Ups and Downs in Buzzes: Life Cycle Modeling for Temporal Pattern Discovery

Yi Chang[‡], Makoto Yamada[†], Antonio Ortega[‡], Yan Liu[‡]
[†]Yahoo Labs, Sunnyvale, CA 94089
[‡]University of Southern California, Los Angeles, CA 90089
{yichang, makotoy}@yahoo-inc.com, ortega@sipi.usc.edu, yanliu.cs@usc.edu

***Abstract*—In social media analysis, one critical task is detecting burst of topics or *buzz*, which is reflected by extremely frequent mentions of certain key words in a short time interval. Detecting buzz not only provides useful insights into the information propagation mechanism, but also plays an essential role in preventing malicious rumors. However, buzz modeling is a challenging task because a buzz time-series usually exhibits sudden spikes and heavy tails, which fails most existing time-series models. To deal with buzz time-series sequences, we propose a novel time-series modeling approach which captures the rise and fade temporal patterns via *Product Life Cycle* (PLC) models, a classical concept in economics. More specifically, we propose a mixture of PLC models to capture the multiple peaks in buzz time-series and furthermore develop a probabilistic graphical model (*K-MPLC*) to automatically discover inherent life cycle patterns within a collection of buzzes. Our experiment results show that our proposed method significantly outperforms existing state-of-the-art approaches on buzzes clustering.**

***keywords:* Time-Series Modeling, Time-Series Clustering**

I. Introduction

Social media is growing at an explosive rate, and hundreds of millions of users generate vast amounts of contents on various social media web sites, such as Twitter and Tumblr. One critical problem in social media analysis is to effectively model buzz events, which are reflected by frequent mentions of certain key words in a short time interval, such as *new iPhone* or *Hurricane Sandy*. Since the buzz events are closely related to human activities, effectively modeling buzzes could help us track mass attention, obtain public opinions, or forecast users' reactions to a particular event in the near future.

Buzz modeling is an extremely challenging problem since many buzz events are unforeseeable (e.g., earthquakes or hurricanes), while others may involve human interventions. As a result, a buzz time-series, which consists of the number of mentions for each time unit, usually exhibits sudden spikes and heavy tails. Thus, it is very hard to represent a buzz time-series using traditional time-series models [3]. Existing buzz modeling methods use simple time-series models (e.g., univariate models), and explore additional feature cues (such as sentiment, spreading measures and controversialist), to predict whether a certain topic can become viral [7], [15], [5]. These methods achieve success to a certain extent, but few of them explicitly capture the inherent temporal patterns of buzzes. Effective buzz models, which not only can capture those sudden spikes and heavy tails patterns but also can differentiate true spikes from jagged noise in a time-series sequence, are in great need.

One important challenge in pursuit of effective buzz models is how to find distinct temporal patterns within a set of buzz time-series. For example, as seen in Figure 1, *Euro 2012* attracted more and more frequent mentions when it came close to its final match day, a sudden increase of frequency happened when the final match started, and the large volume was maintained for a couple of hours during the final match, after which the mention volume dropped dramatically in just a few hours on social media. Therefore capturing the temporal patterns and modeling the inherent structures of these temporal patterns can provide us useful insights into information propagation[22] and buzz prediction in social media.

In this paper, we propose to model rise-and-fade temporal shapes using *product life cycle* (PLC) models [16]. PLC was originally introduced by economists to model the life span of a product: introduction to the market (initial sales), growth (sudden increase of sales), equilibrium (maturity phase defined by approximately constant sales) and decline (when the sales decrease dramatically). In Figure 1, we can observe 4 different stages, which are segmented with dotted lines. In general, for different types of buzz events, the growth and equilibrium stages might be relatively short, while the decline stage varies. Moreover, a buzz sequence often has multiple peaks, and the number of peaks is unknown in advance. To handle a time-series sequence with multiple peaks, we first propose to model each buzz with a mixture of PLC models, which is parameterized by weights, shapes, decay times, and locations of each individual PLC model. Then, we propose an efficient $L1$-regularized approach to achieve a sparse solution for the PLC mixture models. Second, to discover the underlying patterns in a collection of buzz time-series, we develop a probabilistic graphical model, K-Mixture of Product Life Cycle (*K-MPLC*), to automatically group buzz time-series based on the PLC mixture parameters.

