What's in a 'Meme'? Understanding the Dynamics of Image Macros in Social Media

Bodhisattwa Prasad Majumder
IIT Kharagpur
bodhisattwapm2017@email.iimcal.ac.in

Amrith Krishna
IIT Kharagpur
krishnamrith12@gmail.com

Unni Krishnan
The University of Melbourne
unnik@student.unimelb.edu.au

Anil Kumar Boga
IIT Kharagpur
bogaanil.009@gmail.com

Animesh Mukherjee
IIT Kharagpur
animeshm@gmail.com

ABSTRACT
Image memes (simply memes) are becoming increasingly popular over social media. In this work, we for the first time, investigate the evolution and life-cycle of an online community that uses image based memes as their primary means of communication. We find that the frequency of various image memes that are reused within the community follows Zipf-Mandelbrot distribution. The community in its initial days borrows image memes popular over the Internet, but gradually emerges with memes local to the community, finally surpassing the popularity of the former. These memes tend to exhibit temporal variations in their usage patterns. We characterize each day by their temporal activity and popularity. The interaction between memes across days leading to active user engagement within the community is captured using HDP-HMM, a non-parametric Bayesian variation of the infinite HMM. We group the observed days where the states encode possible 'moods' on the observed days. With the model, we also establish the significance of *familiarity* vs. *freshness* which is key to the growth and evolution of the community. Finally, using the adaptor grammar framework, we identify motifs from the community activity and and use it to predict the 'mood' of a day based on the 'moods' of the previous days.

Author Keywords
Social networks, memes, image macros, HDP-HMM, Adaptor Grammar

INTRODUCTION
The rise of Online Social Networks (OSN) has opened up new opportunities for understanding collective human behavior. Emergence of conventions [20] and linguistic style accommodation [8] in OSNs are well established. The evolution of interaction norms in online communities [9] and the temporal variations in content popularity [3, 31] are well studied as well. While, the aforementioned studies focused on communities that interact primarily through a textual medium, we investigate *the evolution of an online community, where its members communicate primarily through image based memes*. In particular, we try to understand 'a typical day in the world of meme'. We look into the interactions and the variations in terms of content usage of *Troll Malayalam*[1], a community centered around a Facebook page with close to 1 million subscribers catering to a specific demography.

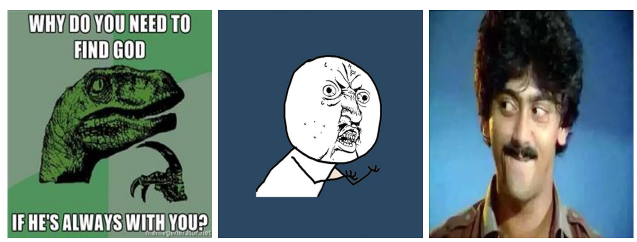


Figure 1: Sample *image macros* overlaid with humorous texts; the third image is a sample used in *Troll Malayalam*, borrowed from local culture.

Content structure in *Troll Malayalam* - A typical image post in *Troll Malayalam* contains two separate entities, (i) an image with (ii) text superimposed on it. The written text is adept in conveying the content of the discussion, but the image sets the tone for the content. We show some typical examples of such image memes posted in such online communities in Figure 1. The image here, henceforth to be referred to as *image macro*, is often reused as can be seen from Figure 2. This is analogous to the concept of 'memes' (cultural correlate of genes) coined by Richard Dawkins [10]. The definitions of all the technical terms used in this paper are presented in Table 1. Similar to popular web-sites like *9gag*[2] and *4chan*[3] which are solely based on troll creation and propagation mostly based on

[1] https://www.facebook.com/Troll.Malayalam/
[2] https://9gag.com/
[3] http://www.4chan.org/

Term	Definition
Content Image	A picture with text super imposed on top of it. The text conveys the content, image sets the tone.
Image macro	A template image over which text is super imposed to create the ‘content image’. Multiple image contents can be generated from a single image macro. Though in general popular image macros are called as meme, we interchangeably both the terms.
Vocabulary	The set of all the image macros in *Troll Malayalam*.
Corpus	The set of all the content images in *Troll Malayalam*.
Post image	The content image posted by the *Troll Malayalam* page. The ‘Post image’ exists independently in the page.
Comment image	The content image posted by a user. A comment image will always be linked to one of the Post Images.
Macro class	Categorization of an image macro based on the temporal properties expressed by the content images generated from the image macro.
Classic macro	Image macros from which content images are generated multiple times and the time gap between the last and the first occurrence is very high, typically above a threshold.
Cult macro	Image macros that are never used to generate a post image, but always a comment image.
Hypergiant macros	Image macros from which content images are generated multiple times but the gap between all the successive enunciations are below a threshold.
Singleton macros	Image macros from which only a single content image is generated.
Global meme	Content images generated from image macros which are popularly used in the Internet not confining to the demographic of the community.
Local meme	Content images generated from image macros which are not categorized as a global meme.

Table 1: Definitions for the technical terms used in this paper.

image or video media. *Troll Malayalam* also focuses mainly on the content dynamics - all published from the page’s own profile. The dynamics of the image memes are not influenced by the popularity of the users interacting with the posts through likes or comments. While in web sites like *9gag* etc., image memes are just one of the means of expressing humor/sarcasm [30], the *Troll Malayalam* community uses image memes as the predominant mode of communication on various topics specific to the demography. Amongst all the responses obtained for the posts by the page, about 57% of the comments were images, and about 76% of the comment likes were shared by the image memes, establishing that image memes are the primary means of communication in this community.

Figure 2: Typical example of *image macro* getting reused with modification.

Figure 3 shows the posting activity in the community for different classes of macros. All the definitions of different class of macros are provided in Table 1. It shows the existence of *image macros* with variations in its temporal patterns over days, similar to the communities based on textual medium [31].

This motivated us to investigate further the temporal characteristic of the *image macros* across the days they were posted. We seek to understand how the joint activity of these various classes of *image macros* define a day, or rather the ’mood’ of the day. To understand these hidden dynamics that create the sustainable user engagement via communication through

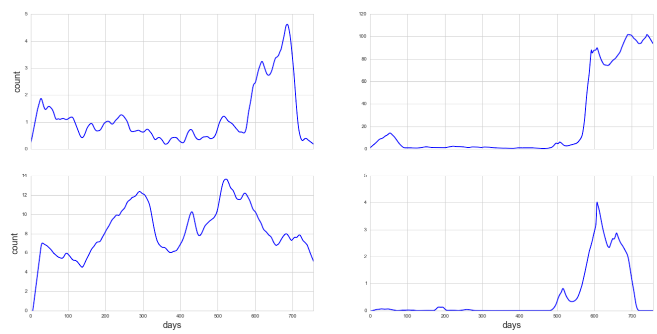


Figure 3: The plots denotes the activity of different classes: Classic, Cult, Hypergiant and Singleton (clockwise, starting from top left).

image macro based memes over different days, we organize our experiments into three major parts as follows.

