mixed transaction sequence is unknown and unfixed.

The main contributions of our work are as follows. First, UNMIX does not need to assign a fixed number of hidden buyers underlying the unlimited number of transactions from one anonymized ID. Second, together with transaction content information, the temporal information provides important clues that improve accuracy in identifying hidden buyers in the same darknet markets. Third, experimental results on three real-world darknet markets indicate UNMIX is able to identify various hidden buyers with different transaction patterns.

II. RELATED WORK

A. Darknet Market Analysis

Darknet markets are online markets hosted on the Tor service and guarantee strong anonymity property to participants. As a result, the darknet markets involve in illegal online transactions. For the sake of public interests, the authorities and researchers have a growing interest to understand the darknet markets. Researchers have collected a large amount of data from darknet markets to analyze the active vendors, buyers, and goods being sold over time so that we can understand the growth of the darknet market ecosystem [5], [6], [7]. Research in [8] also conducts empirical studies to understand the supply chain underlying the markets. Besides analyzing the volumes of whole darknet markets, some studies analyze specific categories in the darknet markets, especially illicit drugs. For example, research in [9] investigates the structure and organization of illicit drug trafficking. Research in [10] describes the motives, perspectives and purchasing experiences

of a darknet market, ‘Silk Road’, users to discover how the darknet market operates. Research in [11] further compares the activities between wholesalers with retail-level distributors on the darknet markets. Some researchers analyze the drugs sold in darknet markets and compare the coherence between digital and physical information [12]. Since the darknet markets have strong correlations with cybercrime, research in [13] focuses on measuring the commoditization of cybercrime via darknet markets.

Many researchers target on the micro-level analysis, which studies the participants in the darknet markets. Due to the anonymity of darknet markets, the challenge of analyzing the behavior of participants in the darknet market is how to link user identities in the markets. Recently, several studies aim to link multiple accounts created by a real-world vendor [3], [1], [2], [14]. The key idea of these studies is based on “stylometry” analysis, which is originally used to attribute authorship to anonymous documents. For example, research in [3] links multiple vendors by analyzing the styles of the product pictures. Based on that, a follower study further considers the product descriptions information as well as the pictures to link multiple vendors [2]. Research in [1] combines random forest classifiers and hierarchical clustering on a set of features to detect whether a pair of vendors is a match or not. Unlike matching vendors that can adopt lengthy product descriptions and photos, the information that can be used for identifying hidden buyers is very limited. To the best of our knowledge, how to identify hidden buyers in the darknet markets has not been studied in the literature.

B. Stream Clustering

Identifying hidden buyers from a mixed transaction sequence can be viewed as a task of stream clustering. Various approaches have been proposed for stream clustering, including partition-based [15], [16], density-based [17], and probability-based approaches [18]. In recent years, probabilistic stream clustering methods have attracted increasing attention. Concretely, the widely-used models for stream clustering from the topic modeling literature are the Latent Dirichlet Allocation (LDA) [19], where the number of topics is fixed. To remove such restriction, researchers extend the LDA model by incorporating the Dirichlet process and propose Hierarchical Dirichlet Process (HDP) [18] that is able to model texts with an unbounded number of topics. Some studies further extend HDP model to nested Hierarchical Dirichlet Process, where adopt the Dirichlet process as the base distribution of another Dirichlet process for several levels [20], [21]. Many models are further proposed to fit the scenarios with online streaming text data [22], [23], [24]. Recently, several studies further incorporate temporal dynamics to group streaming data [4], [25], [26], [27]. For example, the Dirichlet Hawkes Process adopts the temporal point process, e.g., Hawkes process, to model the continuous-time information and the Dirichlet Process to solve the clustering problems [4]. Research in [27] further combines the hierarchical Dirichlet process with temporal point process to model the learning activity on the web. In our work, given

a mixed transaction sequence from various hidden buyers, we propose a novel probabilistic stream clustering algorithm for hidden buyer identification using the Hawkes process to model the temporal dynamics, the multinomial distribution to model texts in product titles and comments, and the categorical distribution to model vendors involved in transactions.

C. Name Disambiguation

Our hidden buyer identification is also related with name disambiguation, also named as entity resolution or name identification. Name disambiguation has been studied in past decades (refer to survey papers [28], [29]) and has many real applications, e.g., identifying users across multiple social networks or separating authors with the same name. The state-of-the-art solutions for name disambiguation problem include two categories: feature-based and linkage-based. The former leverages supervised learning methods to use feature vectors of documents to learn a pairwise similarity function whereas the latter uses co-author relationship to build document graph for each ambiguous name, utilizes graph topology information (node/edge features or embeddings). Our method proposed in this paper focuses on sequence information and considers both physical time, type, and text information of each event.

III. PRELIMINARY

A. Dirichlet Process

The Dirichlet Process (DP) is a Bayesian nonparametric model, which is parameterized by a concentration parameter $\alpha > 0$ and a base distribution G_0 over a space Θ . It indicates that a random distribution G drawn from DP is a distribution over Θ , denoted as $G \sim DP(\alpha, G_0)$. The expectation of the distribution G is the base distribution G_0 . The concentration parameter α controls the variance of G that a larger α leads to a tighter distribution around G_0 . DP is widely used for clustering with the unknown number of clusters.

The Dirichlet Process can also be represented as the Chinese Restaurant Process (CRP). CRP assumes a restaurant with an infinite number of tables, and each of the tables can seat an infinite number of customers. Within the context of clustering, each table indicates a cluster while each customer is a data point. The simulation process of CRP is as follows:

- 1) The first customer always sits at the first table.
- 2) Customer n ($n > 1$) sits at:
 - a) a new table with probability $\frac{\alpha}{\alpha+n-1}$.
 - b) an existing table h with probability $\frac{n_h}{\alpha+n-1}$ where n_h is the number of customers at table h .