Through experiments, we compare the proposed approach with state-of-the-art time-series modeling and clustering methods including SpikeM [19], K-spectral centroid algorithm (K-SC) [25], and the spectral clustering with the dynamic time warping (DTW+Spectral) [26]. As compared to K-SC, DTW+Spectral, and SpikeM, the proposed approach can systematically capture the sharp and heavy-tailed peaks via the decay parameter. The experimental results show that K-MPLC

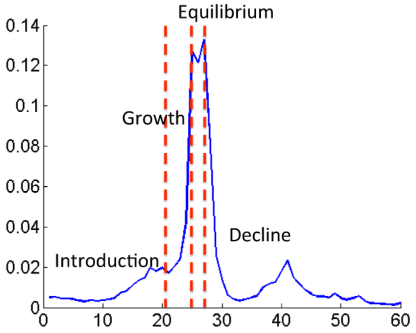


Fig. 1. Buzz time-series of *Euro 2012*.

compares favorably with K-SC and DTW+Spectral for buzzes clustering.

Contributions: The major contributions of this paper include: (i) leveraging the product life cycle concept to model the temporal patterns of buzz events; (ii) a flexible mixture of PLC model to systematically model the possible multiple peaks in a buzz time-series sequence and achieve sparse solutions via the Lasso based approach; (iii) a probabilistic graphical model to discover the inherent clusters within a collection of buzzes.

II. Related Work

Time-series modeling: To model the temporal patterns of online content, Matsubara et al. [19] propose an algorithm to model the rise and fall patterns of influence propagation. Hong et al. [8] attempt to model the time decay of topics on Twitter with Gamma function. Furthermore, the temporal information of online forums [6], blogspace [13], and online groups [12] has been explored and mined under different scenarios, and spatiotemporal patterns of subtopics have been investigated by [20]. However, most of the existing methods can only deal with single spike signal. Thus, existing time-series modeling methods are not suited for buzz events with multiple spikes.

Time-series clustering: Time-series clustering has been an active research topic in recent years, and various time-series clustering algorithms are proposed such as batch mode clustering [21], [17], incremental clustering [18], and anytime clustering [26]. One of the important component of time-series clustering is to choose a distance metric, and Dynamic Time Warping (DTW) and Complexity-Invariant Distance (CID) are the most widely used metrics [26], [1]. Clustering algorithms have been applied to various types of time-series data, including snippets within long time-series [21], different time-series streams [17], and multiple time-series matrix [10]. However, for extremely noisy buzz time-series sequences, DTW scores tend to be inaccurate, and thus clustering performance can be degraded.

The most similar work to ours is that of Yang and Leskovec [25], in which the K-spectral centroid (K-SC) algorithm is proposed to group similar time-series via aligning peaks, shifting over time-axis, and scaling over frequency-axis. The K-SC algorithm can cluster time-series that can be aligned by shifting and scaling operations, but fails for sharp or heavy-tailed temporal sequences. Moreover, if two time-series share similar peak shapes but differ in the intervals between the peaks, K-SC cannot identify these two time-series as the same cluster through simple shifting or scaling.

III. Modeling Buzzes with mixtures of Product Life Cycle Models

A. Product Life Cycle Models

The concept of *Product Life Cycle* (PLC) was originally proposed in economics in the 1960s [16]. The classical view of PLC assumes 4 phases to cover the life span of a product: introduction, growth, equilibrium and decline. Introduction stage refers to low growth rate of sales as the product is newly launched in the market; growth stage implies that the public gains awareness of the product and consumers come to understand its benefits and accept it, so that a company can expect a period of rapid sales growth; equilibrium stage corresponds to the product reaching maturity, so that the sales growth slows and sales volume eventually peaks and stabilizes; decline stage indicates that the product enters into decline, as sales and profits start to fall because the market has become saturated, the product has become obsolete, or customer tastes have changed. Many different versions of PLC models are introduced in [3].

