# Dynamic Theme Tracking in Twitter

Liang Zhao
*Virginia Tech*
*liangz8@vt.edu*

Feng Chen
*University of Albany, SUNY*
*fchen5@albany.edu*

Chang-Tien Lu and Naren Ramakrishnan
*Virginia Tech*
*ctlu@vt.edu, naren@cs.vt.edu*

***Abstract*—Twitter has become a popular social sensor. It is socially significant to surveil the tweet content under crucial themes such as "disease" and "civil unrest". However, this creates two challenges: 1) how to characterize the theme pattern, given Twitter's heterogeneity, dynamics, and unstructured language; and 2) how to model the theme consistently across multiple Twitter functions such as hashtags, replying, and friendships. In this paper, we propose a dynamic query expansion (DQE) model for theme tracking in Twitter. Specifically, DQE characterizes the theme consistency among heterogeneous entities (e.g., terms, tweets, and users) through semantic and social relationships, including co-occurrence, replying, authorship, and friendship. The proposed new optimization algorithm estimates the weight of each relationship by minimizing the Kullback-Leibler divergence. To demonstrate the effectiveness and scalability of DQE, we conducted extensive experiments to track the theme "civil unrest" across 8 Latin American countries.**

***Keywords*-theme tracking; dynamic query expansion; heterogeneous information network.**

## I. Introduction

Twitter is one of the most popular microblogging services and social networks in the world [16]. As of May 2015, there are 646 million users collectively sending over 58 million tweets daily [1]. Compared with traditional media, Twitter has several salient characteristics: 1) Timeliness. Due to their brevity and the widespread use of mobile devices, tweets are commonly posted much faster than traditional media, where hours or even days are spent on compiling, proofreading, typesetting, and publishing. 2) Broad coverage of themes. Tweets cover almost every aspect of our lives, from everyday feelings to breaking news. 3) Diverse channels for information dissemination. Twitter enables "retweeting" for fresh news cascading, "replying" for instant conversations, "hashtag" for theme tagging, and "friendship" for interest sharing. These characteristics make Twitter a highly valuable social sensor for tracking various interesting and crucial themes (e.g., crime and earthquakes), especially when the response times of traditional news outlets are too slow for emergencies and they may be overseen by autocratic governments or threatened by criminal organizations [29]. There is already a considerable body of research on tracking themes in Twitter, and this can be classified into two categories. The first category tracks *general interest themes*, that means it discovers and tracks popular themes among all the themes in Twitter [10], [14], [31]. The second category focuses on tracking only "*targeted themes*" such as earthquakes [29], civil unrest [34], and disease outbreaks [33]. The approaches in this category typically share similar workflow models [19]: given one or more manually selected features (i.e., keywords), a classifier is trained to extract theme-related tweets, whose patterns are then analyzed. Most approaches in this category primarily examine the textual content when tracking a *targeted theme* [19], [29]. However, as a social-psychological behavior, the information diffusion process of theme-related content in Twitter is inevitably influenced by the social relationships (e.g., friendships) [23]. To handle this, a handful of existing works take into account the users' social networks [20], [21], [28].

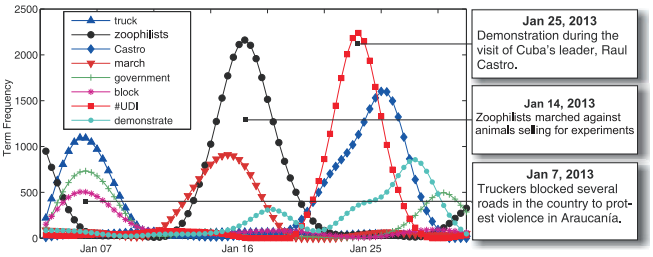


Figure 1: "Civil unrest" related keyterms in Chile on Twitter.

Existing approaches for *targeted theme tracking* suffer from several shortcomings. **First, limited ability to handle theme dynamics.** Existing approaches generally adopt a set of query terms that is manually predefined [20], [21], [28], [29]. However, the effort involved in manually enumerating an exhaustive and unbiased keyword set is usually prohibitive due to its considerable size and the dynamic evolution of trends. Figure 1 illustrates the surveillance of the theme "civil unrest" in Chile over a two week period in January 2013 through Twitter. The top keywords changed continuously on a daily basis, making it difficult to manually determine representative keyterms. **Second, insufficient considerations related to the heterogeneous social relationships.** Existing approaches are generally unable to deal with diverse and dynamic social relationships among tweets and users, such as "friendship" among users, "replying" and "authorship" among tweets. As shown in Figure 2, the likelihood that a user posts a theme-related tweet is typically influenced by conversation contexts, friends' posts, and each individual's preferences. **Third, limited scalability in the**

978-1-4799-9926-2/15/$31.00 ©2015 IEEE

**joint consideration of semantic and social relationships.** The integration of social network and semantic content significantly increases the model complexity, and decreases its scalability. To train the model parameters, multiple variables such as tweets, users, and latent variables are typically coupled in the calculation process, which significantly increases the computation expense.

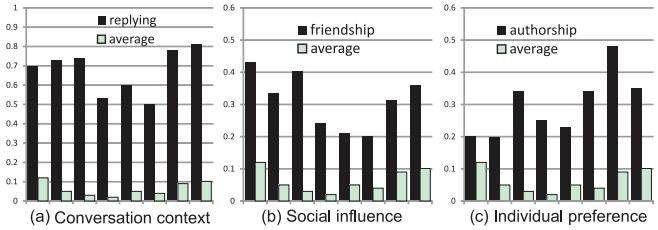


Figure 2: The influence of heterogeneous social relationships on theme patterns. The x-axis represents the datasets for 8 countries[1] on a few randomly selected dates; the y-axis stands for the likelihood that a tweet is related to the theme "civil unrest". "Average" denotes the likelihood that a random tweet is theme-related. "Replying", "friendship", and "authorship" represent the presence of social ties that impact the likelihood that a tweet is theme-related: (a) a tweet is more likely to be theme-related if it replies to a theme-related tweet; (b) a user is more likely to post theme-related tweets if his/her friends posted theme-related tweets; and (c) a tweet is more likely to be theme-related if its user prefers to post theme-related tweets.

To overcome the above challenges, we propose a novel method called dynamic query expansion (DQE) for targeted theme tracking by utilizing the heterogeneous information network in Twitter. Given a tweet stream, our method extracts the top keywords for a targeted theme (e.g., civil unrest). Specifically, for each time interval in tweet streams, we model the Twitter heterogeneous information network by inspecting all the heterogeneous relationships (e.g., co-occurrence, replying, authorship, and friendship) among terms, tweets, and users. Then the theme-related top keywords are extracted as the ones having strong heterogeneous relationships with the entities relevant to the generic concept of the targeted theme. The main contributions are summarized as follows:

- **Proposing a new framework for dynamic theme tracking**: We propose a unified probabilistic framework for theme tracking by jointly considering textual content and heterogeneous social relationships. Neither intensive human labor nor sensitive parameters settings are required.
- **Modeling the Twitter heterogeneous information network**: DQE is applied to track themes by leveraging multiple types of entities (tweets, terms, and users) and their heterogeneous mutual relationships.
- **Designing a scalable optimization method for DQE**: To learn the parameters of DQE, an effective parameter optimization method is designed to minimize theme tracking errors. Linear scalability is achieved by utilizing the conditional independence among entities.
- **Conducting extensive experiments for performance evaluations**: The effectiveness and efficiency of DQE were evaluated on two metrics and compared with existing approaches. Qualitative analysis and case studies further demonstrated its practical usefulness.