States of days - It is evident from Figure 3, that the community has *image macros* exhibiting variation in their usage patterns. We also observe different days have different patterns of usage of these *image macros*. We seek to group all the observed days into a finite set of states defined by the temporal activity and popularity of the *image macros*. Abstracting the observed days to a finite set of groups is beneficial in understanding the *universalities* exhibited by those days, thereby, identifying what kinds of *image macros* make a day popular and are pertinent to the communication happening in the community.

Freshness* vs. *Familiarity - The community has a constant influx of new *image macros*, but the community also re-uses old *image macros* as well. Different levels of usage of such macros in different days makes each day different in nature, i.e., renders a ‘mood’ to a day. In this context, we hypothesize that there is a right mix of freshness vs. familiarity throughout a day that engages the users continuously and leads to success of the community. Further the sequence of states over the days is also interesting since it sets the overall 'mood' of the community.

Motif Identification - We aim to find recurring patterns from the data, that are strong indicators of the community behavior. We see the motifs as functional subunits that characterize the community behavior. The motifs are essentially subsequences of days which are recurring in the data. With the motifs, we can find variable length subsequences, that essentially characterize the community's interaction. We further attempt to understand how, for each individual day, different sub-sequences arise out of interactions between the macros of various temporal profiles. Finally, as an application, we predict the probable 'mood' of the next day observing the 'moods' of previous days.

Our analysis of the community can be essentially categorized into the following three phases

- The interaction among various days is a temporal pattern recognition task, which can be modeled with an HMM once the various states of days based on the temporal activity and popularity of all *image macro* classes are known. However, since the number of states themselves are not known a priori, we jointly solved the state identification and temporal interaction pattern recognition problem using HDP-HMM [15].
- We address the motif identification problem using the adaptor grammar [17] framework which is a non parametric Bayesian approach for learning productions from observations in a PCFG. The observations in this context are the sequences of various states of days. The productions so learned are sub-sequences of varying length from the HMM sequence. We further show that the grammar learns regularities inherent in the data by comparing the likelihood of subsequences of held out data with that of randomly generated subsequences.
- Finally, given the grammar learned, we predict the probable state of the next day given the history of previously observed states. The predicted state of the next day can lead to an understanding of what kind of memes will be appreciated on the given day. We obtain a high MRR of 0.69 in predicting the mood of a given day.

To the best of our knowledge, this is the first ever study to be conducted on evolution of a community that uses a non-textual means for its communication.

RELATED WORK

Study of collective human behavior in collaborative environments has drawn a lot of attention from social sciences community and with the advent of web 2.0, large-scale studies for computational social sciences became feasible. [8] studied the linguistic accommodation in social media and found how users converge to each other's behavior. Overall change in communication style within a community with emphasis on community specific language innovation is accounted in [18]. In [9], the authors observed that the users in a community go through a two stage life-cycle, in which they initially adopt the language of the community and later stop to evolve. In [20], the authors predicted the emergence of dominant conventions in OSNs. [1] presented a large scale analysis of spread of information across OSN and uncovered the regularities in evolution of information content due to imperfect copying mechanisms. [7] studied the characteristics of image memes in isolation, and found that the uniqueness of a meme leads to increased popularity of the meme. In this work, we take a complimentary direction to their work and study the community evolution in terms of content dynamics, and show how the interaction of various memes can lead to better community engagement.

Works which attempt to group contents based on temporal patterns of individual contents over time are abundant in the literature. [2] used a dynamic time warping method to find whether triggering of search query from one source could be traced back to the same at some other source. Yang et. al. [32] proposed a generative model which captures progression stages in an event sequence. In [31], the authors proposed a time series clustering approach to group content based on the similarity of temporal patterns they exhibit. In [3], the authors used an affinity propagation clustering approach by quantifying the rate of change in attention and share of attention over a time window. Both the works [3, 31] used a non-parametric clustering approach which does not require a priori knowledge about the possible number of clusters. Additionally, [3] proposed a prediction model based on estimating transition probabilities between the obtained clusters, which capture the temporal transitions in behavior expressed by the community.

We jointly perform this task of identifying characteristic clusters and temporal transition between the identified clusters in our work. For this purpose we used HDP-HMM [29] that has been employed in tasks like gene expression time course data clustering [5], unsupervised word segmentation [13] in natural languages and audio synthesis [14] among many others. Further, [22] have used HDP-HMM to model movement of vehicles in a traffic junction and then infer meaningful rules from the sequence of latent states in the sequence.

Finally, adaptor grammars have been very effectively used in numerous NLP related tasks [24]. Recently,[34] have used adaptor grammars for identifying entities from shopping related queries . Mark Johnson, has drawn connections between topic models and PCFGs and then proposed a model with combined insights from adaptor grammars and topic models [16].

DATASET

We analyzed data from *Troll Malayalam*, a Facebook page with nearly a million subscribers, primarily from the Malayalam speaking demographic. The page posts are user generated contents curated based on an internal voting within a core group, and the comments are open for any page visitor without any moderation. We consider this community due to the numerous unique opportunities this community provides for understanding the content dynamics within the community.

1. The activity inside the page is self-contained and the content usage is primarily centered around the page itself. Therefore, popularity (in terms of the number of 'likes', con-

tent re-use) of individual contents can be attributed to the activities primarily from within the community itself and exogenous factors are hardly present.

2. Unlike global memes, we can trace back to the origins of the contents generated by the community, as the page preserves complete activity of the community right from its inception in July 2013. This helps us to track the exact time evolution of the contents and how the community preferences changed over time.

3. Since the community uses image macros with texts, we can easily separate out the image macros which has the tendency to spread as opposed to the textual content which is more post-dependent and therefore context-specific. This enables the study of the meme aspect of the image macros within a stricter framework in similar lines as [7].

From *Troll Malayalam*, we crawled the complete page activities starting from its inception in July 2013 to February 2016, over a period of 30 months. From the analysis of the page activities, we find that about 57% of the comments were images. However, strikingly those 57% of comment images account for 76% of the total comment likes. Since we are interested in the engagement dynamics induced by the memes and since the community uses image macros as memes, we removed all the text comments. In total, we collected 31,805 images of which 7,248 were post images and the remaining were comment images. We grouped images that were using the same image macros through a semi-automated mechanism – using the "findimagedups"[4] tool followed by manual supervision. With the grouped images, which are referred to as templates henceforth, we proceed to define a day in terms of temporal activities of these grouped images or macros. In a subsequent section we shall describe how these macros were grouped via a rule based method and then observed per day to identify their activity over days.