To model the long decay trend of the decline stage of a product, Isaic-Maniu and Voda [9] propose to use the Gamma distribution as a PLC model. The probability density function of Gamma distribution is:

$$f(t;\alpha,\beta) \quad = \frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}, \tag{1}$$

where $t \geq 0$, $\alpha > 0$ is the shape parameter, $\beta > 0$ is the rate parameter, and the Gamma function $\Gamma(\alpha)$ is defined as:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} \mathrm{d}t.$$

As compared to other distribution functions, such as the Gaussian distribution, the Gamma distribution can model a sudden rise and long decay trend in the decline stage of a PLC. Note that our method is not limited to the Gamma distribution, and it can be generalized to other base PLC models as well, such as the the Alpha distribution [9].

B. Mixtures of PLC Model

Given a buzz topic, we count its mentions on a given social media channel during a pre-determined time interval (e.g., an hour), and generate a time-series sequence of this topic over a number of intervals during an observation window. Since a buzz sequence may consist of several obvious peaks, it could be modeled with multiple PLC models. As the number of peaks is not known in advance, we model a buzz sequence with a mixture of PLCs as:

$$f(t;\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu}) = \sum_{\ell=1}^{L} w_\ell f(t;\alpha_\ell,\beta_\ell,\mu_\ell), \tag{2}$$

$$f(t;\alpha,\beta,\mu) = \begin{cases} Z^{-1}(t-\mu)^{\alpha-1}e^{-\beta(t-\mu)} & (t \geq \mu) \\ 0 & (\text{Others}) \end{cases},$$

where L is the number of PLC models, $\boldsymbol{w} = [w_1, \ldots, w_L]^\top$ denotes the weight vector, $^\top$ denotes the matrix transpose, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]^\top$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_L]^\top$ are the vectors of Gamma distribution parameters, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_L]^\top$ refers to locations of PLC models, and Z is the normalization factor.

C. Mixtures of PLC Model Estimation

Let us denote buzz time-series as $\boldsymbol{y} = [y_1, \ldots, y_T]^\top$, where T refers to the length of time-series. Note, in this paper, we assume time-series are normalized (i.e., $\sum_{t=1}^{T} y_t = 1$), so that it can be modeled by probability density functions. Then, the optimization problem of fitting the buzz time-series using a mixture of PLC models can be formulated as:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\mu}} \quad & \sum_{t=1}^{T} (y_t - f(t; \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}))^2 \\ \text{s.t.} \quad & \alpha_\ell > 0, \beta_\ell > 0, w_\ell > 0, \mu_\ell > 0, \ell = 1, 2, \ldots, L. \end{aligned}$$

We can solve this optimization problem by using gradient descent. However, since this optimization problem is *non-convex*, it tends to lead to poor local optimum solutions.

To deal with the *non-convexity* problem, we first relax the problem to a *convex* optimization problem and then propose a *Lasso* based approach [23], making it possible to obtain a global solution. The idea of the Lasso based approach is to select L PLC models from a large number of PLC candidates. More specifically, we first define $TM_\alpha M_\beta$ basis functions, where M_α and M_β are the number of candidate values for α and β, so that each basis function corresponds to a PLC model with fixed α and β parameters placed at μ. Note that, since a PLC model can be located at any of T positions, the total number of basis functions is $TM_\alpha M_\beta$.

Then, we solve the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{v}} \quad & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{v}\|^2 + \lambda \|\boldsymbol{v}\|_1 \\ \text{s.t.} \quad & v_\ell \geq 0, \ell = 1, 2, \ldots, TM_\alpha M_\beta, \end{aligned} \tag{3}$$

where $\boldsymbol{v} \in \mathbb{R}^{TM_\alpha M_\beta}$ is the model parameter and $\boldsymbol{K} \in \mathbb{R}^{T \times TM_\alpha M_\beta}$ is the pre-computed Gamma distribution functions:

$$K_{t,\mu \times i \times j} = \begin{cases} Z_{ij}^{-1}(t-\mu)^{\widetilde{\alpha}_i - 1} e^{-\widetilde{\beta}_j (t-\mu)} & (t \geq \mu) \\ 0 & (\text{Others}) \end{cases} .$$