The rest of this paper is organized as follows. Section II reviews existing work in this area. Section III presents the problem formulation. Section IV elaborates the mathematical descriptions of the DQE model, and Section V presents the parameter optimization of DQE. In Section VI, extensive experimental results are analyzed. This paper concludes by summarizing the study's important findings in Section VII.

## II. Related Work

There are several threads of related work of this paper.

**Query expansion in microblogs retrieval.** Query expansion is a process that reformulates the seed query in order to improve the coverage and accuracy of information retrieval [6]. To improve the performance of retrieval in Twitter, a new thread of work utilizes query expansion to dynamically expand keywords [6], retrieve tweets [24], and discover events [32]. The expanded keywords are typically extracted by exploring their co-occurrence with the user-specified initial query in textual content, but information diffusion through social network has not been comprehensively explored.

**Event detection in Twitter.** There exists a large amount of work on event detection in Twitter. Event detection methods utilize supervised (e.g., classifier) or unsupervised (e.g., graph clustering) framework to extract tweets subset related to potential events that can be formalized as spatial burstiness [34], [32], temporal burstiness [4], [27], or spatiotemporal burstiness [17], [33]. This thread of work has a different goal from our paper: it detects the emergence instead of the evolution of events, whereas our paper focuses on continuously tracking the evolutionary dynamics of a theme.

**General theme tracking.** A considerable body of work focuses on characterizing the general pattern of Twitter streams. The pattern is typically conceptualized as a mixture of "latent topics". For example, Blei *et al.* aligned the proportion priors and distributions of latent topics over time [10]. Yang *et al.* proposed an efficient Twitter stream summarization approach that can fit in a limited memory [31]. Hong *et al.* analyzed the inter-relationships of multiple social media streams by considering both local topics and shared topics [14]. Mei and Zhai modeled latent topics through a mixture language model, and discovered the transitions among them [25]. However, because "latent topics" are typically extracted purely statistically based on data without human prior knowledge, they do not necessarily have real-

[1]The 8 countries are Argentina, Brazil, Chile, Colombia, Ecuador, El Salvador, Mexico, and Venezuela, shown from left to right.

world meaning. Hence this thread of work is generally not appropriate to track targeted themes.

**Targeted theme tracking.** A thread of work focuses on tracking targeted themes, such as earthquakes. The majority of research adopts classification framework to extract theme-related tweets based on contextual features only [22], [29]. Hence, it is challenging to select an appropriate set of features. Li *et al.* proposed a generic framework for theme-related feature selection, whereas this approach is specially designed for scrawling two specific types of Twitter APIs and is not appropriate for the task of this paper [19]. A handful of methods have been proposed to take into account social relationships. Lin *et al.* implemented a probabilistic mixture model to characterize the temporal textual pattern and diffusion via friendship [20]. Ratkiewicz *et al.* applied a framework specifically designed to track the so-called "political astroturf" based on mentioning networks [28].

## III. Problem Formulation

In this section, we introduce a few key concepts, and then formally define the task of dynamic theme tracking.

Denote $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_T\}$ as a collection of time-ordered Twitter data separated by $T$ time intervals, where $\mathcal{C}_t \in \mathcal{C}$ represents the subcollection of the $t$th time interval. A Twitter subcollection $\mathcal{C}_t$ can be formulated as a Twitter heterogeneous information network:

*Definition 1:* **Twitter Heterogeneous Information Network**: Given a Twitter subcollection, a ***Twitter heterogeneous information network*** is defined as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V} = W \cup D \cup U$, where $W$, $D$ and $U$ denote the node sets of "terms", "tweets" and "users", respectively. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ stands for the undirected edge set. The edge set $\mathcal{E}$ consists of the relationships among the heterogeneous entities, such as "replying" between "tweets", "authorship" between a "user" and a "tweet", and "containment" between a "term" and a "tweet".

The $i$th word is denoted as $W_i$. Similarly, $D_j$ and $U_k$ stand for the $j$th tweet and the $k$th user, respectively.

*Definition 2:* **Theme**: A ***theme*** is a distribution of terms that characterizes a semantically coherent topic or subtopic (e.g., "crime" or "civil unrest"). Mathematically, a ***theme*** is formulated as a unigram language model $\theta$, i.e., a distribution of terms $\{p(q_i|\theta)\}_{i=1}^{|W|}$.

In general, terms with high probabilities are the most important for indicating what the theme is about. For example, when discussing the theme "earthquake", the terms most likely to be mentioned include "shake", "earthquake", and "temblor". Similarly, the most representative terms for the theme "civil unrest" are "protest", "march", and "strike".

*Definition 3:* **Theme Query**: A ***theme query*** is a set of weighted terms that are most representative for the generic concept of a theme. Mathematically, a theme query is a set of tuples $q = \{(q_i, p(q_i|\theta))|q_i \in Q \subset W\}$, where $Q$ is the set of ***theme query terms***, and $q_i$ is a theme query term.

Empirically, $Q$ can be determined by human domain knowledge or by frequency pattern mining. Theme query terms are a popular way of retrieving the theme-related content from Twitter. For example, Lin *et al.* [29] utilized the terms such as "earthquake" to search for tweets talking about the earthquakes. Ratkiewicz *et al.* [28] took this further by compiling a keyword list to retrieve political tweets.

Our main task is to track a theme dynamically, as formalized below.

**Dynamic Theme Tracking**: Given a collection of Twitter data $\mathcal{C}$, ***dynamic theme tracking*** is to continuously calculate the characterizations of a stream of theme-related tweets: $\{\langle\theta, \mathcal{C}_t\rangle\}_{t=1}^{T}$. $\langle\theta, \mathcal{C}_t\rangle$ designates a ***theme snapshot*** of the subcollection $\mathcal{C}_t$ and is mathematically characterized as a unigram language model, i.e., a distribution of terms: $\{p(W_i|\theta, \mathcal{C}_t)\}_{i=1}^{|W|}$.

Dynamic theme tracking is challenging for several reasons. First, the degree of significance of the theme query terms is difficult to quantify in an unbiased fashion even with human domain knowledge. Second, to achieve an effective characterization of theme snapshots, the influence from dynamic heterogenous social relationships needs to be comprehensively modeled. Finally, it is not trivial to achieve linear scalability when the model jointly considers semantic similarity and heterogenous social network structures.

## IV. Dynamic Query Expansion Model for Theme Tracking

This section describes the dynamic query expansion (DQE) model for dynamic theme tracking. In particular, a theme snapshot is characterized by expanding the theme query through the dynamic heterogeneous term dependencies, as described in Section IV-A. Then the calculations of all types of term dependencies are elaborated respectively in Sections IV-B, IV-C, IV-D, and IV-E.