Let $\{\theta_1, \dots, \theta_n\}$ be a sequence sampled from CRP. The conditional distribution of θ_n can be written as:

$$\theta_n | \theta_{1:n-1} \sim \frac{1}{\alpha + n - 1} (\alpha G_0 + \sum_h n_h \delta_{\theta_h}), \quad (1)$$

where δ_{θ_h} is a point mass centred at θ_h . Equation 1 indicates that a new sample θ_n belongs to a new table with a constant probability or an existing table h with probability proportional to n_h . A larger n_h indicates a higher probability that a

customer will belong to the table h . Hence, DP has a clustering property that the rich gets richer.

B. Temporal Point Process

Temporal point process is a random process that models the observed random event patterns along the time. Given an event time sequence $\mathcal{T} = \{t_1, \dots, t_n\}$, a temporal point process can be characterized by the *conditional intensity function* which indicates the expected instantaneous rate of the next event at time t ($t > t_n$):

$$\lambda^*(t) = \lambda(t | \mathcal{H}_{t_n}) = \lim_{dt \rightarrow 0} \frac{\mathbb{E}[N([t, t+dt)) | \mathcal{H}_{t_n}]}{dt}, \quad (2)$$

where $N([t, t+dt))$ indicates the number of events occurred in a time interval dt ; $\mathcal{H}_{t_n} = \{t_i | t_i \leq t_n\}$ is the collection of historical events until time t_n .

Let $f^*(t) = f(t | \mathcal{H}_{t_n})$ be the conditional density function of the event happening at time t given the historical events up to time t_n , which is defined as

$$f^*(t) = \lambda^*(t) \cdot S^*(t) = \lambda^*(t) \cdot \exp\left(-\int_{t_n}^t \lambda^*(\tau) d\tau\right), \quad (3)$$

where $S^*(t) = S(t | \mathcal{H}_{t_n}) = \exp(-\int_{t_n}^t \lambda^*(\tau) d\tau)$ is the *survival function* that indicates the probability that no new event has ever happened up to time t since t_n .

With an observation window $[0, T]$, the joint likelihood of the observed sequence \mathcal{T} is formalized as

$$\mathcal{L} = \prod_{t_i \in \mathcal{T}} f^*(t_i) = \prod_{t_i \in \mathcal{T}} \lambda^*(t_i) \cdot \exp\left(-\int_0^T \lambda^*(\tau) d\tau\right). \quad (4)$$

Hawkes process. A Hawkes process is one type of temporal point processes and captures the self-excitation phenomenon among events [30]. In the Hawkes process, the conditional intensity function is defined as:

$$\lambda^*(t) = \lambda_0 + \sum_{t_i \in \mathcal{T}} \gamma(t, t_i), \quad (5)$$

where $\lambda_0 > 0$ is the base intensity that indicates the intensity of events triggered by external signals instead of previous events; $\gamma(t, t_i)$ is the triggering kernel that is usually a monotonically decreasing function which ensures the recent events have higher influences on the intensity of next event. The Hawkes process models the self-excitation phenomenon that a new event arrival increases the conditional intensity of the upcoming event immediately and then decreases back towards λ_0 in the long term. Recently, the Hawkes process is widely used to model event patterns which are clustered, such as the information diffusion on social networks or the earthquake occurrences [31], [32], [33].

IV. HIDDEN BUYER IDENTIFICATION

In a darknet market, a buyer purchases products from vendors and then publishes comments about the products. Especially, we can't see the *real user names* of buyers. Instead, what we can observe are some anonymized IDs, each of which contains an unbounded number of real buyers. Given a

sequence of transactions marked by one specific anonymized ID, our goal is to uncover the real buyers by grouping the transactions into subsequences, where each subsequence is from a real buyer. In our scenario, these distinctive real buyers are named as *hidden buyers*. Given a sequence of transactions $\mathcal{S} = \{e_1, \dots, e_n\}$ underlying one specific anonymized ID, its corresponding sequence of real buyers is denoted as $\mathcal{U} = \{u_1, \dots, u_n\}$ with one set of real buyers as $\{u_i\}$. Then, the *hidden buyer* associated with one transaction e is expressed as $u^* \in \{u_i\}$.

Formally, transaction e in \mathcal{S} is denoted as $e := (t, u, v, p, c)$, which means that at time t , a buyer u purchases a product p from a vendor $v \in \mathcal{V}$, where $\mathcal{V} = \{v_1, \dots, v_n\}$ is the corresponding vendor sequence, and publishes a comment c . Since product titles and comments are both text information, we further combine them as a *content* vector \mathbf{w} by a bag of word model. Finally, we define one transaction in \mathcal{S} as $e := (t, u, v, \mathbf{w})$. Note that since we only observe the time to publish a comment, in our scenario, we assume the operations, purchasing a product and publishing a comment, are synchronous.

To identify hidden buyers, we assume that different hidden buyers have their unique hidden transaction patterns. For example, buyer \mathcal{A} always buys fentanyl from one certain vendor without comments, while buyer \mathcal{B} often takes fentanyl from the same vendor as well but likes to leave the comments. Given this toy example, we are wondering if transactions with a similar purchasing pattern are associated with the same hidden buyer. To further explore and solve this problem, in this work, we aim to uncover the mixed transaction sequence marked by one anonymized ID and propose a novel identification framework named as UNMIX.

UNMIX is a Dirichlet process framework with Chinese restaurant process as implementation. In UNMIX, each table encapsulates a marked Hawkes process model, which is for time and type information, and a bag-of-words model, which is for textual comment and title information. Here, each table corresponds to a real hidden buyer in our scenario. For one specific transaction, its hidden buyer assignment is based on a discrete probability distribution that is derived by a posterior predictive distribution. The estimated likelihoods are related to the historical transactions from these hidden buyers. Hence, transactions with the similar patterns are easily going to the same hidden buyer, and an upcoming transaction tends to be assigned to a hidden buyer (table) where the majority of previous transactions (customers) are similar.