Preliminary statistics from the dataset

In this section, we perform a preliminary analysis of the dataset and observe that *image macros* exhibit similarities to natural language units. These observations lay the foundations of the linguistically grounded framework that we use in the subsequent sections.

Frequency of *image macro* usage: Analogous to natural languages, we consider the entire set of *image macros* to be the *vocabulary* of the community, where each unique template is a *type* and each instance of the *image macro* used, is a *token*. We find that the frequency of usage of *image macros* follows a finite Zipf-Mandelbrot (ZM) law, a generalization of Zipf's law.

$$g(p_i) = \frac{1-\alpha}{(B^{1-\alpha} - A^{1-\alpha})} p_i^{(-\alpha-1)} \quad (1)$$

Equation 1 denotes an *LNRE* (Large Number of Rare Events) [19] re-formulation of *ZM* with a finite vocabulary which is given by the type density function $g(.)$ [11]. Figure 4 illustrates the distribution of vocabulary usage that follows the finite Zipf-Mandelbrot law. The parameters α, A and B here are equal to 0.0409, $8.6553e-08$ and 0.0042 respectively.

Local and global memes: We further analyze the community behavior to understand the evolution of the local culture. The vocabulary so built from our dataset shows the presence of two major classes of image types. The first type comprises image macros which are popular globally and are borrowed by the community, henceforth to be referred to as the *global memes*. The second type comprises the new image macros generated by the community members based on the images from movies and mainstream media relevant to the demography. Such community generated image macros will be termed as *local memes*. The entire dataset was annotated by *six* annotators independently. The annotators mark whether each image macro is a global meme or a local meme. If a particular token contains modification of existing global memes, the annotators still mark it as global. The global memes were further validated by a set of *two* annotators, of which one annotator is an author. We retained only those images, where both the annotators agreed that the meme was global. The annotations resulted in altogether a set of 28730 local memes and 1448 global memes.

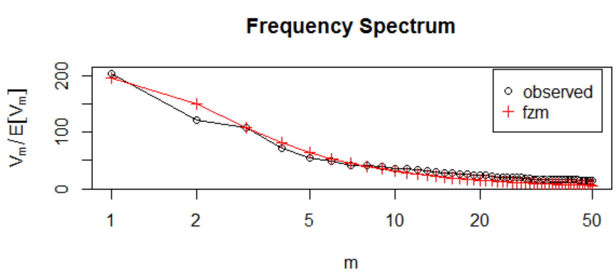


Figure 4: Distribution of the vocabulary usage follows the Zipf-Mandelbort law. The fitted curve is marked in red and the the observed distribution in black.

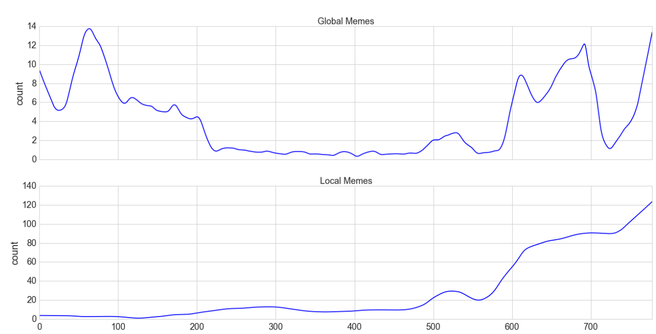


Figure 5: Temporal plot of global and local vocabulary.

Further analysis on overall activity shown that the community activity close to the middle of June 2015 observed a giant leap. It was observed that while the rate at which the post memes

[4]https://github.com/opennota/findimagedupes

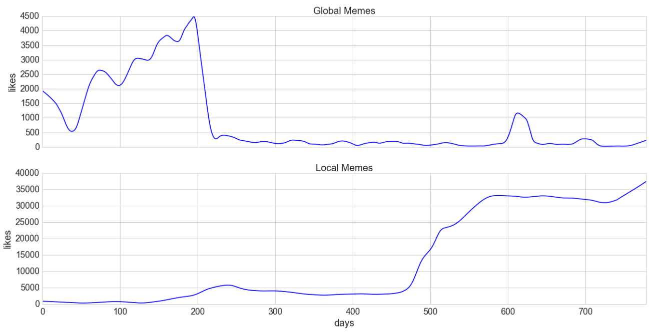


Figure 6: Temporal plot of likes gained by global and local macros.

published remained constant, while the average number of comments posted in the page increased. For the month of may the average number of comments posted per day was 6.32, while that on the month of June was 84.40. This is indicative of the acceptance gained by the community to a larger audience. Figure 5 shows the induction of new memes, i.e. per day vocabulary induction, for both the local and the global meme categories. It can be observed that the community started by borrowing memes popular over internet, which we call as global memes, but eventually the number of local memes outgrew while the vocabulary growth of global memes became constant. We find that the daily induction of local memes became more than that of global memes from the 211th day, where a dip in the share global memes is visible. During the first 210 days, the vocabulary size for global memes were 1314, while that of local memes were 144. Figure 6 presents the popularity of the image macros based on the likes each of them gathered. We see that while the local images eventually becoming the norm, there is still a non-negligible number of likes going to the global memes.

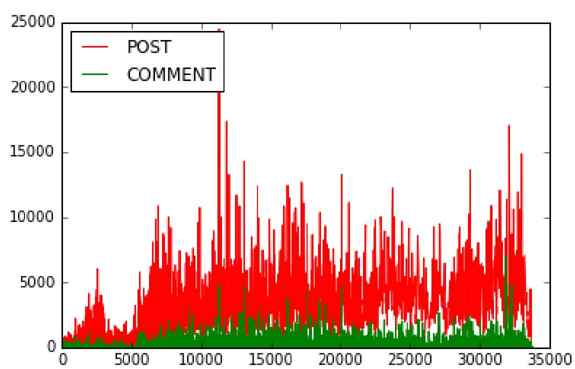


Figure 7: Normalized likes for post vs. comment macros.

To measure popularity, we compute the volume of likes obtained from the post and from the comments to a post separately. Naturally, even though both these entities are visible on the same page, the visibility of the original post is way more higher than the comments generated from the post. Thus, the volume of likes for the post itself is usually much higher than for the comments. For better visualization, we made these two temporal patterns (i.e., number of likes to the post vs. its comments over time) scale invariant using an approach outlined in [31]. The scaling co-efficient is ~ 6.9. The result after this scaling is presented in Figure 7. Further, in this figure, to neutralize the effect of content that arises due to relevant social contexts we apply a moving average normalization which enables further smoothing. These observations motivate to model the regularities present in the observed sequence day-wise and finalize over finite number of states based on the various composition of temporal activities from various image macros which collectively describe various 'moods' of the observed days.