### A. Calculation of Theme Snapshots

Given a theme, it is common to utilize its keywords (i.e., theme query) to retrieve theme-related tweets [20], [21], [29]. Though the theme query represents the generic concept of a theme, the retrieved theme-related content can dynamically evolve when different theme-related occurrences become popular in Twitter. This process is demonstrated in Figure 1, where under the theme "civil unrest", "truck", "government", and "block" trended on Jan 07; "zoophilists", "march", and "demonstrate" were popular on Jan 16; and "#UDI", "Castro", and "demonstrate" became keywords on Jan 25. Theme query terms (e.g., "march") typically appear together with some dynamic keyterms (e.g., "zoophilist") to collectively describe what is happening now, i.e., the theme snapshot. To capture the dynamic relationships between theme query terms and all the other terms, a *translation model* is typically utilized [8]. Given a prior distribution on

the theme query terms, the derivation of the theme snapshot is formulated as follows:

$$p\left(W_j\,|\theta,\mathcal{C}_t\right)=\sum\nolimits_{i=1}^{|q|}p\left(W_j\,|q_i,\mathcal{C}_t\right)K\cdot p(q_i|\theta),\qquad(1)$$

where $p(q_i|\theta)$ is the prior probability of theme query term $q_i$ for the theme $\theta$, and $K$ is the normalization coefficient to ensure the distribution $\{p(q_i|\theta)\}_{i=1}^{|q|}$ is normalized. In the subcollection $\mathcal{C}_t$, $p(W_j|q_i,\mathcal{C}_t)$ is the translation model quantifying the dependency between terms $q_i$ and $W_j$.

A Twitter heterogeneous information network contains multiple edge types that lead to term dependencies. As shown in Figure 2, the theme consistency and semantic proximity of tweets can be strongly reflected by these heterogeneous social relationships. Two terms can have a relationship when they both frequently appear in a tweet, in a conversation context (i.e., replying tweets), in a user's historic posts, or in the collective tweets from a community. This means the calculation of term dependencies should take into account all the term dependencies based on their co-occurrence, replying, authorship, and friendship:

$$p\left(W_j\,|q_i,\mathcal{C}_t\right)=\sum_{k\in\{C,R,A,F\}}\lambda_k\cdot p_k\left(W_j\,|q_i,\mathcal{C}_t\right),\qquad(2)$$

where $\{p_k\left(W_j\,|q_i,\mathcal{C}_t\right)\}_{k\in\{C,R,A,F\}}$ denotes the conditional probability of term $W_j$ given $q_i$, based on different relationships including **C**o-occurrence (C), **R**eplying (R), **A**uthorship (A), and **F**riendship (F). $\lambda_k$ is the weight of the $k$th type of relationship. For simplicity, only the subscript "$t$" of "$\mathcal{C}_t$" will be explicitly shown here. For example, the notation "$p_k\left(W_j\,|q_i,\mathcal{C}_t\right)$" is simplified as follows:

$$p_k^{(t)}(W_j|q_i)=p_k(W_j|q_i,\mathcal{C}_t).$$

### B. Term Dependencies Based on Co-occurrence

Term co-occurrence is generally deemed to be an indicator of semantic proximity and utilized to derive the statistical dependencies between terms [12]. Compared to conventional long documents, the short-length tweet messages cover a narrow and topic-coherent content, which further emphasizes the significance of co-occurrences. The term dependence based on co-occurrence can be formulated as:

$$p_C^{(t)}(W_j|q_i)=\frac{p_C^{(t)}(W_j,q_i)}{p(q_i)}=\sum_k p^{(t)}(W_j,D_k,q_i)/p(q_i),\quad(3)$$

where $p_C^{(t)}(W_j,q_i)$ denotes the probability that the terms $W_j$ and $q_i$ co-occur in the same tweet in time interval $t$, and is calculated by marginalizing tweet variable $D_k$ out of the joint distribution $p(W_j,D_k,q_i)$, which denotes the probability that "$W_j$ and $q_i$ co-occur in the tweet $D_k$".

It can be readily seen that the retrieved tweet $D_k$ is dependent on the theme query term $q_i$, while the term $W_j$ is dependent on $D_k$ that contains it. Hence, "$W_j$" is called "d-separated" [9] by "$D_k$" from "$q_i$", which leads to the conditional independence: $W_j \perp\!\!\!\perp q_i \,|\, D_k$. Equation 3 can then be rewritten as:

$$p_C^{(t)}(W_j|q_i)=\sum\nolimits_{k=1}^{|D|}p^{(t)}(W_j|D_k)p^{(t)}(D_k|q_i),\qquad(4)$$

where $p^{(t)}(D_k|q_i)$ denotes the probability that the tweet $D_k$ is selected among all the tweets retrieved by the theme query term $q_i$ in time interval $t$:

$$p^{(t)}(D_k|q_i)=f^{(t)}(D_k,q_i)/f^{(t)}(q_i),\qquad(5)$$

where $f^{(t)}(q_i)$ is the total frequency of the term $q_i$ while $f^{(t)}(D_k,q_i)$ is the frequency of $q_i$ in tweet $D_k$ in time interval $t$.

Similar to [12], $p^{(t)}(W_j|D_k)$ is defined as the normalized weight of the term $W_j$ in the tweet $D_k$ in time interval $t$:

$$p^{(t)}(W_j|D_k)=s^{(t)}(W_j|D_k)\Big/\sum_{W_l\in D_k}s^{(t)}(W_l\,|D_k\,),\qquad(6)$$

where $s^{(t)}(W_j|D_k)$ denotes the weight of the term $W_j$ in the tweet $D_k$ in time interval $t$, which is defined by Ponte and Croft [26][2].

### C. Term Dependencies Based on Tweet Replying

The "Reply" function is an important and popular feature of Twitter: the percentage of replied and replying tweets is 23% [3]. "Replying" facilitates user conversations on particular themes. Conventionally, a tweet and its replying tweets are causal in context, similar in semantics, and consistent in theme. As shown in Figure 2(a), the tweet messages in the same conversation context generally exhibit theme consistency and semantic similarity, which enables a new channel to derive term dependence:

$$p_R^{(t)}\left(W_j\,|q_i\,\right)=\frac{p_R^{(t)}(W_j,q_i)}{p(q_i)}=\frac{\sum_{l,k}p^{(t)}(W_j,D_k,D_l,q_i)}{p(q_i)},\quad(7)$$

which can be simplified into Equation 8, given $D_k \perp\!\!\!\perp q_i \,|\, D_l$ and $W_j \perp\!\!\!\perp q_i \,|\, D_k$.

$$p_R^{(t)}(W_j\,|q_i\,)=\sum_{k=1}^{|D|}p^{(t)}(W_j|D_k\,)\sum_{l=1}^{|D|}p^{(t)}(D_k|D_l\,)p^{(t)}(D_l|q_i\,).\quad(8)$$

where $p^{(t)}(D_k\,|D_l\,)$ denotes the probability that the tweet $D_k$ is selected from all the tweets having replying relationships with $D_l$, in time interval $t$:

$$p^{(t)}\left(D_k\,|D_l\,\right)=f^{(t)}\left(D_k,D_l\right)\Big/N_{D_l},\qquad(9)$$

where $N_{D_l}$ denotes the number of tweets having replying relationships with the tweet $D_l$. $f^{(t)}(D_k,D_l)$ is a boolean value such that $f^{(t)}(D_k,D_l)=1$ denotes "the tweets $D_k$ and $D_l$ have a replying relationship"; $f^{(t)}(D_k,D_l)=0$, otherwise.