A. Modeling Buyer Transactions

From the perspective of features, we consider three categories of information: time, content (product titles and comments) and vendor. Each of them has its own distinctive characteristics and should be captured by different models. For instance, due to the drug addiction effects, once a user starts to purchase illicit drugs, he may keep purchasing constantly in a short period of time. Since the behavior of purchasing drugs is self-exciting, it is natural to adopt the Hawkes process to

model the purchasing behavior in terms of time. Meanwhile, vendor and content information are characterized by categorical and multinomial distributions, respectively. Given the unbounded number of hidden buyers in a dynamic transaction sequence, we adopt the Dirichlet process as a prior probability distribution to model the generation of hidden buyers.

Generally, UNMIX is a hierarchical framework with two layers: in the outer layer, it employs Dirichlet process to capture the diversity of hidden transaction patterns for distinctive hidden buyers; in the inner layer (inside the hidden buyers), it makes use of Hawkes process, multinomial distribution and categorical distribution to model the time, content and vendor information, respectively.

Intensity of the buyer transaction activity. We adopt the Hawkes process to model buyer transactions over time. In our scenario, the sequence of transactions with the same anonymized ID are actually conducted by different hidden buyers. For each hidden buyer, we adopt one Hawkes process to model its temporal information. As a result, the intensity function of Hawkes process over the whole transaction sequence from all of existed hidden buyers is defined as:

$$\lambda(t) = \lambda_0 + \sum_{h=1}^H \lambda_h(t), \quad (6)$$

where H is the total number of identified hidden buyers until time t .

$\lambda_h(t)$ is the intensity of one certain hidden buyer h and it can be expressed as follow:

$$\lambda_h(t) = \sum_{t_i \in \mathcal{T}} \gamma_h(t, t_i) \mathbb{1}[u_i = u_h^*], \quad (7)$$

where $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ is the corresponding event time sequence of \mathcal{S} ; $\gamma_h(t, t_i)$ is the triggering kernel associated with one hidden buyer u_h^* ; u_i is the index of hidden buyer associated with the i -th transaction, and $\mathbb{1}[u_i = u_h^*]$ denotes the i -th transaction has been assigned to the h -th buyer in Chinese restaurant process. Here, the triggering kernel function with K base kernel functions is in the form as $\gamma_h(t, t_i) = \sum_{l=1}^K \alpha_h^l \kappa(\pi_l, t - t_i)$, where $\alpha_h^l > 0$ controls the self-excitation of the Hawkes process with $\sum_l \alpha_h^l = 1$, and π_l is typical reference time point that controls the event decay. We adopt the Gaussian RBF kernel as the base kernel function.

Distribution of content information (product titles and comments). Since both product titles and comments are text information, we represent them as a bag-of-word language model. We call both the product title and comment in a transaction as the *content* of the transaction. As a result, we use a vector \mathbf{w}_i to represent the content in transaction e_i , where each dimension refers to the frequency of the corresponding word sampled from a vocabulary \mathcal{W} . In particular, \mathbf{w}_i provided by hidden buyer h , is described as follow

$$\mathbf{w}_i \sim \text{Multi}(\theta_h), \quad (8)$$

where θ_h is the prior of multinomial distribution with size $|\mathcal{W}|$, which indicates the occurrence likelihood of each word in the content given the hidden buyer u_h^* .

Distribution of vendors. In this work, we use vendor ID to indicate each vendor. Due to its discreteness property, at each time t_i , the vendor is sampled from a categorical distribution with the sample space size as $|\mathcal{V}|$:

$$v_i \sim \text{Cat}(\eta_h), \quad (9)$$

where η_h is the prior of categorical distribution with size $|\mathcal{V}|$, which refers to the occurrence likelihood of each vendor given the hidden buyer u_h^* .

B. The Generative Process

We can describe our model as a generative process similar to CRP. At time t , the upcoming transaction e may be from either a new buyer or an existing buyer. To give a proper hidden buyer assignment of event e , our proposed framework UNMIX, which is running on a Dirichlet process, will dynamically reuse an existing hidden buyer or generate a new one to adapt the upcoming event e . Concretely, hidden buyer u of the upcoming event can be chosen in a metropolis sampling-based way

$$u = \begin{cases} u_{H+1}^* & \text{with probability } \frac{\lambda_0}{\lambda(t)} \\ u_h^* & \text{with probability } \frac{\lambda_h(t)}{\lambda(t)} \end{cases}, \quad (10)$$

where H is the number of hidden buyers up to but not including time t ; $\lambda_h(t)$ indicates the intensity of a Hawkes process for the hidden buyer u_h^* defined in Equation 7. Meanwhile, λ_0 plays the similar role as the concentration parameter α in DP and the probability of u belonging to u_h^* is proportional to the intensity function $\lambda_h(t)$ from a Hawkes process.

The algorithm of the generative process is shown in Algorithm 1, where λ_0 is the base intensity, α_0 is the initial parameter setting of trigger kernels in Equation 7, η_0 and θ_0 are the initial prior for the categorical and multinomial distributions. For the transaction e_i , Line 2 first samples time t_i via a Hawkes process. Based on temporal dynamics of historical events, Line 3 chooses a proper hidden buyer for the current event at time t_i . If u_i belongs to the hidden buyer u_h^* , Line 5 reuses α_h , θ_h and η_h for α_i , θ_i and η_i as the parameters for the Hawkes process, multinomial and categorical distributions; otherwise, Line 7 samples α_i , θ_i and η_i from Dirichlet distributions. Given the priors (α_i , θ_i and η_i), Lines 8 and 9 illustrate how to draw the corresponding content and vendor information. The output of the algorithm is a generated transaction sequences with H hidden buyers.