TIME SERIES ANALYSIS OF PAGE ACTIVITY

Since we study the community activity from a Facebook page, we assume that every visitor to the page gets to see the page contents in a sequential fashion. In addition, the posts on the page always appear in the name of the community itself. Hence, there is no notion of a separate contribution of the influence of an user to the popularity of a post and therefore this factor can be eliminated. Further, from our manual inspection of the data, we did not find any traces of user specific influence on the popularity of a content. In particular, there were no strong indicators for (i) the like patterns, (ii) the user-mentions, or (iii) cliquish interactions across the community validating our hypothesis that the content of the meme is a key factor in its overall reach within the community.

We formulate the page activity as a time series T_p on the *image macros*. Let t_i denote the time-stamp at which an *image macro* m_{t_i} appeared in the page. Then $T_p = \{m_{t_i} : t_i \in T\}$ where $T = \{t_i : t_i < t_{i+1} \forall t_i\}$.

We perform a daywise segregation of T_p and derive a second time series T_D where each day is represented by a combination of posting activities of various *image macro classes*. Formally

$$T_D = \{f : \mathrm{MC}_d \to \mathbb{F}_d : d \in D\}$$

such that,

$$\mathrm{MC}_d = \{\mathrm{mc}_{t_i} \ldots \mathrm{mc}_{t_{i+n_d}} : t_i \in T, \bigcup_{k \in 0 \ldots n_d} t_{i+k} = d\}$$

where, D represent the set of days so that $t_i \ldots t_{i+n_d}$ cover the span of a day d, mc_{t_i} denote the *image macro class of* m_{t_i} and the function f maps the set MC_d to the set of features $\mathbb{F}_d$ for d as given in Table 2.

This approach is analogous to a sequence of characters in natural language processing tasks, or sequence of genomic strands. By definition, each day constitute points that are uniform in time. Within a day, we preserve the order of the contents that are posted in the page, whether it be a page post, or a comment to a post. We characterize a day by the temporal postage behavior of these macros along with their popularity.

Sequence modeling

With the definition of T_D in place, we propose a generative model where the days are seen as manifestations of one of the latent moods. Hidden Markov Models (HMM) [25] are an effective means of modeling such instances of time series data. But, the cardinality of the latent state space is unknown. Prior to modeling the HMM, We need to find clusters of days where days with similar temporal profiles are grouped into a finite set of states S. But we choose to solve the problem of clustering the days and modeling the state transitions jointly using Hierarchical Dirichlet process HMM (HDP-HMM) [29], a non parametric Bayesian extension of HMM with infinite states. This helps us to eliminate the need to have apriori knowledge on the number of latent states(types of moods).

In an infinite HMM, where there are potentially unbounded sets of states, the transitions are determined by Dirichlet process (DP) mixtures rather than by finite mixture distributions. To make sure the DP mixtures of states share components between them, a Hierarchical Dirichlet Process (HDP) [29] is used. Equation 2 summarizes the HDP-HMM model [12].

$$\begin{aligned} &\beta \mid \gamma \sim GEM(\gamma) && \\ &\pi_j \mid \beta, \alpha \sim DP(\alpha, \beta) && j = 1, 2... \\ &\theta_j \mid H, \lambda \sim H(\lambda) && j = 1, 2... \\ &z_t \mid \{\pi_j\}_{j=1}^{\infty}, z_{t-1} \sim \pi_{z_{t-1}} && t = 1, ..., T \\ &y_t \mid \{\theta_j\}_{j=1}^{\infty}, z_t \sim F(\theta_{z_t}) && t = 1, ..., T \end{aligned} \tag{2}$$

Here $GEM(\cdot)$ represents the stick breaking construction of the Dirichlet process [27]. β, the root probability vector which defines a multinomial distribution over the set of states. For each state, a vector π_j is drawn from a DP, which is the transition vector from the state j to all other states. For the emission probabilities from the latent state to the observed state the parameter θ_j is drawn from a distribution prior H. Here we assume that the model generates the observed sequence of days y_t and the latent state sequence z_t. Each observation y_t is drawn from the emission probabilities $F(\theta_{z_t})$, where z_t is drawn from $\pi_{z_{t-1}}$, i.e., from the previous latent state from the state sequence. We assume z_1 is drawn from β [14, 12, 15]. Finally, we obtain the sequence of latent states $z_{1...t}$, where each state ideally should represent a group of days exhibiting the similar 'mood'.

Feature construction for HDP-HMM: Based on the temporal characteristics such as lifespan, postage behavior e.g., post or comment images, we grouped all the *image macros* into four macro classes as defined in Table 1. Here we explicitly mention how different classes have been derived with the exact thresholds used based on the available data.

1. *Classic macros* - *image macros* having lifespan (i.e., time of last occurrence - time of first occurrence, normalized) greater than 0.05 are termed as classic macros. We consider only top 7%-ile macros in terms of lifespan setting the threshold as 0.05. This group corresponds to the set of macros that persisted in the community for a very long time.

2. *Cult macros* - *image macros* that only appeared as a comment corresponding to a post are classified as cult macros. This set of macros signify that they have not been curated by the community administrator but spontaneously appeared to advance the flow of the discussion under a certain post.

3. *Singletons* - *image macros* that appeared only once throughout the time span considered. These are very specific macros that were used for some very particular reason and failed to retain their popularity in the subsequent activities within the community.

4. *Hypergiant macros* - *image macros* that are not classic, cult or singleton have been grouped as hypergiant macros. These macros are curated more than once but could not sustain as long as the classic ones.

Different classes of macros exhibit a variety of behavior even within a day. Hence to represent a particular day, we seek to incorporate the features corresponding to the activity of the different classes of macros posted on that day. We have considered five temporal features for each class of macros per day which include (i) the total number of appearances normalized by the total macros posted on that day, (ii) the mean lifespan during the day, (iii) the log of the mean arrival time, (iv) the standard deviation of the inter arrival time and (v) the skewness of the inter arrival time. The definitions and the intuition behind the applicability of the features is described in Table 2. Apart from the temporal features, we have included the log of the number of likes for each macro category which denotes category wise and overall popularity of the day. We also considered the entropy considering all classes of macros per day. The multi-class entropy calculation is performed based on the following formula -

$$Entropy = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \tag{3}$$

where n is the number of classes present (four here). We calculate the $P(x_i)$ empirically from the sequence of appearances of the different classes of *image macros*. In total, we have 25 features to represent a typical day which in consequence, evoke a multivariate time series appropriate to model via the HDP-HMM framework.