### D. Term Dependencies Based on Authorship

Tweet content is determined by its user's posting behavior, and is generally confined by the user's limited types of personal interests [30]. The importance of "authorship" is well recognized in reflecting the potential similarity in vocabulary, semantics, and theme among all documents from the same author. As shown in Figure 2(c), a tweet is more likely to be theme-related if its user usually posts tweets on this theme and statistically there is likely to be greater theme

[2]The details are elaborated in the supplementary materials at: http://people.cs.vt.edu/liangz8/materials/papers/DTTAddon.pdf

proximity and the semantic similarity among the tweets with the same user. Hence, the term dependence within a single user's posts can be formulated as:

$$p_A^{(t)}(W_j|q_i) = \frac{p_A^{(t)}(W_j,q_i)}{p(q_i)} = \frac{\sum_{k,m,l} p^{(t)}(W_j,D_k,U_m,D_l,q_i)}{p(q_i)},$$

which can be reduced to Equation 10 by considering Equation 8 and $D_k \perp\!\!\!\perp D_l \,|\, U_m$.

$$p^{(t)}(W_j|q_i) = \sum_{k=1}^{|D|} p^{(t)}(W_j|D_k) \sum_{m=1}^{|U|} p^{(t)}(D_k|U_m) \cdot \cdot \sum_{l\neq k}^{|D|} p^{(t)}(U_m|D_l) \cdot p^{(t)}(D_l|q_i), \quad (10)$$

where $p^{(t)}(U_m|D_l)$ denotes the probability that the tweet $D_l$ is posted by $U_m$ in time interval $t$, and $p^{(t)}(D_k|U_m)$ denotes the probability that the tweet $D_k$ is selected among all the posts from user $U_m$ in time interval $t$.

### E. Term Dependencies Based on Friendship

In Twitter, the relationship between two users who follow each other reflects their interactive relationship, and is generally recognized as friendship [13]; a user will befriend others because they are close/similar in interests, belief, geo-location, or social relation [15]. This similarity may result in relevant posting content from friends. More importantly, being friends fosters the dissemination of information, especially on topics of common interest. Figure 2(b) verifies this phenomenon from the statistical point of view, illustrating that a user is more likely to post tweets about a theme if his/her friends also post theme-related tweets. This fact reveals the potential theme proximity and the semantic similarity among the tweets discussed in a friendship community, which supports the utility of a term dependence calculation utilizing friendship. Following the same logic as that described for Equation 10, we get:

$$p_F^{(t)}(W_j|q_i) = \sum_{k,n,m,l} p^{(t)}(W_j,D_k,U_n,U_m,D_l,q_i) \Big/ p(q_i). \quad (11)$$

As with the deduction of the conditional independencies for Equations 3, 7, and 10, we obtain $U_n \perp\!\!\!\perp D_l \,|\, U_m$ and $D_k \perp\!\!\!\perp U_m \,|\, U_n$, allowing Equation 11 to be reformulated as follows:

$$p^{(t)}(W_j|q_i) = \sum_{k=1}^{|D|} p^{(t)}(W_j|D_k) \cdot \sum_{n=1}^{|U|} p^{(t)}(D_k|U_n) \cdot \sum_{m=1}^{|U|} p^{(t)}(U_n|U_m) \cdot \sum_{l=1}^{|D|} p^{(t)}(U_m|D_l) p^{(t)}(D_l|q_i), \quad (12)$$

where $p^{(t)}(U_n|U_m)$ denotes the probability that the user $U_n$ is selected among all the friends of user $U_m$, in time interval $t$:

$$p^{(t)}(U_n|U_m) = f^{(t)}(U_n,U_m) \Big/ \sum_{U_i \in U_m} f^{(t)}(U_i), \quad (13)$$

where the notation "$U_i \in U_m$" signifies that user $U_i$ is a friend of $U_m$, $f^{(t)}(U_n,U_m)$ is the frequency of postings by $U_n$, who is a friend of $U_m$ and $f^{(t)}(U_i)$ denotes the posting frequency of user $U_i$.

## V. Parameter Estimation

This section presents the parameter estimation for the proposed DQE. First, the objective function of the parameter optimization is formulated. Then an effective algorithm is proposed to solve this optimization problem. Finally, the time complexity is analyzed.

### A. Parameter Optimization

In DQE, two sets of parameters need to be estimated. The first set of parameters $\{\lambda_k\}_{k\in\{C,R,A,F\}}$ (see Equation 2) measures the mixture weights of the four heterogeneous relationships, co-occurrence, replying, authorship, and friendship, in the Twitter heterogeneous information network. The second set of parameters is the prior distribution of theme query terms: $\{p(q_i)\}_{i=1}^{|q|}$ in Equation 1. Neither of these two sets of parameters can be directly set manually without bias, but can be estimated by optimizing the model performance. To achieve good performance on dynamic theme tracking, the inferred theme snapshots should be as close as possible to the "gold standard of theme snapshots (GSTS)", which is described in Section VI-A1. Here the GSTS on training set is utilized to estimate the parameters of DQE. Specifically, the parameters are optimized by minimizing the Kullback-Leibler divergence [25] between the inferred theme snapshots and GSTS:

$$\min KL(P|Q) = \min \sum_{t=1}^{T}\sum_{j=1}^{|W|} \log\frac{P^{(t)}(j)}{Q^{(t)}(j)} P^{(t)}(j) = -\sum_{t=1}^{T}\sum_{j=1}^{|W|} \left(P^{(t)}(j)\log p(W_j|\theta,\mathcal{C}_t)\right) + \kappa, \quad (14)$$

where $KL(P|Q)$ denotes the Kullback-Leibler divergence between *GSTS* (denoted as $P$) and *the inferred theme snapshots* (denoted as $Q$). $P^{(t)}(j)$ and $Q^{(t)}(j) = p(W_j|\theta,\mathcal{C}_t)$ denote the probabilities of the $j$th term in time interval $t$. $\kappa = \sum_{t=1}^{T}\sum_{j=1}^{|W|}\left(P^{(t)}(j)\log P^{(t)}(j)\right)$ is a constant value and thus can be discarded in this optimization function.

Considering Equations 1 and 2 and omitting the constant term $\kappa$, Equation 14 is re-arranged as:

$$\min_{p(q_i),\lambda_k} -\sum_{t=1}^{T}\sum_{j=1}^{|W|} P^{(t)}(j)\log(\sum_i p(q_i)\sum_k \lambda_k p_k(W_j|q_i,\mathcal{C}_t)) \quad s.t.\ \textstyle\sum_i p(q_i) = 1,\ p(q_i)\geq 0,\ \sum_k \lambda_k = 1,\ \lambda_k \geq 0 \quad (15)$$

where $p(q_i)$ is the prior probability of the theme query term $q_i$ and $p_k(W_j|q_i,\mathcal{C}_t)$ is the term dependence based on the $k$th type of relationship with the weight $\lambda_k$ in the Twitter heterogeneous information network.