C. Inference

Given a transaction sequence $\mathcal{S} = \{e_1, \dots, e_{n-1}\}$ from an anonymized ID, our target is to infer the unique hidden buyer u_h^* for the upcoming transaction e_n . According to the probabilistic graphical model shown in Figure 2, we can formulate the sequential posterior of latent variable hidden buyer for the upcoming transaction (t_n, w_n, v_n) as follows

$$\begin{aligned} P(u_n|t_n, w_n, v_n, \text{rest}) &\sim \\ P(v_n|u_n, \text{rest}) \cdot P(w_n|u_n, \text{rest}) \cdot P(u_n|t_n, \text{rest}). \end{aligned} \quad (11)$$

Algorithm 1: The generative process of UNMIX

Input : $\lambda_0, \alpha_0, \theta_0, \eta_0$
Output: $\{e_i := (t_i, u_i, v_i, w_i)\}_{i=1}^N$ where N is the total number of transactions produced by the generative process algorithm.

- 1 **for** $i = 1, \dots, N$ **do**
- 2 Sample the time $t_i \sim \text{Hawkes}(\lambda^*(t_i))$;
- 3 Sample the hidden buyer u_i of transaction e_i by Eq. 10 ;
- 4 **if** $u_i == u_h^*$ **then**
- 5 Reuse η_h and θ_h for η_i and θ_i ;
- 6 **else**
- 7 Sample η_i from $\text{Dir}(\eta|\eta_0)$, θ_i from $\text{Dir}(\theta|\theta_0)$, and α_i from $\text{Dir}(\alpha|\alpha_0)$ for the new user ;
- 8 Sample each word w_i in the content of transaction e_i by Eq. 8 ;
- 9 Sample the vendor v_i of transaction e_i by Eq. 9 ;
- 10 **end**

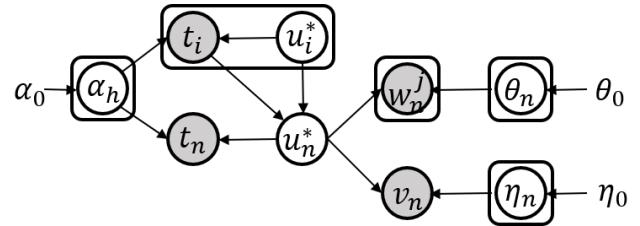


Figure 2: Graphical representation of UNMIX

where *rest* refers to the transaction sequence until t_{n-1} . In Equation 11, the prior $P(u_n|t_n, \text{rest})$ is given by:

$$P(u_n|t_n, \text{rest}) = \begin{cases} \frac{\lambda_0}{\lambda(t)} & \text{for new buyer} \\ \frac{\lambda_h(t)}{\lambda(t)} & \text{for observed buyer } u_h^* \end{cases} \quad (12)$$

where $\lambda_h(t) := \sum_{t_i \in \mathcal{T}} \gamma_h(t, t_i) \mathbb{1}[u_i = u_h^*]$ indicates the intensity from buyer u_h^* .

Dirichlet-Multinomial distribution is employed to formulate the content information. Based on the conjugate relation between the multinomial and Dirichlet distributions, the likelihood $P(\mathbf{w}_n|u_n, \text{rest})$ is given by

$$\begin{aligned} &P(\mathbf{w}_n|u_n, \text{rest}) \\ &= \frac{\Gamma(C^{u_n} + 1)}{\prod_w^{\mathcal{W}} \Gamma(C_w^{u_n} + 1)} \cdot \frac{\Gamma(C^{u_n \setminus w_n} + \sum_w^{\mathcal{W}} \theta_0^w)}{\prod_w^{\mathcal{W}} \Gamma(C_w^{u_n \setminus w_n} + \theta_0^w)} \\ &\quad \frac{\prod_w^{\mathcal{W}} \Gamma(C_w^{u_n \setminus w_n} + C_w^{w_n} + \theta_0^w)}{\Gamma(C^{u_n \setminus w_n} + C^{w_n} + \sum_w^{\mathcal{W}} \theta_0^w)}, \end{aligned} \quad (13)$$

where $C^{u_n \setminus w_n}$ and $C_w^{u_n \setminus w_n}$ refer to the total word count and the count of word w appeared in the content from buyer u_n excluding w_n , respectively; C^{w_n} and $C_w^{w_n}$ refer to the total word count and the count of word w in content w_n , respectively; θ_0^w is the value in Dirichlet prior for word w .

Table I: Statistics of three darknet markets

Darknet Markets	Vendors	Anonymized Buyer IDs	Transactions
Wall Street Market	440	1896	18603
Empire Market	273	1492	12937
Dream Market	606	2587	102378

For the vendor information, we make use of Dirichlet-Categorical distribution to formulate the likelihood $P(v_n|u_n, rest)$

$$P(v_n = v|u_n, rest) = \frac{C_v^{u_n \setminus v_n} + \eta_0^v}{C^{u_n \setminus v_n} + \sum_{v'} \eta_0^{v'}}, \quad (14)$$

where $C_v^{u_n \setminus v_n}$ is the count of the vendor v from unique buyer u_n excluding the current vendor v_n ; $C^{u_n \setminus v_n}$ is the total number of vendors associated with the buyer u_n excluding the current vendor v_n ; η_0^v is the Dirichlet prior for vendor v .

In this work, we adopt the Sequential Monte Carlo (SMC) [34] to infer the sequence posterior, in which a set of particles are maintained and each of them represents a hypothesis of one hidden buyer. Additionally, we make use of $P(u_n|u_{n-1}, t_{1:n}, \mathbf{w}_{1:n}, v_{1:n})$ as the proposal distribution to minimize the variation of particle weights that are used to measure how well the particle's hypothesis can explain the data.