In the above formula, $Z_{ij} = \max(\{(t-1)^{\widetilde{\alpha}_i - 1} e^{-\widetilde{\beta}_j (t-1)}\}_{t=1}^{T})$ is the normalization factor, $\widetilde{\alpha}_i$, $i = 1, \ldots, M_\alpha$ and $\widetilde{\beta}_j$, $j = 1, \ldots, M_\beta$ are pre-defined Gamma parameters, μ is the location index of a PLC model ($t - \mu$ indicates the shift of a PLC along x-axis), and $\|\boldsymbol{v}\|_1$ is the $L1$ regularizer to avoid overfitting. Since the optimization problem of Eq.(3) is *convex* with respect to $\boldsymbol{v}$, Eq.(3) can be solved by using a state-of-the-art Lasso optimization solver. In this paper, we employ the *dual augmented Lagrangian* (DAL) based approach [24]. Since we use the $L1$ regularizer, we can select a small number of non-overlapped basis functions.

After the Lasso fitting, we first obtain M PLC models ($M > L$) with using a small regularization parameter, and then select L PLC models from them. After fixing individual PLC parameters with Lasso, we solve the following quadratic programming (QP) problem to obtain the corresponding weight $\boldsymbol{w}$ for each PLC model:

$$\min_{\boldsymbol{w}} \quad \|\boldsymbol{y} - \widetilde{\boldsymbol{K}}\boldsymbol{w}\|^2, \quad \text{s.t.} \quad w_\ell \geq 0, \sum_{i=1}^{L} w_i = 1,$$

where $\widetilde{\boldsymbol{K}} \in \mathbb{R}^{T \times L}$ is the selected L basis functions from $\boldsymbol{K}$. We call this entire Lasso based fitting framework as *PLC Lasso*.

IV. Clustering Buzzes using K-MPLC

In this section, we first propose three types of features for characterizing buzz time-series. Then, in order to discover underlying similar patterns in a set of buzz time-series, we propose a novel probabilistic graphical model, K-Mixture of Product Life Cycle (K-MPLC), to cluster N time-series based on their PLC parameters.

A. Feature Extraction

We obtain estimated model parameters $\boldsymbol{w}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\mu}$ for each buzz time-series sequence using the PLC Lasso. However, some of those parameters may not be useful for buzzes clustering. For example, if buzz time-series are not aligned according to their peak positions, location of each PLC model varies, and it may not be possible to obtain meaningful clusters using absolute PLC locations. Thus, in this paper, we propose three types of effective and robust features for characterizing buzz time-series sequences.

Weight parameters: To capture the shape information of a time-series sequence, we use the normalized weight parameters obtained by PLC Lasso fitting $\widehat{\boldsymbol{w}}$. The number of weight parameters relies on the number of PLC models.

Inverse β of the maximum PLC: To discriminate sharp-tailed and heavy-tailed buzz sequences, we propose the following feature:

$$\tau = \beta_{i_{\max}}^{-1},$$

where $i_{\max}$ is the index of the largest PLC with respect to weight $\boldsymbol{w}$.

Standard deviation of PLC locations: To discriminate buzz time-series with single peak from multiple peaks, we propose the following feature:

$$\sigma = \sqrt{\frac{1}{L} \sum_{\ell=1}^{L} \widehat{w}_\ell (\mu_\ell - \mu')^2},$$

where $\mu_1, \ldots, \mu_L$ are locations of PLCs and $\mu' = \sum_{\ell=1}^{L} \widehat{w}_\ell \mu_\ell$ is their weighted mean.

B. K-MPLC

To discover the underlying similar patterns of buzz time-series, we propose a probabilistic graphical model, K-Mixture of Product Life Cycle (K-MPLC), to cluster N time-series into K groups. Throughout this paper, we assume K is known.

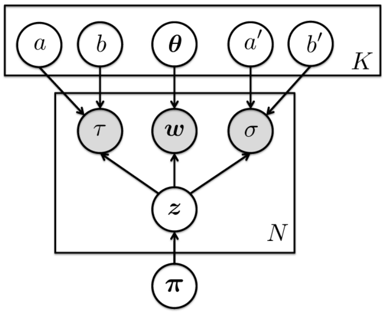


Fig. 2. Graphical Model of K-MPLC.