### B. Optimization Problem Solution

There is no closed-form solution to the minimization problem in Equation 15. Instead, this optimization problem can be addressed by iteratively minimizing two convex sub-problems w.r.t. $p(q_i)$ and $\lambda_k$, respectively. Algorithm 1 illustrates the procedures involved in the optimization

**Algorithm 1:** Parameter Estimation for DQE

**Input**: Twitter data collections $\{\mathcal{C}_t\}_{t=1}^{T}$.
**Output**: optimized parameters$\{p(q_i)\}_{i=1}^{|q|}$ and $\{\lambda_k\}_{k\in\{C,R,A,F\}}$.

1 *// calculate term dependencies via heterogeneous relationships*;
2 **for** $t \leftarrow 1$ **to** $T$ **do**
3  **for** $k \in \{C, R, A, F\}$ **do**
4  Calculate term dependence $p_k\,(W_j\,|q_i, \mathcal{C}_t\,)$, $j \in \{1, 2, \cdots, |W|\},\ i \in \{1, 2, \cdots, |q|\}$;
5  **end**
6 **end**
7 *// optimize* $\{p(q_i)\}_{i=1}^{|q|}$ *and* $\{\lambda_k\}_{k\in\{C,R,A,F\}}$. ;
8 **repeat**
9  **for** $i \leftarrow 1$ **to** $|q|$ **do**
10  Calculate $p(q_i)$ based on Equation 16;
11  **end**
12  **for** $k \in \{C, R, A, F\}$ **do**
13  Calculate $\lambda_k$ based on Equation 17;
14  **end**
15 **until** *Convergence*;

problem solution. For each subcollection, the term dependencies based on heterogeneous relationships are calculated (in Steps 2-6). Treating the calculated term dependencies as constants, the objective function in Equation 15 is minimized by iteratively optimizing the prior probability $\{p(q_i)\}_{i=1}^{|q|}$ and mixture weights $\{\lambda_k\}_{k\in\{C,R,A,F\}}$ (in Steps 8-15). The total time complexity is approximately linear to the number of tweets, of which the deduction is in the supplementary material. The optimizations of prior probability and mixture weights are elaborated below.

*1) Estimating prior probability* $\{p(q_i)\}_{i=1}^{|q|}$*:* To Minimize the objective function in Equation 15 w.r.t. $p(q_i)$, Equation 15 is thus re-formulated as the following objective function:

$$\min_{p(q_i)} -\sum_{t=1}^{T}\sum_{j=1}^{|W|} P^{(t)}(j)\log(\sum_i p(q_i)\cdot g_{tij}) \quad s.t.\begin{cases}\sum_i p(q_i)=1\\ p(q_i)\geq 0\end{cases}, \tag{16}$$

where $g_{tij} = \sum_k \lambda_k p_k\,(W_j|q_i, \mathcal{C}_t)$. Equation 16 is equivalent to a *weighted analytic center* problem that is convex [5]. As a commonly utilized solution for this type of problem, Newton's Method can be applied here.

*2) Estimating mixture weights* $\{\lambda_k\}_{k\in\{C,R,A,F\}}$*:* The optimization of mixture weights is formalized as follows:

$$\min_{\lambda_k} -\sum_{t=1}^{T}\sum_{j=1}^{|W|} (P^{(t)}(j)\log(\sum_k \lambda_k\cdot h_{tkj})) \quad s.t.\begin{cases}\sum_k \lambda_k=1\\ \lambda_k\geq 0\end{cases}, \tag{17}$$

where $h_{tkj} = \sum_i p_k\,(W_j|q_i, \mathcal{C}_t)\,p(q_i)$. As with Equation 16, Equation 17 is also solved using Newton's method.

## VI. Experiments

In this section, the empirical performance evaluations of the proposed DQE method are presented. First, the effectiveness of the theme tracking is validated and compared with other methods based on two metrics. Second, the scalability of DQE is evaluated. Case studies and qualitative analyses are provided to demonstrate the practical usefulness of the method proposed here. All the experiments were conducted on a computer with 2.60 GHz Intel i7 CPU and 8.0 GB RAM.

### A. Experimental Setup

*1) Dataset and Gold Standard:* The dataset was constructed by randomly sampling 10% (by volume) of the Twitter data from July 2012 to May 2013 in 8 countries in Latin America, as shown in Table I. The data for the period July 1, 2012 to October 31, 2012 is used as the *training set* for estimating the parameters of our method and the other comparison methods and the data for the second half of the period, from November 1, 2012 to May 31, 2013, is used as *testing set* for the performance evaluation. For each country, the Twitter data collection is partitioned into a sequence of date-interval subcollections. The terms in tweets are stemmed into their roots in corresponding language and stop words are eliminated.

For the purposes of this empirical study, the evaluation of the results obtained is based on the performance for tracking the theme "civil unrest" on a daily basis. We also tested approach on an hourly basis and observed similar patterns. Due to the space limit, only results on a daily basis is reported here. All the methods are validated against a label set known as the "*gold standard of theme snapshots (GSTS)*", which originated from authoritative news outlets. The detailed generation process for GSTS is as follows. First, the most influential international news outlets and the top 3 newspapers in each country are selected based on the rankings given by International Media and Newspapers [2], as listed in Table I. A theme-related report counts if it was published by any of these news sources. For each news report, the representative words were extracted, each of which must simultaneously satisfy the following 3 criteria: 1) must appear over 5 times in the article; 2) must not be a stopword; and 3) must appear in the title. Then, for each date, we retrieved the theme-related tweets – the tweets that contain all the representative words for each of the news reports whose reported event was on that date. For retrieved tweets on each date, the distribution of terms is calculated based on their proportion of frequencies. Consequently, GSTS is designated as the set of these distributions of terms for all the dates.

*2) Metrics:* Two metrics were adopted to evaluate the results of all the methods tested:

- **Cosine Similarity**: As a well-recognized metric for measuring semantic proximity, cosine similarity is used to evaluate the similarity between theme snapshots and GSTS:

$$\cos(\vec{X}_t, \vec{Y}_t) = \vec{X}_t\cdot\vec{Y}_t \Big/ (|\vec{X}_t|\cdot|\vec{Y}_t|),$$

[3]In addition to the top 3 domestic news outlets in each country, the following news outlets were included: The New York Times; The Guardian; The Wall Street Journal; The Washington Post; The International Herald Tribune; The Times of London; Infolatam.

Table I: Dataset and gold standard

| Country | #Tweets (million) | News source[3] | #Events |
|---|---|---|---|
| AR | 52 | Clarín; La Nación; Infobae | 365 |
| BR | 57 | O Globo; O Estado de São Paulo; Jornal do Brasil | 451 |
| CH | 28 | La Tercera; Las Últimas Notícias; El Mercurio | 252 |
| CO | 41 | El Espectador; El Tiempo; El Colombiano | 298 |
| EC | 13 | El Universo; El Comercio; Hoy | 275 |
| EL | 7 | El Diáro de Hoy; La Prensa Gráfica; El Mundo | 180 |
| ME | 51 | La Jornada; Reforma; Milenio | 1217 |
| VE | 45 | El Universal; El Nacional; Ultimas Notícias | 678 |

where $\vec{X}_t$ and $\vec{Y}_t$ are the document vectors formalized for theme snapshots and GSTS in the time interval $t$. Specifically, $\vec{X}_t(i)$ and $\vec{Y}_t(i)$ are assigned the probability values of $i$th term endowed by theme snapshot and GSTS, respectively, in the time interval $t$.