We follow the literature [35], [36] and update α_h^l by maximum likelihood estimation (Equation 4). In addition, following the work [4], we also set up a constant observed window in online inference. In other words, for the hidden buyer sampling and triggering kernel parameter updating, we only consider the transactions in the observed window and skip the ones which are far away.

V. EXPERIMENTS

A. Datasets and Baselines

Datasets. To evaluate our approach, we have crawled the data from three popular darknet markets, i.e., *Dream Market*, *Wall Street Market*, and *Empire Market*. The statistics of the crawled darknet markets are shown in Table I. Dream Market is the most popular darknet market and has the largest number of vendors and transactions among these three darknet markets. Meanwhile, due to the anonymity of buyer IDs, all three darknet markets have a similar number of anonymized IDs.

Note that in the Dream Market, buyers comment on vendors instead of products. Hence, for the Dream Market, we only adopt texts from comments as the content information.

Baselines. We compare our approach with two baselines.

- Hierarchical Dirichlet Process (**HDP**) is a nonparametric Bayesian approach for topic modeling [18]. We adopt DBSCAN to group transactions, each of which is represented as a topic distribution. HDP only considers the information of product titles and buyer comments.
- Dirichlet Hawkes Process (**DHP**) [4] is a simplified version of our approach and does not adopt the vendor information for clustering.

Implementation details. In all experiments, we set the base intensity of Hawkes process $\lambda_0 = 0.1$, the number of RBF

Table II: Results of hidden buyer identification on transaction sequences without ground-truth

	Approaches	C_v	Silhouette
Wall Street Market	HDP	0.4941	-0.0940
	DHP	0.7361	-0.0040
	UNMIX	0.7668	0.0063
Empire Market	HDP	0.4749	-0.0784
	DHP	0.6659	-0.0154
	UNMIX	0.6726	0.0026
Dream Market	HDP	0.5367	0.1101
	DHP	0.6667	0.1202
	UNMIX	0.6735	0.1439

Kernels $K = 4$ and the corresponding reference time points τ_l being 2, 7, 14, 28, respectively. The hyperparameters α_h , θ_h and η_h are respectively initialized and updated by three symmetric Dirichlet distributions which are parameterized by α_0 , θ_0 and η_0 and whose concentration parameter values are 0.1, 0.01 and 0.01. The code is available at Github¹.

B. Experiments on Transaction Sequences

Experimental setup. We apply our algorithm on the transaction sequences from various anonymized IDs. For each darknet market, we select anonymized IDs with at least 50 transactions. We then have 28, 16, and 579 anonymized IDs for Wall Street Market, Empire Market, and Dream Market.

We adopt the *Silhouette coefficient* (Silhouette) and the *topic coherence* (C_v) to measure the consistency of clustering results [37], [38]. Both of these metrics evaluate the clustering performance without ground-truth. Originally, topic coherence evaluates topic models via top-k topic words. In this work, we extract the top-k frequent words from each cluster and evaluate their coherence. If the transactions in a cluster have high coherence in product titles and comments, we can then reasonably consider that transactions from the same cluster are conducted by one hidden buyer. The metric of topic coherence is implemented by Gensim². For Silhouette coefficient, we use the word distribution as the feature vector for each transaction. We report the mean value of each metric over various anonymized IDs in each market.

Experimental results. As shown in Table II, UNMIX achieves the best performance in terms of topic coherence and Silhouette coefficient. Specifically, compared with DHP that does not adopt the vendor information, our approach achieves higher values in C_v and Silhouette coefficient, which indicates the usefulness of incorporating vendor information for hidden buyer identification. HDP has the lowest values of C_v and Silhouette coefficient among three approaches over three datasets. This is because HDP does not capture the temporal dynamics information. Since the intensity of the transaction is critical for hidden buyer identification, ignoring the temporal information in modeling could lead to poor performance.

Case study. Figure 3 shows an instance of the hidden buyer identification. Given a transaction sequence with 94 transactions from an anonymized ID “s**y” in Empire Market, our

¹<https://github.com/PanpanZheng/UNMIX>

²<https://radimrehurek.com/gensim/index.html>

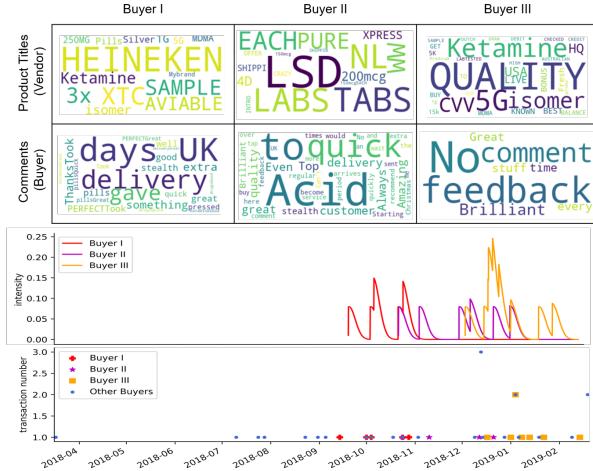


Figure 3: The hidden buyers identified from an anonymized ID “s**y” in Empire Market. The first and second rows show the frequent words in product titles and comments associated with each hidden buyer, respectively. The third row shows the intensity values of the 3 hidden buyers, while the bottom row shows the distribution of transactions over time.

proposed approach detects 22 hidden buyers. We show the top 3 hidden buyers who have the highest transaction numbers, i.e., 21, 13, 13. The first and second rows show the top words in product titles and comments from these 3 hidden buyers, respectively. The third row shows the intensity values of the 3 hidden buyers. The last row shows the distribution of transactions from “s**y” over time. We can observe that the transactions from these 3 hidden buyers roughly spread over time. In particular, the time ranges of transactions from Buyers I, II and III are 2018-09-14 to 2018-11-03, 2018-10-24 to 2019-01-22, and 2018-12-13 to 2019-02-14, respectively. The intensities of the three identified hidden buyers also lie in these areas. Meanwhile, the products bought by the three buyers are different. For example, Buyer I buys products with the frequent word “HEINEKEN” in titles, while Buyers II and III buy products with frequent words “LSD”, and “Ketamine”, respectively. Although “Ketamine” appears in product titles bought by both Buyers I and III, the two buyers have different comment styles, i.e., Buyer I prefers to write detailed comments while Buyer III seldom comments on the products. Moreover, from the time aspect, Buyers I and III are active in different months. Overall, we can notice that the three hidden buyers detected from the anonymized ID “s**y” have different styles in time, product or comment perspective.