Motif identification

In order to capture the regularities expressed in the temporal pattern of the state predictions, we further investigate whether we can identify conventions and norms that has organically evolved within the community. We essentially look for patterns exhibited by the sequences of the states with a grammar based formalism, and treat it as a grammar induction task. The objective is to find motifs or sub-sequences the can succinctly describe the major activity patterns within the community. We use adaptor grammars [17] for motif identification, where the framework learns a grammar structure from the state sequence obtained. The grammar A_g obtained from the adaptor Grammar is a probabilistic context free grammar, where the productions from a set of fixed non-terminals and the probabilities for the productions to be invoked are learned from the data [21]. An adaptor grammar is a six tuple (N, w, R, S, θ, C), where θ is the probability vector for each of the productions, and C

Feature name	Description
Normalized macro count	For each *image macro class*, the total number of appearance divided by the total number of appearances of macros from all the classes.
Mean Lifespan over the day	Mean Lifespan of each *image macro classes* . For an *image macro class* *mc*, $$\text{Meanlifespan} = \frac{t_n - t_0}{\text{Count}(mc_d)}$$ where, t_n is the timestamp of the last appearance of *mc* in a day and t_0, the first. For cases where the number of appearances of the specific class of macro is zero or one, this value is set to 0.
Log(mean inter arrival time)	If two consecutive time points when a particular class of image macro appears are t and $t+\Delta t$ then the inter arrival time in Δt. For cases where $\text{Count}(mc) = 0$ or 1, this value is set to some large value (10 here).
Standard deviation of inter arrival time	The standard deviation of the inter arrival times for all occurrences of the specific class of macro throughout the day. Captures the variability of the several appearances of the class of macros. For cases where the number of appearances of the specific class of macro is zero or one, this value is set to 0.
Skewness of the inter arrival time	Captures the skewness of the inter arrival time. On some days the activity of a specific class of macros could be heavy at the start or the end of the day, which can be nicely captured by this third order moment. For cases of zreo or one appearances, this value is set to 0.
Log of likes gained	The log of total likes gained by the specific class of *image macro*s over the day
Entropy of the day	The multi-class entropy of the day considering presence of all class of macros in a day (Equation 3)

Table 2: Features given to HDP-HMM along with their description and applicability

is a vector of adaptors indexed by the non-terminals N [17]. The productions are sub-sequences from the HDP-HMM state sequence.

```
@motif -> states
states -> state
states -> states state
```

Figure 8: unigram grammar

```
@motifs -> motif
@motifs -> @motifs motif
motif -> states
states -> state
states -> states state
```

Figure 9: collocation grammar

The grammar so learnt describes the dominant activity patterns in the community. In Adaptor Grammar, a skeletal context free grammar is defined as shown in Figure 8 and Figure 9, where the set of non-terminals to be adapted is fixed a priori and will be a subset of the entire set of non-terminals in the skeletal grammar. For each of the adapted non-terminal, marked with a '@', the grammar learns a distribution over trees rooted at each of the adapted non-terminal [33, 21]. For the adaptor grammar to learn the productions we used a sliding window approach and provided sequences of a fixed length at a time. We used an unigram grammar and a collocation grammar both of which are provided in [17]. The unigram grammar looks for sub-sequences of varying length based on its likelihood of presence from the input sequences, while the collocation calculates a co-occurrence likelihood of a sub-sequence given the likelihood of another sub-sequence in the context.

State prediction

Using the grammar learned in the previous section, we predict the next state given k previous state sub-sequences. We formalize the state sequence prediction as the problem of predicting the next state given the a set of previous states. If S is the set of (unique) states and given a grammar model A_g, our state sequence prediction problem can be defined as:

DEFINITION 1 (STATE SEQUENCE PREDICTION). *Predict* s_{i+k+1} *given* A_g *and* $s_i \ldots s_{i+k}$, $\forall i, s_i \in S$

We generate sub-sequences from the HDP-HMM state sequence and for each sub-sequence $s_i \ldots s_{i+n}$ of length n, a set of $|S|$ subsequences of length $n+1$ are generated by appending each of the states s_t from S. The modified sub-sequence will be $s_i, \ldots, s_{i+n}, s_t, \forall s_t \in S$.

These sub-sequences were parsed by the grammar A_g and were ranked based on decreasing likelihood. The state s_t from the highest ranked sub-sequence is given as the predicted next state in our model. The performance of this prediction model, thus, also depends on the output of the HDP-HMM and how well it captures the temporal activity and popularity of various classes of macros. In other words, having observed a certain sequence of days, we attempt to predict how the next day should shape up in terms of the 'mood' or tone of the discussion inside the community. The community administrator can follow these rules to inject and propagate ideas in more efficient way by channeling right content which can bind the users to the community. Given the original state sequence of all the days in the data, it is a hard problem to identify the maximum repeating sub-sequences of any length which could have been used as norms or rules of the community activity. Our prediction process provides most likely sub-sequences which also convey the overall community behavior. These sub-sequences can be thought of as encoding the *universal* norms and conventions characterizing this community.

RESULTS

We form our time series T_D of 802 observations (days) with data from Troll Malayalam. We first input T_D to the HDP-HMM to obtain the hidden state predictions corresponding to each of the observation. We then use the state predictions from HDP-HMM to learn a grammar A_g using adaptor grammar. We perform a series of experiments to validate the effectiveness of our proposed approach in capturing the community dynamics. First, We perform feature analysis over the states predicted by HDP-HMM, which brings out the trends regarding segregation of the days to each of the hidden state. To evaluate whether we are able to capture the regularities in the community dynamics, we use A_g to predict the likelihood of sequences held-out from the data and then compare it from a set of randomly generated sequences sampled from a normal distribution. We use the likelihood prediction as an intrinsic evaluation approach to show the effectiveness of our overall framework, as we lack any ground truth for our dataset.

States of the days

We experimented with first order autoregressive Gaussian likelihood and Gaussian observation models to fit our data with HDP-HMM. We find that our data is better explained by the Gaussian models. Stickiness is a parameter introduced by [12], which encourages a sequence to remain in the previous state for successive observations. In our case, we find that reducing the stickiness improves the segregation of the days to different states. Table 3 shows the parameter settings as explained in [15] for our best performing model.

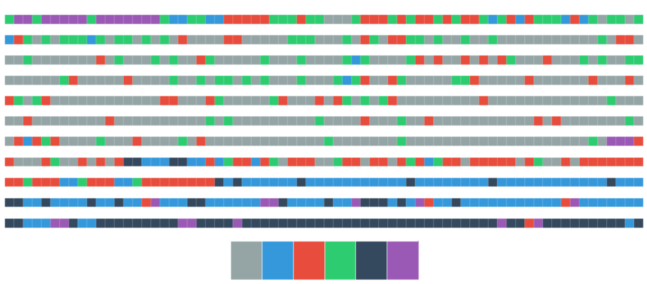


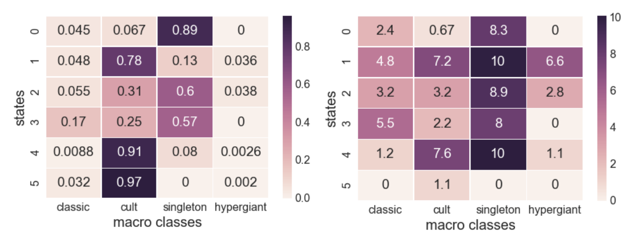
Figure 10: Occurrences of various states over all the days; the second image denotes the color coding for six states (0 through 5) starting from the left.