- **The Jaccard Index**: The semantic meaning indicated by a distribution of terms is generally represented by a set of high-probability keyterms [25]. To measure the quality of the keyterm set against GSTS, the Jaccard index, a popular metric for set similarity, is typically utilized [7]:

$$J(A_t(N), B_t(N)) = A_t(N) \cap B_t(N) / A_t(N) \cup B_t(N),$$

where $A_t(N)$ denotes the set of $N$ top-ranked (i.e., highest probability) keyterms based on the theme snapshot in the time interval $t$. $B_t(N)$ denotes the set of $N$ top-ranked keyterms based on GSTS in time interval $t$.

Each metric score is calculated for individual time intervals and then averaged to obtain the overall metric score.

*3) Initial Settings:* The proposed method requires the determination of theme query terms. In particular, theme query terms were designated as the terms that appear (i.e., with non-zero probability) on more than 10 dates in the period of July 1, 2012 to Oct 31, 2012 in GSTS. Consequently, for each country, hundreds of words were extracted as theme query terms.

*4) Comparison Methods:* Our DQE model is compared with 9 methods, including 6 existing methods: *Supervised topic models (STM)* [11], *Query Expansion (QE)* [24], *Dynamic topic models (DTM)* [10], *TEDAS* [18], *Language model-based (LMB) approach* [22], and *Earthquake detection (ED)* [29]. Parameter settings of them are described in supplementary materials. In addition, 3 baselines are compared, including *DQE-C*, *DQE-R*, and *DQE-A*. These are the 3 baselines of the proposed DQE and are the same as DQE except for the calculation of term dependencies such that: 1) "DQE-C" only considers the "co-occurrence" relationship; 2) "DQE-R" considers both the "co-occurrence" and "replying" relationships; and 3) "DQE-A" considers "co-occurrence", "replying", and "authorship". DQE-C, DQE-R, and DQE-A are trained the same way as DQE. The initialization of the theme queries for the baselines is the same as that for DQE, as described in Section VI-A3.

### B. Qualitative Analysis of Effectiveness

As discussed in Section V, the theme query terms' prior probabilities $\{p(q_i)\}_{i=1}^{|q|}$ and the mixture weights $\{\lambda_k\}_{k\in\{C,R,A,F\}}$ must be optimized. The optimization results are shown in Table II and Table III.

The results of the optimized mixture weights shown in Table III indicate that "co-occurrence", which has the highest weights on average, is most significant in characterizing theme snapshots, although "replying" and "authorship" also contribute considerably. It seems large countries such as Brazil and Mexico tend to have relatively high weights for "replying" relationship, while smaller ones such as El Salvador and Ecuador tend to have higher weights for authorship. This phenomenon may originate from the difference in the scales of the social networks in different countries. Finally, although "friendship" generally possesses the least weights, it still exerts an important smoothing effect on the distribution of terms.

Table II lists the top 10 "civil unrest" theme query terms optimized for each country. The terms are in Spanish, Portuguese or English. Amongst all the countries, the terms semantically related to "civil unrest" are typically high-ranked, e.g., "protest" and "movement". Also included are the names of well-known protest organizations, e.g., "MST"[4] and "@epn"[5].

### C. Quantitative Analysis of Effectiveness

To evaluate the effectiveness of our DQE, the performance of dynamic theme tracking is evaluated against GSTS via two metrics: cosine similarity and the Jaccard index.

*1) Evaluation based on cosine similarity:* The theme tracking results for all the methods are validated against GSTS based on cosine similarity score, as shown in Table IV. The proposed DQE performs best, obtaining a score of 0.48 on average. It outperforms the best existing methods ED and DTM by about 40%, and also outperforms the 3 baselines methods: DQE-A by about 10%, DQE-R by about 20%, and DQE-C by about 40%. This is because DQE considers all three of the social relationships: replying, authorship, and friendship, which can provide good heuristics for dynamic keyterms and effective smoothing on theme snapshots. The methods LMB and TEDAS have the worst performance, resulting in scores lower than 0.2 for several countries. The reason for this is that LMB only considers unigrams for classification, but omits the relationships among them, and the query expansion in TEDAS is only able to retrieve a portion of all the theme-related tweets. STM also exhibits a poor performance, with a score of 0.25 on average. Compared with STM, DTM performs much better, achieving

[4]MST: Brazil's Landless Rural Workers' Movement

[5]epn: The Twitter account of the president of Mexico

Table II: "Civil unrest" top-10 theme query terms (stemmed) optimized by DQE.

| Brazil (BR) | Colombia (CO) | Mexico (ME) | El Salvador (EL) | Chile (CH) | Argentina (AR) | Venezuela (VE) | Ecuador (EC) |
|---|---|---|---|---|---|---|---|
| manifestaca | protest | protest | enfrent | protest | protest | protest | enfrent |
| protest | enfrent | manifest | protest | carabin | par | manifest | protest |
| tom | manifiest | part | univers | enfrent | gobiern | president | movimient |
| congress | march | PRI | movimient | manifest | manifestacion | chavez | march |
| massacr | movimi | reform | Salvador | movimi | manifiest | apoy | educacion |
| march | educacion | @epn | trabaj | manifestacion | call | call | libert |
| grev | colombi | maestr | ano | Chile | pais | huelg | grup |
| MST | patriot | mexic | vinotint | alamed | party | trabaj | president |
| moviment | president | educacion | comalap | tom | tom | educacion | grevist |
| polic | manifestacion | enfrent | escol | estudiantil | polic | CNE | demostracion |