C. Experiments on Transaction Sequences with Ground-truth

Experimental setup. Due to the anonymity of darknet markets, it is infeasible to get the ground-truth regarding the actual buyers with the same anonymized ID. To quantify the performance of our proposed approach, we propose a procedure to generate transaction sequences with ground-truth. Specifically, based on our observations, transactions conducted by one anonymized ID from one vendor in a short time are very likely from one single real-world buyer due

Table III: Results of hidden buyer identification on transaction sequences with ground-truth

Dataset (length, # of IDs)	Approach	ARS	NMI	V-score	H-score	# of clusters
Wall Street Market (42,6)	HDP	0.1612	0.3380	0.3316	0.2777	3
	DHP	0.9675	0.9804	0.9802	0.9999	7
	UNMIX	0.9385	0.9627	0.9621	1.000	8
Empire Market (188,27)	HDP	0.0422	0.2537	0.2197	0.1464	7
	DHP	0.4874	0.8282	0.8281	0.8192	41
	UNMIX	0.5236	0.8549	0.8549	0.8588	44
Dream Market (229,36)	HDP	0.0215	0.3127	0.2773	0.1896	10
	DHP	0.1391	0.6171	0.6151	0.5697	45
	UNMIX	0.1831	0.6881	0.6878	0.6707	59

to the consistent transaction behavior. Therefore, for each darknet market, we first select H anonymized IDs, where each anonymized ID has around five to eight transactions from one vendor in a month. Then, we combine all the transactions from these H anonymized IDs to compose one transaction sequence and sort the sequence by transaction time. Hence, in this setting, we generate one transaction sequence for each darknet market, while each transaction sequence is actual from various anonymized IDs. We expect the ideal algorithm can group transactions from one anonymized ID into one cluster. The statistics of transaction sequences with ground-truth are shown in the first column of Table III.

We evaluate the performance by four clustering metrics, *adjusted rand score (ARS)*, *normalized mutual information score (NMI)*, *V-measure score (V-score)*, and *homogeneity score (H-score)*, all of which are computed by comparing with ground-truth labels.

Experimental results. Table III shows the clustering results on various transaction sequences. Overall, by incorporating the content, vendor, and time information for hidden buyer identification, UNMIX achieves best performance in terms of various clustering metrics. HDP has the worst performance over three datasets, which indicates that the temporal information is critical to identify hidden buyers. DHP achieves a significant performance boost compared with HDP since DHP involves the temporal information. UNMIX outperforms DHP, which demonstrates the importance of vendor information for hidden buyer identification. Meanwhile, we can observe that the performance of three approaches are reduced when the sequences become complicated. For example, UNMIX achieves the highest scores in Wall Street Market and the lowest scores in Dream Market. This is because the sequence of Wall Street Market is simple with sequence length 42 and 6 anonymized IDs, while the sequence of Dream Market has length 229 and 36 anonymized IDs. Another reason is that we do not have product title information in Dream Market.

We notice that for Wall Street Market, DHP achieves a slightly better performance than UNMIX. This is because the number of hidden buyers identified by DHP is close to the ground truth number. However, we argue that although we combine the sequence from different anonymized IDs to compose the sequence, such sequence is only weakly-labeled since the short sequence from one anonymized ID could be actually from various hidden buyers. Based on our observation, our approach groups the subsequence from one anonymized

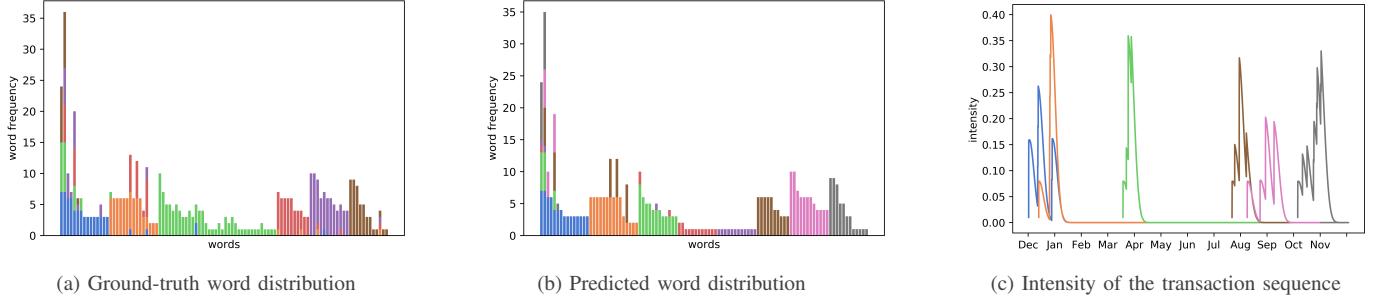


Figure 4: Word and intensity distributions of the transaction sequence from Wall Street Market. Each color indicates a detected hidden user.

ID into three clusters. However, these three hidden buyers do not share any common words in product titles and comments, which indicates the identified hidden buyers have different patterns such that they buy different products and have different comment styles. The identified three hidden buyers based on our approach are likely three different real world buyers from the content aspect.