With the parameter settings as mentioned in Table 3, we obtain a model with 6 hidden states from our dataset. These six states describe the different 'moods' of the observed days. We perform our feature analysis based on the segregation of days to each of the hidden state. Figure 10 shows the sequential appearances of various states over the complete time span.

Hyperparameter	Value
gamma	5.0
alpha	0.2
startAlpha	5.0
stickyKappa	5.0
nu	0
sF	1.0
kappa	$1e-07$

Table 3: Hyperparameters used for the best run of HDP-HMM.

States 4 and 5 are dominated by days where cult macros have been actively used with macro average of occurrence of 90.85% and 96.63% respectively. In Figure 10, we observe that almost 88% of the initial three months belong to state 5. Toward the final days in our dataset, we find that the days mostly belong to state 4. It is to be noted that the results are obtained after setting a low value of stickiness. We observe that in the last two months where the overall community activity is very high, 80% of the days belong to state 4. In spite of the overwhelming presence of cult macros in state 4, the singleton macros enjoy 13 times more likes than the cult macros. From Figure 11a, it can be noted that the share of singleton macros in state 4 amounts only to 8 %. In fact, it can also be observed from Figure 11b that singletons enjoy high value of likes in all the states wherever they are present. This is primarily becasue of the fact that about 94.15% of post images are singletons. The post images tend to garner more likes than comment images. In contrast, Cult macros never appeared as a post image.



(a) Heatmap showing the mean of count of macros across all states and all macros classes.

(b) Heatmap showing the mean of the log likes across all states and all macros classes.

Figure 11: Comparison of different states and macro classes via average values of counts and log of likes

The distinguishing criteria between the states 4 and 5 primarily arises from the mean inter arrival time between the cult macros. In the later days, due to large subscriber base there was increased activity in the community compared to the initial days resulting in a large difference in the macro averaged mean inter arrival times for the cult macros in those days. State 5 has a average inter arrival time of 3.57, while that of state 4 is -2.49 on log scale, for the cult macros. The lower the value, higher the activity in the community. This clearly demarcates the activity happening at days on state 4 and state 5, though the cult macros have an overwhelming presence in those days as compared to other categories. Moreover, figure 13 shows that state 5 encodes days having minimum entropy denoting sole presence of cult macros in those days.

Hypergiants account for 0.88% of the macro vocabulary (220 unique macros). By definition, they appear for short span of time but enjoy high popularity during that timespan. Days with considerable presence of hypergiants are mostly confined to state 1 and state 2 though on average they amount to only 3-4% of those days activity. Hypergiants macros tend to be the most active macro class by virtue of lowest mean inter arrival time for the days in state 1. Most importantly, in these days of state 1, short-lived macros enjoy 43 times more likes than their appearances in days belonging to other states. State 1 essentially contains those days where images with high likes were posted. Images in state 1 days have on average 7 times more likes than those in the next highest state, state 4. So, state 1 clearly indicates the most popular days. The trend is evident also from Figure 11b. Also, Figure 12 shows that considering activity from all macros classes, state 1 captures most active days since the mean inter arrival time considering all macros classes is lower compared to all other states. Amongst the days in the state 1, the hypergiants has the least of the lifespan with a score of 0.40 as compared to other categories of 0.49, 0.9 and 0.72.

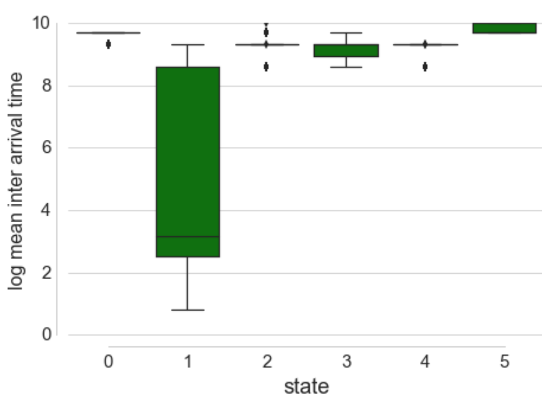


Figure 12: Boxplots denoting the distribution of combined inter arrival time over all the states.

State 0 contains days where singleton appears almost 88% times in a day. In spite of large share of singletons, the state has an average lifespan of 0.21 and is only the second lowest, after state 5 with a lifespan of 0.12. Singletons in these days garner on average 350 times more likes than classic macros, the next highest class based on likes in state 0. Compared to the other classes the singletons in state 0 show very high activity as their mean inter arrival time per day is lowest compared to the others. The remaining 2 states, states 2 and 3 are dominated by singleton macros with 59.73% and 57.37%. They also show highest values of entropy with values 1.06 and 1.01 respectively as also evident from Figure 13. Considering all days where classic macros appear, half (49%) of those days belong to state 3. Classic macros in these days enjoy 3 times more likes on average compared to other days from other states. Thus, activity and popularity of different classes of macros have indicated the possible 'mood' of a day via the obtained hidden states.

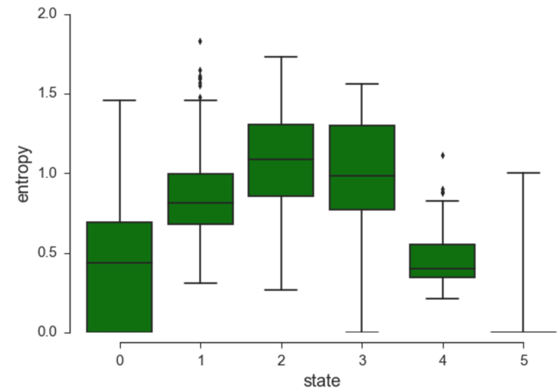


Figure 13: Boxplots denoting the distribution of entropy over all the states.

Freshness vs. *Familiarity*

Each state exhibit mixture of activities for various classes of macros which sequentially appear over the period of a day. This mixture of freshness and familiarity rightly define the current user engagement in the community. While it is known that the old macros get reused along with new macros, majority of the macros die out, not to be used again by the community. However, some persist and get reused again. Here, we look for something more specific - whether the community's appreciation for usage of familiar macros and fresh macros vary from one day to another or not.

We perform a trend analysis of average likes garnered by the fresh and the familiar macros considering only post images. The rationale behind using post macros is their independent existence in the community. By definition, any macro when used for the first time is a fresh macros, and from subsequent usages become familiar macros. We calculate the average likes garnered by posts of the said two categories with a sliding window of five and look for the maximal subsequences of monotonically increasing and decreasing likes. In this way, we find all such maximal subsequences of lengths 3, 4 and 5.