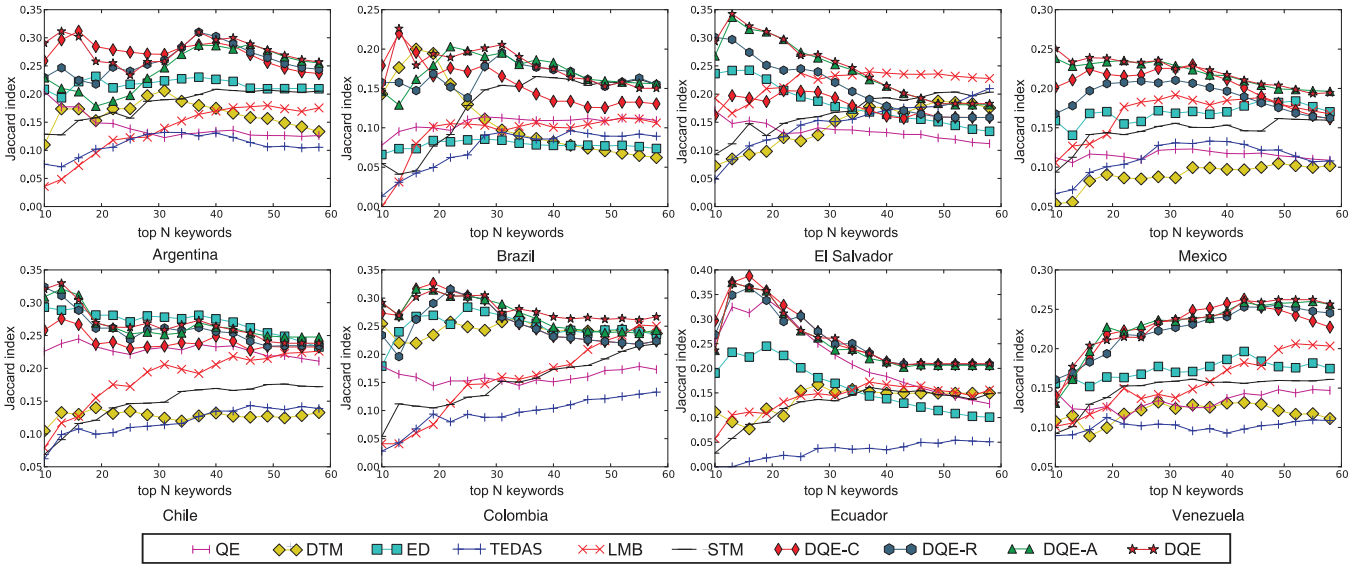


Figure 3: The Jaccard index of theme tracking results against the gold standard when the number of top keywords varies. DQE achieves highest overall scores in most countries, and is more effective in predicting the most significant keywords (e.g., top 20 keywords).

Table III: Mixture weights of heterogeneous relationships optimized by DQE. The trade-offs reflect the relative importance of distinct relationships in calculating term dependencies for dynamic theme tracking.

| Country | Co-occur | Replying | Authorship | Friendship |
|---|---|---|---|---|
| BR | 0.81 | 0.12 | 0.07 | 0.00 |
| CO | 0.85 | 0.05 | 0.08 | 0.02 |
| ME | 0.72 | 0.12 | 0.13 | 0.03 |
| EL | 0.47 | 0.00 | 0.49 | 0.04 |
| CH | 0.85 | 0.00 | 0.11 | 0.02 |
| AR | 0.52 | 0.33 | 0.10 | 0.05 |
| VE | 0.67 | 0.25 | 0.06 | 0.02 |
| EC | 0.65 | 0.03 | 0.32 | 0.00 |

scores of around 0.3-0.5 in most countries. This is because DTM models the theme's temporal evolution, leading to more effective smoothing on the theme snapshots over time. ED and QE also achieve good performance, although still not as good as DQE, because they omits the social ties that impact the theme patterns.

*2) Evaluation based on the Jaccard index:* Figure 3 shows the dynamic theme tracking evaluation results based on the Jaccard index by varying $N$, the number of top-ranked keyterms, from 10 to 60. Performing consistently the best, DQE does especially well in predicting the top 10-20 keyterms, which demonstrates its important capacity to predict the keyterms that matter most. The performances of DQE, DQE-A, DQE-R, and DQE-C are close and outperform all the existing methods. This is because the theme query terms have been well estimated, which enables the dynamic keyterms to be retrieved very effectively via the heterogenous relationships in Twitter. QE is also good at predicting the top 20 keyterms, achieving a 0.20 Jaccard index on average. ED performs reasonably well, achieving a score of 0.20 on average in the top 20 keyterms. DTM, TEDAS, and STM perform the worst, especially when $N$ is small. This is because they are incapable of extracting most theme-related content. LMB achieves a higher score when $N$ is larger, with a score of 0.21 in predicting the top 60 keyterms, but is still lower than DQE because it is smoothed by a background model that does not take into account more accurate heuristics like social relationships.

DQE achieves the best overall performance on both metrics. The effectiveness of utilizing the heterogeneous social relationships is also clearly demonstrated by these results.

### D. Scalability Study

To examine the scalability of DQE, Figure 4 plots the running times of all the methods when the sizes of their input data are varied. As can be seen from Figure 4(a), the running times of DQE, DQE-A, DQE-R, and DQE-C all

Table IV: The cosine similarity scores of theme tracking

| | BR | CO | ME | EL | CH | AR | VE | EC |
|---|---|---|---|---|---|---|---|---|
| QE | 0.23 | 0.3 | 0.22 | 0.36 | 0.46 | 0.26 | 0.31 | 0.27 |
| TEDAS | 0.14 | 0.18 | 0.25 | 0.26 | 0.23 | 0.23 | 0.07 | 0.07 |
| STM | 0.15 | 0.2 | 0.2 | 0.32 | 0.21 | 0.22 | 0.19 | 0.26 |
| DTM | 0.41 | 0.53 | 0.27 | 0.11 | 0.35 | 0.33 | 0.26 | 0.35 |
| LMB | 0.16 | 0.1 | 0.24 | 0.35 | 0.11 | 0.16 | 0.27 | 0.19 |
| ED | 0.23 | 0.47 | 0.27 | 0.44 | 0.51 | 0.26 | 0.38 | 0.36 |
| DQE-C | 0.43 | 0.46 | 0.4 | 0.3 | 0.43 | 0.36 | 0.34 | 0.31 |
| DQE-R | 0.36 | 0.47 | 0.34 | 0.61 | 0.57 | 0.38 | 0.37 | 0.31 |
| DQE-A | 0.3 | 0.6 | 0.42 | 0.63 | 0.59 | 0.45 | 0.31 | 0.39 |
| DQE | 0.43 | 0.61 | 0.43 | 0.63 | 0.59 | 0.46 | 0.32 | 0.39 |

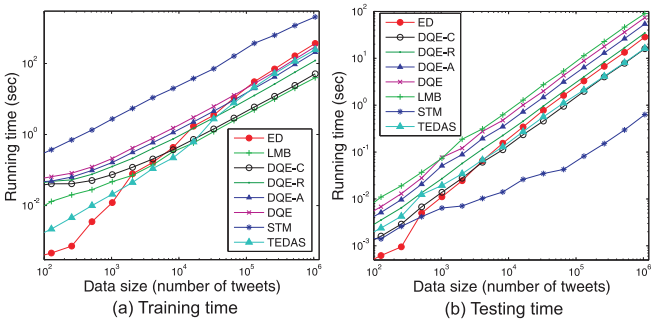


Figure 4: Running times of all the supervised methods. Running time for DQE increases linearly and is among the smallest.

increase linearly with the size of the training data. This result verifies our theoretical analysis of the time complexity in Section V-B. In addition, LMB and STM also achieves linear scalability. LMB consumes least training time, while STM's time cost is the largest at about 50 times larger than any of the other methods. Being implemented based on linear SVM, both TEDAS and ED have super-linear scalability. For TEDAS, this is because its number of "social network" features such as hashtags and mentioning symbols (e.g., @nfl) increases as the data size becomes larger. For ED, its features encompass all the words in the corpus, which means its number of features grows along with the input data size. As shown in Figure 4(b), for the test phase, all the methods have linear scalability except STM. LMB has the largest time consumption, while STM consumes least.