Suppose that we are given N buzz time-series and their corresponding parameters $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N] \in \mathbb{R}^{L\times N}$, $\boldsymbol{\tau} \in \mathbb{R}^N$, and $\boldsymbol{\sigma} \in \mathbb{R}^N$. Given distinct characteristics over different parameter vectors, we use a Dirichlet distribution to model the weight vector $\boldsymbol{w}$ as its sum needs to be 1. Since τ and σ take positive values, we use the Gamma distribution to model them separately. The graphical model we propose is shown in Figure 2.

Specifically, the probability for each instance is:

$$p(\boldsymbol{w},\tau,\sigma|\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}')=\sum_{k=1}^{K}\pi_k p(\boldsymbol{w}|\boldsymbol{\theta}_k)p(\tau|a_k,b_k)p(\sigma|a'_k,b'_k).$$

Here, $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^\top$ are the mixture weights, and

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{\theta}_k) &= C(\boldsymbol{\theta}_k)\prod_{i=1}^{L} w_i^{\theta_{ki}-1}, \\ p(\tau|a_k,b_k) &= \frac{b_k^{a_k}}{\Gamma(a_k)}\tau^{a_k-1}e^{-b_k\tau}, \\ p(\sigma|a'_k,b'_k) &= \frac{{b'_k}^{a'_k}}{\Gamma(a'_k)}\sigma^{a'_k-1}e^{-b'_k\sigma} \end{aligned}$$

are Dirichlet and Gamma distributions, $C(\boldsymbol{\theta}) = \frac{\Gamma(\sum_{i=1}^{L}\theta_i)}{\Gamma(\theta_1)\Gamma(\theta_2)\ldots\Gamma(\theta_L)}$, and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\mathrm{d}t$.

In this paper, we use maximum likelihood estimation to estimate the model parameters. Specifically, the optimization problem for the proposed model can be given as:

$$\begin{aligned} \max_{\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}'} \quad & \sum_{j=1}^{N}\log p(\boldsymbol{w}_j,\tau_j,\sigma_j|\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}') \\ \text{s.t.} \quad & \sum_{k=1}^{K}\pi_k = 1, \theta_{ik} > 0, a_k > 0, b_k > 0, a'_k > 0, b'_k > 0 \end{aligned}$$

In this paper we use the expectation-maximization (EM) algorithm to solve this problem.

V. Clustering Experiments

Methods: We compare the proposed K-MPLC algorithm with K-means algorithm, Gaussian mixture model (GMM) [2], and K-SC [25]. K-means, which is a general clustering algorithm, is considered as a baseline; if we assume each time-series vector $\boldsymbol{y}$ is generated i.i.d. from a multivariate normal distribution, then GMM would be a reasonable baseline as well; K-SC, which is the state-of-the-art algorithm for buzz time-series clustering, provides a more competitive comparison point. SpikeM [19], which is the state-of-the-art algorithm for buzz time-series modeling, is considered another baseline. As SpikeM does not handle time-series clustering, we concatenate all parameters extracted by SpikeM as features for K-means and GMM, and we denote them as *SpikeM+K-means* and *SpikeM+GMM* respectively. We also compare with a DTW-based anytime clustering algorithm [26], which is denoted as *DTW+Spectral*, whose basic idea is first to generate a DTW matrix to measure the non-metric distance between any two time-series, then apply a spectral clustering algorithm on the DTW matrix. In addition, we concatenate all the PLC Lasso features $\boldsymbol{w}$, τ, and σ and use them as features for K-means and GMM, and we denote them as *Lasso+K-means* and *Lasso+GMM*. We denote our proposed method as *Lasso+K-MPLC*, where we set α to be 1 and β can take values of $[0.1, 0.2, \ldots, 1.0]$ in our experiments. To avoid fluctuation due to random factors, we set the initialization of GMM and K-MPLC to be the clustering indexing of K-means. In addition, we run all experiments 10 times, and report the average evaluation metrics.