We observe 394 such subsequences where a trend is evident. There are 260 such subsequences with an increasing trend of likes for the familiar macros, 273 of such subsequences have an increasing trend for fresh macros and 139 of the subsequences show a trend for increase in both fresh and familiar macros. In particular, we find that about 121 subsequences exist, where familiar macros show an increase in the trend while fresh macros show a decrease in trend.

These trends clearly capture the mood of the community on different days. In the stretch of days where likes gained by familiar ones increase monotonically, we find that the number of appearances of classic and hypergiants macros are of the order of their usual appearances. This is probably because, the community is not conscious of this phenomena. Nevertheless, in spite of the normal appearance, the classic macros enjoy two times more likes than the usual, which overall increase the likes gained by familiar macros. Further, among the days observed with this increasing trend, we find 70.23% stretches reflect monotonic increasing trend of likes for classic macros themselves. Hypergiants show sporadic increasing trends of like which covers 40.15% of stretches where familiar macros show increasing trend of popularity.

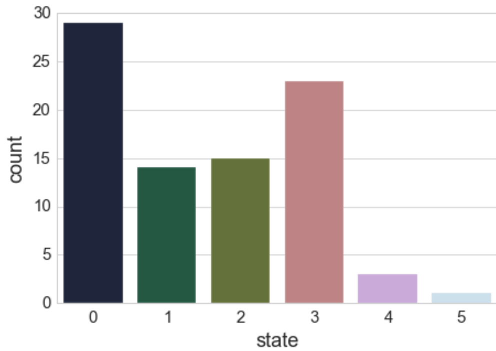


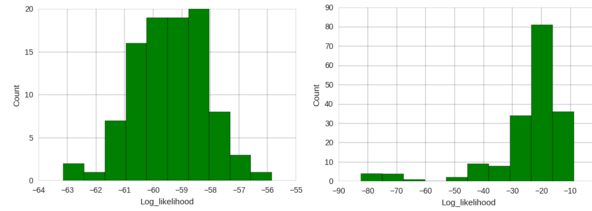
Figure 14: Barplot shows the distribution of states over the 85 days where familiar macros gained more average likes than fresh macros.

There are 85 days where familiar macros gained more average likes than the fresh macros. It is observed that, classic macros appear 1.32 times more in these days than on other days. Further, hypergiant macros appear 20.8% more than on other days to contribute to the likes gained by the familiar macros. In addition, classic macros enjoy 50 times more likes in these specific days compared to other days. Hypergiants also enjoy two times more likes than usual days. These corroborate the fact that the classic and hypergiants macros contributes primarily to the popularity of the familiar macros. Figure 14 shows the state distribution over these 85 days. The most observed state is state 0 which mostly contains days where singletons are prevalent.

In spite of the prevelance of singletons in the state which by default belongs to the fresh memes, the days had higher average likes for familiar memes. The familiar memes were essentially the ones belonging to classic macros as the other two classes are not part of the analysis. From Figure 11b,it is to be noted that hypergiant macros fail to garner any likes. Since we perform our analysis only on the post images, cult macros plays no role here as well. Hence, popularity of familiar macros in these days indicate that classic macros contribute to that popularity. Also, the second most frequent state here is state 3 which contains days where classic ones were prevalent and popular. This observation support our hypothesis that different classes of macros play an important role in determining the acceptance of *familiarity* vs. *freshness* in community across different days.

The grammar framework

We use adaptor grammar framework for motif identification from the predicted state sequence. We aim to find recurring patterns from the data, that are strong indicators of the community behavior. We see the motifs as functional subunits that characterize the community behavior and we used it to



(a) random data

(b) heldout

Figure 15: Histogram of loglikelihood for held-out and random sequences for the collocation grammar.

predict the mood of the subsequent days. We train the adaptor grammar A_g with 600 sub-sequences of length 25 each for both the unigram and the collocation grammars. The input data set is shuffled for the training. The terminals in both the cases are considered as the hidden states we observed from the HDP-HMM model, i.e., states 0, 1, 2, 3, 4 and 5. To calculate the effectiveness of the grammar, we provide two sets of data to the grammar to be parsed. We had 179 sub-sequences data each of length 25 from the HDP-HMM state sequence held out from the grammar during its learning. We also generate 96 random sequences to be given to the grammar. Since, the framework essentially learns a Probabilistic Context Free Grammar, for each sequence, the grammar outputs a parse tree along with a likelihood for the tree. If the likelihood of the parse tree is high, it implies that the string is more likely to be part of the grammar learned. The collocation grammar performed better in terms of generating likelihood for the held out sub-sequences compared to the unigram grammar. Figure 15 shows log-likelihood for both the held-out and random set of sequences where the collocation grammar have been used to parse the sub-sequences. It can be observed that, the random sequences sampled from a normal distribution show less likelihood with a mean log likelihood of -59.39. The held-out sequences from the data as shown in Figure 15b shows higher likelihoods for the sequences and is skewed towards the higher likelihoods. This experiment shows two important aspects. One, the community follows a certain trend in its behavior. Second, our model captures the trend for this behavior, using our framework of HDP-HMM followed by adaptor grammar.

State sequence prediction

We use the grammar framework, to predict the possible state for a day in future. We use the state sequence for the past n days and get the prediction for the next subsequent day. Since, we essentially learned a grammar framework, we frame our recommendation task as follows. For a single day, we obtain k sequences, where k is the number of states. Each sequence of length $n+1$, has the first n entries in the sequence same for all. The last entry is each possible state. Now, we calculate the likelihood of all the sequences and rank the most likely sequence as per the grammar.

To measure the performance of our prediction framework, we use a variation of Mean Reciprocal Rank (MRR) motivated from [4]. From the ground truth state assignment, we calculate the rank of the actual sub-sequence in our likelihood based rank. We only consider the ranking valid if the actual sub-sequence appears within rank 3 and the rest are considered *no hit* with rank= infinity. Given a grammar $A_g$and rank list R_l, if rank of a sub-sequence *seq* is k, PredictionRank is defined as:

$$\text{PredictionRank}(A_q, R_l, seq) = \begin{cases} k & \text{if } k \leq 3 \\ \infty & \text{otherwise} \end{cases} \quad (4)$$

Now MRR can be calculated as:

$$\text{MRR} = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{\text{PredictionRank}(A_q, R_l, seq_i)} \quad (5)$$

MRR for various sub-sequence lengths is reported in table 4.

Length of input sequences	MRR
3	0.56218
5	0.69262
10	**0.69845**
15	0.68077
20	0.68750
25	0.67597
30	0.67528

Table 4: MRR for the sequence prediction with different length input sequences.