Visualization. We further show the visualization results on the transaction sequences from Wall Street Market to illustrate the effectiveness of the proposed approach. We investigate our approach for hidden buyer identification from content and temporal dynamics aspects. To show the content information, we select the top 15 words from each predicted hidden buyer (cluster) and compare the word distributions between the ground-truth and predicted hidden buyers. Figures 4a and 4b show the word distribution of the sequence from Wall Street. Each color indicates the word distribution of one anonymized ID, while each bar indicates one word. We can observe that word distributions of predicted hidden buyers are very close to those of ground-truth, which indicates the importance of adopting content information for hidden buyer identification.

To show the information of temporal dynamics, we plot the intensity values of six identified hidden buyers (λ_h) over time. For the other two identified hidden buyers, since each of them only has one transaction, we omit their intensity curves for simplicity. We observe that these six hidden buyers are active at different months. Meanwhile, due to the self-excitation property of Hawkes process, once a transaction occurs, the intensity increases. Hence, when a hidden buyer becomes active, the following transactions have high chance to be from the same hidden buyer based on Equation 10.

D. Experiments on DBLP Publication Sequences with Ground-truth

We further adopt the *Name Disambiguation Dataset* [39] to conduct evaluation.

Experimental setup. We focus on identifying unique authors from a publication sequence with the same author name. We consider each publication as a transaction, each author ID as a hidden buyer, and each sequence of papers with the same author name as a transaction sequence. For each publication, we treat its venue as mark and paper title as text content. Our goal is to detect the true author ID for the upcoming

Table IV: Results of online name disambiguation on the DBLP dataset

Name (length, # of authors)	Approach	ARS	NMI	V-score	H-score	# of clusters
M. Yang (52,16)	HDP	0.0180	0.2609	0.1275	0.0681	2
	DHP	0.0268	0.6902	0.6796	0.8241	36
	UNMIX	0.0206	0.7323	0.7166	0.9025	42
L. Zhao (70,22)	HDP	0.0070	0.1719	0.0574	0.0295	2
	DHP	0.0755	0.6550	0.6493	0.7473	35
	UNMIX	0.035	0.6814	0.6726	0.8013	45
Kun Zhang (306,28)	HDP	0.0022	0.1384	0.1374	0.1559	8
	DHP	0.1268	0.2904	0.2853	0.3507	34
	UNMIX	0.0965	0.3495	0.3332	0.4775	42
Average	HDP	0.0393	0.2014	0.1757	0.1491	-
	DHP	0.1382	0.4298	0.4245	0.4549	-
	UNMIX	0.1517	0.4861	0.4808	0.5474	-

paper in an online manner. The dataset contains 29 author names, which lead to 29 sequences. We show the statistics and detection results of three author names, M. Yang, L. Zhao, and Kun Zhang, in Table IV. For example, there are 52 papers published by 16 authors under the name M. Yang.

Experimental results. We report in the last row of Table IV the average detection results of 29 sequences. We can see UNMIX performs much better than the two baselines in terms of all four metrics. For example, the H-score from UNMIX is 0.5474 and is much better than HDP (0.1491) and DHP (0.4594). For three specific authors, UNMIX outperforms the two baselines in terms of NMI, V-Score and H-Score. For example, UNMIX achieves H-score 0.9025, which is much better than HDP (0.0681) and DHP (0.8241).

As shown in the last column of Table IV, we see the number of detected clusters (unique author IDs) by each model. We can observe that the number of clusters from HDP tends to be less than the number of real unique authors, while the numbers of clusters from UNMIX and DHP are usually more than the number of real unique authors. Furthermore, the unique authors detected by UNMIX are more than the one by DHP. The reason is that UNMIX can capture venue information in its modeling process while DHP cannot.

We would point out that it is beyond the scope of this work to compare the proposed algorithm with solutions developed for name disambiguation. As discussed in Section II-C, existing solutions for name disambiguation problem often use feature vectors derived from abstracts (or whole papers) and/or exploit graph topology information based on co-author relationship. Our method focuses on how to model both physical time and type information in sequence and identify individuals in an online manner. We consider the comparison

with name disambiguation solutions as our future work.

VI. CONCLUSIONS

In this paper, we have developed UNMIX for hidden buyer identification in darknet markets by grouping continuous-time transactions for each hidden buyer. Due to the unfixed number of hidden buyers, UNMIX adopts the Dirichlet process to group transactions from one hidden buyer into a cluster. In order to capture the hidden behavior of different buyers, UNMIX uses the Hawkes process to model the transaction time information, the multinomial distribution to model the text information in product titles and comments, and the categorical distribution to model vendors involved in transactions. Experimental results on three darknet markets show that UNMIX achieves the best performance for hidden buyer identification. The case studies also indicate that different hidden buyers identified by UNMIX have different behaviors. In the future, we plan to study how to incorporate buyer ratings into the framework to improve the performance of hidden buyer identification. We also plan to investigate linking hidden buyers across different darknet markets.

ACKNOWLEDGEMENTS

This work was conducted when Panpan Zheng was a PhD student at the University of Arkansas. Dr. Shuhan Yuan was supported in part by NSF 2103829. Dr. Xintao Wu was supported in part by NSF 1564250, 1937010, and 1946391. Dr. Yubao Wu was supported in part by NSF 2039949 and 2030636, and a DHS grant entitled “Open Source Intelligence in Online Stolen Data Markets - Assessment of Network Disruption Strategies”.