Dataset: To evaluate the effectiveness of the K-MPLC algorithm, we select thousands of high frequency search queries as candidate buzz topics and collect mentions from social media sites from June 22nd to August 8th, 2012. Considering the number of mentions of a topic per hour, we generated a time-series sequence for the topic within a time window. If the number of mentions at time t in a topic is 10 times higher than the average mention numbers in the past 48 hours, we regard the topic at that time as a buzz topic. According to [14], [4], hot topics on social media lose their attraction quickly, therefore, we select a time window of 72 hours, and obtain a 72-dimension time-series sequence $\boldsymbol{y}$ for each buzz topic. Finally, we collected the general buzz dataset, which contains 534 buzz time-series: most of them are celebrity names, such as *Michael Phelps* and *Tyler Perry*; the rest are event names, such as *Euro 2012* and *UFC 148*. In our experiments, time-series sequences are not aligned according to their peaks. That is, the data set used in this paper is a more challenging dataset than the one used for K-SC [25]. All time-series in the dataset are manually labeled into 5 clusters based on their distinct temporal shapes: time-series containing single sharp peak are labeled as 1; sequences with single heavy-tailed peak are labeled as 2; curves with double sharp peaks are labeled as 3; time-series containing double peaks (at least one is heavy-tailed) are labeled as 4; time-series with more than 3 peaks are labeled as 5.

Evaluation Metrics: We use Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI)[11], where both metrics are the larger the better.

Evaluation Results: Figure 3 (a) shows the clustering results of different algorithms on the general buzz dataset. In this experiment, each time-series sequence is modeled by a mixture of 3 PLC models. It clearly shows that *Lasso+K-MPLC* performs the best among all algorithms: its ARI and NMI scores are 0.2714 and 0.2628, which is much better than others. K-SC performs the second best among all algorithms. Although SpikeM provides good results for time-series modeling, its parameters are not effective as features for K-means or GMM. In order to figure out the effectiveness of PLC Lasso and K-MPLC separately, we also combine PLC Lasso parameters as features with existing clustering methods. With the help of PLC Lasso based features, *Lasso+K-means* and *Lasso+K-means* compare favorably with K-SC, and their metrics gaps to *Lasso+K-MPLC* indicates the effectiveness of K-MPLC.

To further compare the proposed algorithm with other baseline approaches, we change the number of existing cluster labels in the data set to obtain more experimental results.

- **4-label clustering ground truth:** we combine the sequences originally labeled as 3 and 4 into the same cluster, while the remaining is the same.
- **3-label clustering ground truth (different number of peaks):** we combine the sequences originally labeled as 1 and 2 into one cluster, combine the sequences originally labeled as 3 and 4 into another same cluster, and leave the remaining the same.
- **2-label clustering ground truth (Single peak vs. multiple peaks):** we combine the sequences originally labeled as 1 and 2 into one cluster, and combine the sequences originally labeled as 3, 4 and 5 into the other cluster.

Comparing with Figure 3 (a)-(d), overall, the observations are quite consistent: the Lasso based approach shows its effectiveness on K-means and GMM; DTW + Spectral usually performs the 2nd best among all, and the DTW based approach is more robust than K-SC on this dataset; K-MPLC always performs the best. Notice that, with decreasing of cluster numbers K, the difficulty of clustering also decreases. In our experiment, the absolute evaluation metrics of K-MPLC is increasing (ARI is 0.2930 when $K = 5$, and ARI is 0.3857 when $K = 2$), which also shows the effectiveness of our proposed method. On the other hand, K-SC and DTW+Spectral perform worse when K is smaller, such as $K = 2$. One possible reason is that one of two clusters consists of single sharp peak and single fat peak sequences. More specifically, for K-SC and DTW+Spectral, the distance between single sharp peak sequences and single fat peak sequences can be larger than the one between single sharp peak sequences and multiple peaks sequences, and thus single peak and multiple peaks tends to be clustered together.

Figure 4 illustrates the clustering results with different PLC numbers L. In this experiment, it is clear that each sequence with a mixture of 3 PLC models consistently outperforms that of using 2 or 4 PLC models. One possible explanation is that in our general buzz dataset, as the ground truth is to differentiate curves with 1 peak or 2 peaks from 3 or more peaks, representing each time-series sequence as a mixture of 3 PLC models is the most appropriate, while leveraging 4 PLC models might bring in more noisy information, and leveraging 2 PLC models are least robust.