CONCLUSION AND FUTURE WORK

In this work we proposed a framework for studying the temporal properties exhibited by an online community which uses *image macros* as their primary means of communication. We jointly solve the problem of clustering the days that have similar temporal properties and modeling the interactions between successive days using HDP-HMM. The feature analysis of the state space so obtained confirms the effectiveness of our model. We further find regularities from the state prediction sequences i.e. the mood sequences by motif identification using adaptor grammar. The grammar so learnt was used for prediction of the possible ‘mood’ for a day based on the moods for the preceding days. Our analysis also shows how a community makes a sustainable progress by first starting with a set of borrowed popular ideas and subsequently modifying the content specific to the local context that cater to the specific information requirements for the users of a demographic.

Our prediction framework will be beneficial to the community administrators to strengthen the user engagement within the community. By strategically placing different classes of *image macros* based on the ‘mood’ of the day, the community can occupy larger attention share of its users. This is particularly relevant at a time, where metrics like ‘time biased gain’ [6] centered around the concept of “time well spent” [28] is gaining popularity in online systems.

On a wider perspective, we are intrigued by the dynamics that is evident in the image meme based interactions and its resemblance with text based interactions at a community level. We already find that the frequency of usage of macros follow zipf-mandelbrot law. We also find that the community borrows macros from the global context, as in the case of natural languages. We were able to observe regualrities in the state sequence, i.e. the sequence do not appear randomly but based on some inherent rules, for which a semantic interpretation is yet to be formed [26]. Compounding is a common phenomena observed in text memes and hashtags [23] in social media. Empirically, we find instances of image macros that merge together to form new macros, which is promising to explore. We expect that a deeper analysis of image based memes and the community will be an interesting addendum to the findings mostly performed on text based communication platforms.

REFERENCES

1. Lada A Adamic, Thomas M Lento, Eytan Adar, and Pauline C Ng. 2016. Information evolution in social networks. In *Proc ACM Int. Conf. on Web Search and Data Mining*.
2. Eytan Adar, Daniel S Weld, Brian N Bershad, and Steven S Gribble. 2007. Why we search: visualizing and predicting user behavior. In *Proc. ACM Int. Conf. on World Wide Web*.
3. Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. 2013. A peek into the future: predicting the evolution of popularity in user generated content. In *Proc. ACM Int. Conf. on Web search and Data Mining*.
4. Ziv Bar-Yossef and Naama Kraus. 2011. Context-sensitive query auto-completion. In *Proc. ACM Int. Conf. on World Wide Web*.
5. Matthew Beal and Praveen Krishnamurthy. 2012. Gene expression time course clustering with countably infinite hidden Markov models. *arXiv Preprint arXiv:1206.6824* (2012).
6. Charles LA Clarke and Mark D Smucker. 2014. Time well spent. In *Proceedings of the 5th Information Interaction in Context Symposium*. ACM, 205–214.
7. Michele Coscia. 2014. Average is boring: How similarity kills a meme’s success. *Scientific reports* (2014).
8. Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proc. ACM Int. Conf. on World Wide Web*.
9. Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proc. ACM Int. Conf. on World Wide Web*.
10. Richard Dawkins and others. 2016. *The selfish gene*. Oxford university press.
11. Stefan Evert. 2004. A simple LNRE model for random character sequences. In *Proceedings of JADT*.
12. Emily B Fox. 2009. *Bayesian nonparametric learning of complex dynamical phenomena*. Ph.D. Dissertation. Massachusetts Institute of Technology.

13. Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proc. ACL Annual Meeting of the Association for Computational Linguistics*.

14. Matthew Hoffman, Perry Cook, and David Blei. 2008. Data-driven recomposition using the hierarchical Dirichlet process hidden Markov model. In *Proc. Int. Computer Music Conference*.

15. Michael C Hughes, William T Stephenson, and Erik Sudderth. 2015. Scalable adaptation of state complexity for nonparametric hidden Markov models. In *Advances in Neural Information Processing Systems*. 1198–1206.

16. Mark Johnson. 2010. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *Proc. ACL Annual Meeting of the Association for Computational Linguistics*.

17. Mark Johnson, Thomas L Griffiths, Sharon Goldwater, and others. 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in neural information processing systems* 19 (2007), 641.

18. Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proc. ACM Int. Conf. on Web Search and Data Mining*. ACM.

19. E. V. Khmaladze. 1988. The statistical analysis of a large number of rare events. In *CWI Report MS-R8804*.

20. Farshad Kooti, Winter A Mason, Krishna P Gummadi, and Meeyoung Cha. 2012. Predicting emerging social conventions in online social networks. In *Proc. ACM Int. Conf. on Information and Knowledge Management*.

21. Amrith Krishna, Pavankumar Satuluri, Shubham Sharma, Apurv Kumar, and Pawan Goyal. 2016. Compound Type Identification in Sanskrit: What Roles do the Corpus and Grammar Play?. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*. The COLING 2016 Organizing Committee, Osaka, Japan, 1–10. http://aclweb.org/anthology/W16-3701

22. Daniel Kuettel, Michael D Breitenstein, Luc Van Gool, and Vittorio Ferrari. 2010. What's going on? Discovering spatio-temporal dependencies in dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*.

23. Suman Kalyan Maity, Ritvik Saraf, and Animesh Mukherjee. 2016. # Bieber+# Blast=# BieberBlast: Early Prediction of Popular Hashtag Compounds. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 50–63.

24. John K Pate and Mark Johnson. 2016. Grammar induction from (lots of) words alone. In *Proc. COLING 2016, Int. Conf. on Computational Linguistics*.

25. Lawrence Rabiner and B Juang. 1986. An introduction to hidden Markov models. *ieee assp magazine* 3, 1 (1986), 4–16.

26. Rajesh PN Rao, Nisha Yadav, Mayank N Vahia, Hrishikesh Joglekar, R Adhikari, and Iravatham Mahadevan. 2009. A Markov model of the Indus script. *Proceedings of the National Academy of Sciences* 106, 33 (2009), 13685–13690.

27. Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica sinica* (1994), 639–650.

28. Mark D Smucker and Charles LA Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 95–104.

29. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet Processes. *J. Amer. Statist. Assoc.* (2006).

30. Albin Wagener. 2014. Creating identity and building bridges between cultures: The case of 9gag. *Int. Journal of Communication* (2014).

31. Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proc. ACM Int. Conf. on Web Search and Data Mining*.

32. Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. 2014. Finding progression stages in time-evolving event sequences. In *Proc. ACM Int. Conf. on World Wide Web*.

33. Ke Zhai, Jordan Boyd-Graber, and Shay B Cohen. 2014. Online adaptor grammars with hybrid inference. *Transactions of the Association for Computational Linguistics* 2 (2014), 465–476.

34. Ke Zhai, Zornitsa Kozareva, Yuening Hu, Qi Li, and Weiwei Guo. 2016. Query to Knowledge: Unsupervised Entity Extraction from Shopping Queries using Adaptor Grammars. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*.