### *E. Case Studies*

During the experiment, we observed numerous interesting cases of theme tracking by DQE. Looking at the example of January 2013 in Chile, the theme tracking results shown in the first column of Figure 5 are compared to the "civil unrest" related news reports in the second column. All the terms have been translated into English. The graph on the left hand side of the figure plots the dynamics of keyterms that are among the top-10 for more than 5 days. The curves are smoothed by spline functions [9]. Also shown are the 5 top-ranked keyterms for each date and the second column lists the news' headlines/bylines of all the events. Two interesting facts can be observed:

1) *Theme-related keyterms on different dates exhibit the dynamic nature of theme snapshots.* As shown in Figure 5, the top-ranked keyterms in theme snapshots change dramatically from day to day. Various words, such as a person's name (e.g., Castro), an object (e.g., truck), and hashtags (e.g., #UDI) can become keyterms temporarily due to the trending events under this theme.

2) *The theme-related keyterms discovered by DQE are a good match for those appearing in news reports.* In Figure 5, matching terms in the first and second columns are shown in the same colors. For example, the top ranked keyterms discovered by DQE were "truck, "movilizacion", and "road" on January 7 when the event "truckers' strikes on several roads" occurred, while around Jan 27, the top keyterms discovered by DQE included "Castro", "embassy", and "UDI", which corresponds to reports from news outlets about the demonstrations against Raul Castro[6] outside the Embassy of Cuba in Santiago, Chile.

## VII. Conclusion

This paper presents a dynamic query expansion (DQE) model for dynamic theme tracking. Specifically, DQE characterizes the theme snapshots on time-ordered subcollections by utilizing the heterogeneous social relationships in Twitter. The proposed optimization algorithm effectively estimates two sets of model parameters by minimizing the Kullback-Leibler divergence. Though modeling complex semantic and social relationships among heterogeneous entities, DQE achieves linear scalability due to the effective utilization of conditional independencies. DQE's effectiveness for theme tracking is demonstrated by its ability to outperform existing methods on two different metrics. Additionally, the time consumption of DQE is empirically validated to be linear in the data size. Finally, real-world case studies on tracking "civil unrest" in Chile demonstrate the practical usefulness of the proposed approach.

## Acknowledgment

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

## References

[1] 2015 twitter statistics. http://www.statisticbrain.com/twitter-statistics/. Accessed 2015 Jun 30.
[2] International Media and Newspapers. http://www.4imn.com/.
[3] Sysomos Inc. https://www.sysomos.com/insidetwitter/engagement/. Accessed sep 22, 2015.
[4] C. C. Aggarwal and K. Subbian. Event detection in social streams. In *SDM*, volume 12, pages 624–635. SIAM, 2012.

[6]Raul Castro: The leader of Cuba.

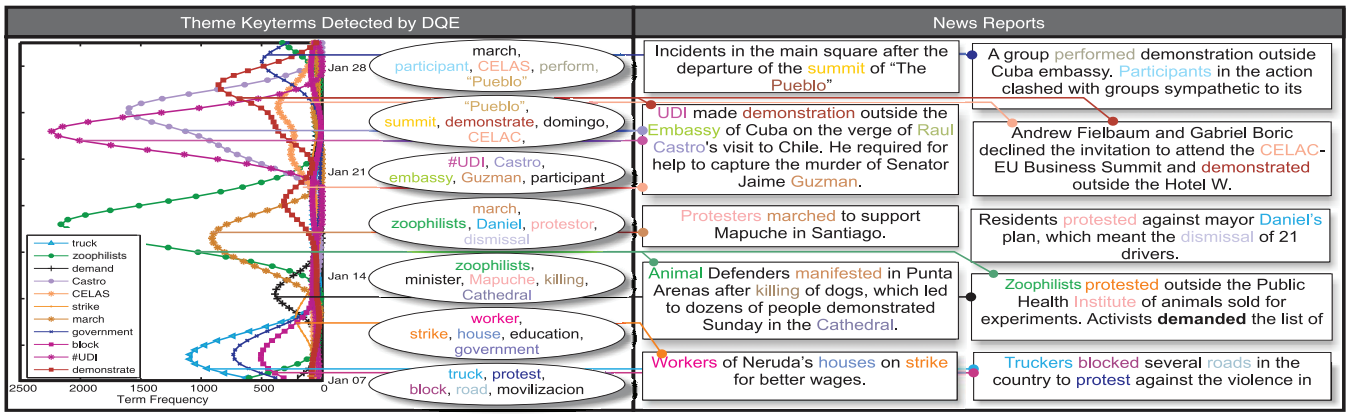


Figure 5: "Civil unrest" theme tracking in Jan 2013, Chile by DQE. The detected keyterms are compared with news headlines/bylines.


[5] D. S. Atkinson and P. M. Vaidya. A scaling technique for finding the weighted analytic center of a polytope. *Mathematical Programming*, 57(1-3):163–192, 1992.

[6] A. Bandyopadhyay, K. Ghosh, P. Majumder, and M. Mitra. Query expansion for microblog retrieval. *International Journal of Web Science*, 1(4):368–380, 2012.

[7] R. Bekkerman, M. Scholz, and K. Viswanathan. Improving clustering stability with combinatorial MRFs. In *KDD*, pages 99–108. ACM, 2009.

[8] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR*, pages 222–229. ACM, 1999.

[9] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[10] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120. ACM, 2006.

[11] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, volume 7, pages 121–128, 2007.

[12] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *WWW*, pages 325–332. ACM, 2002.

[13] C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de L Arcanjo. Inferring the location of Twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[14] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *KDD*, pages 832–840. ACM, 2011.

[15] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *WSDM*, pages 537–546. ACM, 2013.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW*, pages 591–600. ACM, 2010.

[17] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras. On the spatiotemporal burstiness of terms. *VLDB*, 5(9):836–847, 2012.

[18] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang. TEDAS: a Twitter-based event detection and analysis system. In *ICDE*, pages 1273–1276. IEEE, 2012.

[19] R. Li, S. Wang, and K. C.-C. Chang. Towards social data platform: automatic topic-focused monitor for Twitter stream. *VLDB*, 6(14):1966–1977, 2013.

[20] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky. The joint inference of topic diffusion and evolution in social communities. In *ICDM*, pages 378–387. IEEE, 2011.

[21] C. X. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *KDD*, pages 929–938. ACM, 2010.

[22] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *KDD*, pages 422–429. ACM, 2011.

[23] M. H. MacRoberts and B. R. MacRoberts. Problems of citation analysis. *Scientometrics*, 36(3):435–444, 1996.

[24] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in Information Retrieval*, pages 362–367. Springer, 2011.

[25] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD*, pages 198–207. ACM, 2005.

[26] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281. ACM, 1998.

[27] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, et al. 'beating the news' with embers: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1799–1808. ACM, 2014.

[28] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, pages 297–304, 2011.

[29] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860. ACM, 2010.

[30] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *KDD*, pages 1049–1058. ACM, 2010.

[31] X. Yang, A. Ghoting, Y. Ruan, and S. Parthasarathy. A framework for summarizing and analyzing Twitter feeds. In *KDD*, pages 370–378. ACM, 2012.

[32] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS one*, 9(10):e110206, 2014.

[33] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Spatiotemporal event forecasting in social media. In *SDM*, pages 963–971. SIAM, 2015.

[34] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *KDD*, pages 1503–1512. ACM, 2015.