VI. Conclusion and Future Work

In this paper, we presented our study to model the sudden spikes and heavy-tailed patterns of buzz events on social media with product life cycle models (PLC). Specifically, we model a time-series sequence with a mixture of PLC models, and propose an efficient Lasso based parameter estimation approach. Then, we proposed a novel probabilistic model to cluster buzz time-series with similar temporal patterns into the same group based on the features obtained from the mixture of PLC models. A novelty of the proposed approach is that it can distinguish a buzz time-series sequence with a sharp peak from a buzz sequence with a heavy-tailed peak, and our proposed method significantly outperforms the current state-of-the-art algorithms on the buzz clustering tasks.

Our work can be extended as follows. First, we can use a group regularizer instead of the $L1$ regularizer to model buzz time-series, which is the most promising future direction. Second, since the Gamma distribution function is not suited for modeling the slow rising trend in the introduction stage or growth stage of a product, it would be interesting to explore other PLC models. Third, in this paper, we only deal with a segment of time-series within a fix size of time window. In the future, we will leverage PLC model to split a long time-series sequence into different segments, and combine with social media content to mine more interesting findings.

ACKNOWLEDGMENTS

This research was supported by the NSF research grants IIS-1134990, and U.S. Defense Advanced Research Projects Agency (DARPA) under Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-1-0034. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency, or the U.S. Government.

References

[1] G. Batista, X. Wang, and E. J. Keogh. A complexity-invariant distance measure for time series. In *Proceedings of SDM*, 2011.

[2] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. springer New York, 2006.

[3] W. R. Blischke and D. N. P. Murthy. *Reliability Modeling, Prediction and Optimization*. John Wiley and Sons Inc, 2000.

[4] Y. Chang, L. Tang, Y. Inagaki, and Y. Liu. What is tumblr: A statistical overview and comparison. *SIGKDD Explorations*, 16(1), 2014.

[5] H. Choi and H. Varian. Predicting the present with google trends. In *Technical report, Google*, 2009.

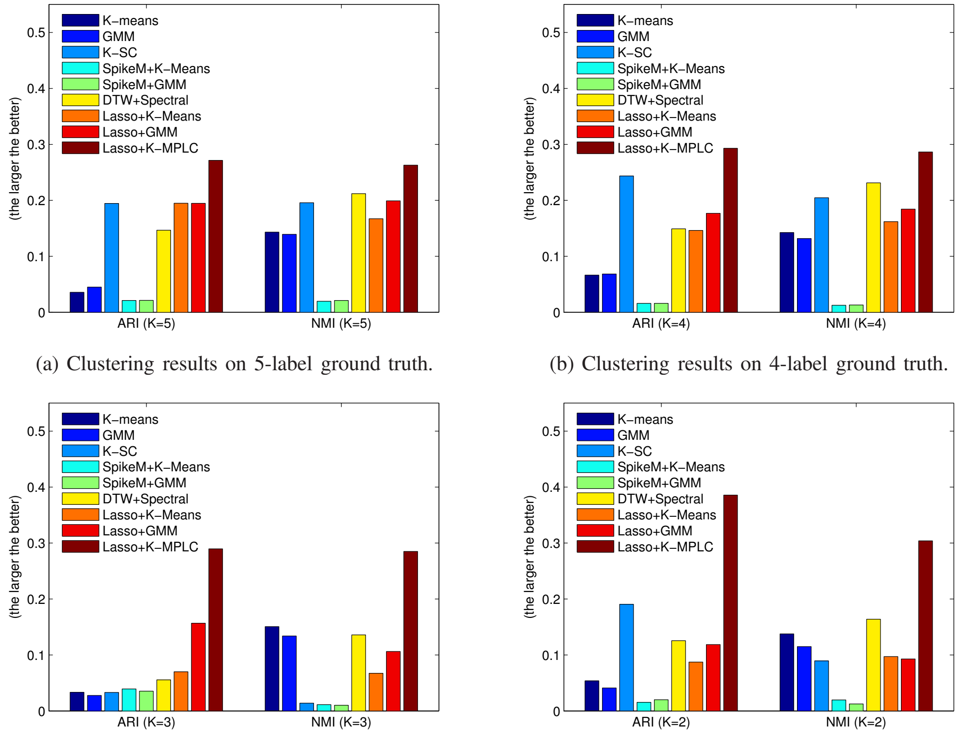
[6] D. Gruhl, R. V. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of KDD*, 2005.