REFERENCES

- [1] X. H. Tai, K. Soska, and N. Christin, “Adversarial matching of dark net market vendor accounts,” in *KDD*, 2019, pp. 1871–1880.
- [2] Y. Zhang, Y. Fan, W. Song, S. Hou, Y. Ye, X. Li, L. Zhao, C. Shi, J. Wang, and Q. Xiong, “Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network,” in *WWW*, 2019.
- [3] X. Wang, P. Peng, C. Wang, and G. Wang, “You are your photographs: Detecting multiple identities of vendors in the darknet marketplaces,” in *ASIACCS*, 2018.
- [4] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song, “Dirichlet-hawkes processes with applications to clustering continuous-time document streams,” in *KDD*, 2015.
- [5] N. Christin, “Traveling the silk road: A measurement analysis of a large anonymous online marketplace,” in *WWW*, 2013.
- [6] K. Soska and N. Christin, “Measuring the longitudinal evolution of the online anonymous marketplace ecosystem,” in *USENIX*, 2015.
- [7] S. Miller, A. El-Bahrawy, M. Dittus, M. Graham, and J. Wright, “Predicting drug demand with wikipedia views: Evidence from darknet markets,” in *Proceedings of The Web Conference 2020*, 2020.
- [8] M. Dittus, J. Wright, and M. Graham, “Platform criminalism: The ‘last-mile’ geography of the darknet market supply chain,” in *WWW*, 2018.
- [9] J. Broséus, D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino, and D. Décar-Hétu, “Studying illicit drug trafficking on darknet markets: structure and organisation from a canadian perspective,” *Forensic science international*, vol. 264, pp. 7–14, 2016.
- [10] M. C. Van Hout and T. Bingham, “‘surfing the silk road’: A study of users’ experiences,” *International Journal of Drug Policy*, vol. 24, no. 6, pp. 524–529, 2013.
- [11] J. Aldridge and D. Décar-Hétu, “Hidden wholesale: The drug diffusing capacity of online drug cryptomarkets,” *International Journal of Drug Policy*, vol. 35, pp. 7–15, 2016.
- [12] D. Rhumorbarbe, L. Staehli, J. Broséus, Q. Rossy, and P. Esseiva, “Buying drugs on a darknet market: A better deal? studying the online illicit drug market through the analysis of digital, physical and chemical data,” *Forensic science international*, vol. 267, pp. 173–182, 2016.
- [13] R. Van Wegberg, S. Tajalizadehkhooob, K. Soska, U. Akyazi, C. H. Ganan, B. Klievink, N. Christin, and M. Van Eeten, “Plug and prey? measuring the commoditization of cybercrime via online anonymous markets,” in *USENIX*, 2018.
- [14] R. Kumar, S. Yadav, R. Daniulaityte, F. Lamy, K. Thirunarayanan, U. Lokala, and A. Sheth, “Edarkfind: Unsupervised multi-view learning for sybil account detection,” in *Proceedings of The Web Conference 2020*, 2020, p. 1955–1965.
- [15] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O’Callaghan, “Clustering data streams: Theory and practice,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 15, no. 3, pp. 515–528, 2003.
- [16] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A framework for clustering evolving data streams,” in *VLDB*, 2003.
- [17] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: Ordering points to identify the clustering structure,” in *SIGMOD*, 1999.
- [18] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Sharing clusters among related groups: Hierarchical dirichlet processes,” in *NIPS*, 2005.
- [19] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *JMLR*, vol. 3, pp. 993–1022, 2003.
- [20] D. M. Blei, T. L. Griffiths, and M. I. Jordan, “The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies,” *Journal of the ACM (JACM)*, vol. 57, no. 2, p. 7, 2010.
- [21] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, “Nested hierarchical dirichlet processes,” *IEEE TPAMI*, vol. 37, no. 2, pp. 256–270, 2014.
- [22] C. Wang, D. Blei, and D. Heckerman, “Continuous time dynamic topic models,” in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008, pp. 579–586.
- [23] A. Ahmed, Q. Ho, C. H. Teo, J. Eisenstein, A. Smola, and E. Xing, “Online inference for the infinite topic-cluster model: Storylines from streaming text,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 101–109.
- [24] S. Liang, E. Yilmaz, and E. Kanoulas, “Dynamic clustering of streaming short documents,” in *KDD*, 2016.
- [25] C. Mavroforakis, I. Valera, and M. Gomez-Rodriguez, “Modeling the dynamics of learning activity on the web,” in *WWW*, 2017.
- [26] H. Xu and H. Zha, “A dirichlet mixture model of hawkes processes for event sequence clustering,” in *NIPS*, 2017.
- [27] Y. Seonwoo, A. Oh, and S. Park, “Hierarchical dirichlet gaussian marked hawkes process for narrative reconstruction in continuous time domain,” in *EMNLP*, 2018.
- [28] D. G. Brizan and A. U. Tansel, “A survey of entity resolution and record linkage methodologies,” *Communications of the IIMA*, vol. 6, no. 3, p. 5, 2006.
- [29] L. Getoor and A. Machanavajjhala, “Entity resolution: theory, practice & open challenges,” *Proceedings of the VLDB Endowment*, vol. 5, no. 12, 2012.
- [30] A. G. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.
- [31] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, “Seismic: A self-exciting point process model for predicting tweet popularity,” in *KDD*, 2015.
- [32] A. Reinhart, “A review of self-exciting spatio-temporal point processes and their applications,” *arXiv:1708.02647 [stat]*, 2017.
- [33] M. Farajtabar, “Point process modeling and optimization of social networks,” Ph.D. dissertation, 2018.
- [34] J. S. Liu, “Nonparametric hierarchical bayes via sequential imputations,” *Ann. Statist.*, vol. 24, no. 3, pp. 911–930, 06 1996.
- [35] O. Cappe, S. J. Godsill, and E. Moulines, “An overview of existing methods and recent advances in sequential monte carlo,” *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.
- [36] C. M. Carvalho, M. S. Johannes, H. F. Lopes, and N. G. Polson, “Particle learning and smoothing,” *Statistical Science*, vol. 25, no. 1, pp. 88–106, 2010.
- [37] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
- [38] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *WSDM*, 2015, pp. 399–408.
- [39] Y. Zhang, F. Zhang, P. Yao, and J. Tang, “Name disambiguation in aminer: Clustering, maintenance, and human in the loop,” in *KDD*, 2018.