[7] M. Guerini, C. Strapparava, and G. Ozbal. Exploring text virality in social networks. In *Proceedings of ICWSM*, 2011.

[8] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *Proceedings of KDD*, 2011.

[9] A. Isaic-Maniu and V. G. Voda. On a model regarding the product life-cycle. *Management and Marketing*, 3(3):87–96, 2008.

[10] S. Jiang, J. Ferreira, and M. C. Gonzalez. Clustering daily patterns of human activities in the city. In *Data Mining and Knowledge Discovery*, pages Volume 25, Number 1, 2012.



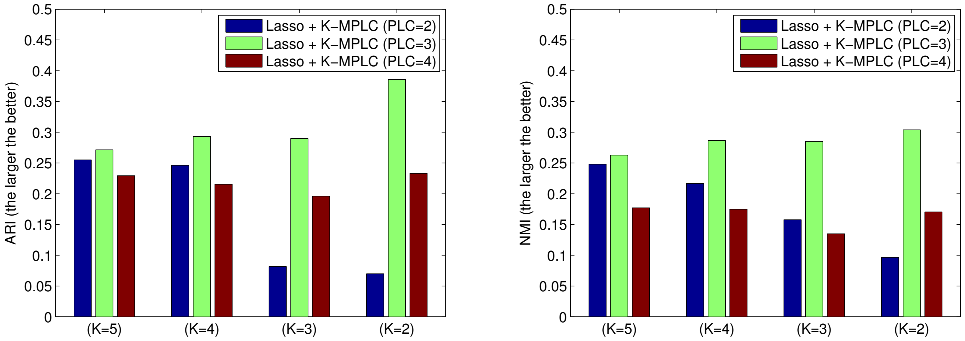
(a) Clustering results on 5-label ground truth.

(b) Clustering results on 4-label ground truth.

(c) Clustering results on 3-label ground truth.

(d) Clustering results on 2-label ground truth.

Fig. 3. Clustering results on General Buzz Dataset with different Labels.



(a) ARI over different number of PLC numbers.

(b) NMI over different number of PLC numbers.

Fig. 4. Clustering Results on the General Buzz Dataset with Different PLC Numbers.

[11] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 2005.

[12] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *Proceedings of KDD*, 2010.

[13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of WWW*, 2003.

[14] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media. In *Proceedings of WWW*, 2010.

[15] K. Lerman and R. Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In *Proceedings of ICWSM*, 2010.

[16] T. Levitt. Exploit the product life cycle. *Harvard Business Review*, 1, November 1965.

[17] L. Li and B. A. Prakash. Time series clustering: Complex is simpler. In *Proceedings of ICML*, 2011.

[18] J. Lin, M. Vlachos, E. J. Keogh, and D. Gunopulos. Iterative incremental clustering of time series. In *Proceedings of EDBT*, 2004.

[19] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of KDD*, 2012.

[20] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of WWW*, 2006.

[21] T. Rakthanmanon, E. J. Keogh, S. Lonardi, and S. Evans. Time series epenthesis: Clustering time series streams requires ignoring some data. In *Proceedings of ICDM*, 2011.

[22] J. Sun and J. Tang. A survey of models and algorithms for social influence analysis. In *Social Network Data Analysis*, pages 177–214, 2011.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[24] R. Tomioka, T. Suzuki, and M. Sugiyama. Super-linear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation. *JMLR*, 12:1537–1586, 2011.

[25] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of WSDM*, 2011.

[26] Q. Zhu, G. Batista, T. Rakthanmanon, and E. Keogh. A novel approximation to dynamic time warping allows anytime clustering of massive time series datasets. In *Proceedings of SDM*, 2012.