Modeling Morphology of Cascades in Online Social Networks using Multi-Order Markov Chains

M. Zubair Shafiq Alex X. Liu Hayder Radha
Michigan State University, East Lansing, MI, USA
{shafiqmu,alexliu}@cse.msu.edu, radha@egr.msu.edu

ABSTRACT

Cascades represent an important phenomenon across various disciplines such as sociology, economy, marketing, and epidemiology. An important property of cascades is their morphology, which encompasses their structure, shape, and size. However, we believe that cascade morphology has not been thoroughly studied in prior literature. The goal of this paper is to develop a model that allows us to quantitatively and rigorously analyze, characterize, and classify cascade morphology. Towards this end, we propose $M4$, a Multi-order Markov Model for the Morphology of cascades in online social networks. $M4$ can represent cascades with arbitrary structures, shapes, and sizes, and also allows us to classify cascades based on their underlying attributes. $M4$ essentially provides a quantitative tool for analyzing and classifying cascades solely based on their morphology. For validation, we apply $M4$ model to solve the following cascade size prediction problem: given the first τ_1 edges in a cascade, can we predict if the cascade will have a total of more than τ_2 edges over its lifetime, where $\tau_2 > \tau_1$? The results of our experiments, conducted on a Twitter dataset, show that $M4$ achieves classification accuracy of up to 91.2% for this prediction problem.

1. INTRODUCTION

The term *cascade* describes the phenomenon of something propagating along the links in a social network. That something can be information such as a URL, action such as donating to a campaign against cancer, influence such as buying a product, discussion such as commenting on a blog article, and a resource such as a torrent file. Cascade phenomenon has been a fundamental topic in many disciplines such as sociology, economy, psychology, political science, marketing, and epidemiology with research literature tracing back to the 1950s [5].

The focus of this paper is to study the morphology of cascades in online social networks. Cascade morphology encompasses many aspects of cascades such as their structures, shapes, and sizes. Specifically, we aim to develop a model that allows us to quantitatively and rigorously analyze, characterize, and classify cascade morphology; which are extremely difficult without a model. Despite the numerous publications regarding different aspects of online social networks, there has been relatively small amount of effort addressing the morphology of cascades in these networks. Recently some researchers have studied the structure of cascades [3, 4]; however, their analysis of cascade structures is limited to basic structural properties such as degree distribution, size, and depth.

In this paper, we propose $M4$, a multi-order Markov chain based model for formally representing the morphology of cascades in online social networks. $M4$ has two key components: a cascade encoding algorithm and a cascade modeling method. The cascade encoding algorithm uniquely encodes the morphology of a cascade for quantitative representation. It encodes a cascade by first performing a depth-first traversal on the cascade graph and then compressing the traversal results using run-length encoding. The cascade modeling method models the run-length encoded sequence of a cascade as a discrete random process. This random process is further modeled as a Markov chain, which is then generalized into a multi-order Markov chain model. Note that $M4$ can represent cascades with arbitrary structures, shapes, and sizes. It can also characterize cascades with different attributes using the state information from the underlying multi-order Markov chains.

In this paper, we utilize $M4$ to solve the cascade size prediction problem. This prediction problem not only serves the purpose of validating the relevance of our model but also can be used in many real-life applications. We use a dataset collected from Twitter containing more than 8 million tweets involving more than 200 thousand unique users to evaluate the accuracy of our model. Our experimental results show that $M4$ indeed can capture the characteristics of cascade morphology. Using the parameters derived from $M4$, we achieve classification accuracy of 91.2% for the cascade size prediction problem.

2. PROPOSED MODEL

$M4$ model consists of two major components. The first component encodes a given cascade graph so that its morphological information is retained. The second component models the encoded sequence using a multi-order discrete Markov chain.

The first step in cascade encoding is to encode the constructed cascade graph that uniquely represents the structure of the cascade graph. Towards this end, we use a modified version of the graph encoding algorithm in [2]. Our

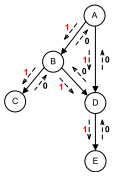

Figure 1: Traversal example

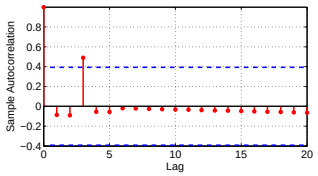

Figure 2: Autocorrelation analysis

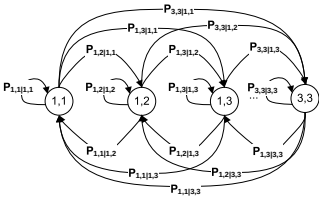

Figure 3: Markov model

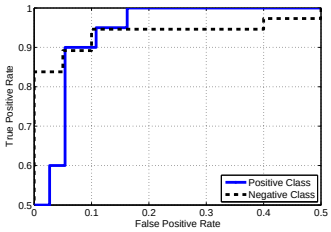

Figure 4: ROC plot

algorithm first conducts a depth-first traversal of the constructed cascade graph starting from the root node, which results in a spanning tree. To result in a unique spanning tree, at each node in the cascade graph, we sort the outgoing edges in the increasing order of their time stamps, *i.e.*, sort the outgoing edges $e_1, e_2, \cdots, e_k$ of a node so that $T(e_1) < T(e_2) < \cdots < T(e_k)$; and then traverse them in this order. For each edge, we use 1 to encode its downward traversal and 0 to encode its upward traversal. Figure 1 shows an example of the traversal process.

The second step in cascade encoding is to convert the binary sequence into the corresponding run-length encoding. By replacing each run in a binary sequence with the length of the run, we obtain the run-length encoding of the binary sequence. For example, for the binary sequence 11011000, the corresponding run-length encoding is 2123. Consider the run-length encoded sequence $\hat{C}$ of a cascade graph G. We can jointly model this sequence using a discrete random process $\{\hat{C}_k\}$, $k = 1, 2, ..., |\hat{C}|$. Basic analysis of this process would reveal that there is some level of dependencies among the consecutive symbols emitted by the random process. Meanwhile, to balance between capturing some of the dependencies within the process and to simplify the mathematical treatment of this encoded sequence, we resort to invoking the Markov property assumption [1]. This assumption can be reasonably justified (at least to some extent) by analyzing the autocorrelation function of the underlying process $\{\hat{C}_k\}$. For a first order Markov process, this implies the following assumption: $Pr[\hat{C}_n = c_n|\hat{C}_1 = c_1, \hat{C}_2 = c_2, ..., \hat{C}_{n-1} = c_{n-1}] = Pr[\hat{C}_n = c_n|\hat{C}_{n-1} = c_{n-1}]$. Given the Markov assumption with homogeneous time-invariant transition probabilities, $\hat{C}$ can be represented using a traditional Markov chain. We can generalize a Markov chain model by incorporating multiple consecutive transitions as a single state in the state transition matrix, which will allow us to specify arbitrary sized subgraphs of cascades. The order of a Markov chain represents the extent to which past states determine the present state. We use the autocorrelation-based analysis to select the order of Markov chain. Figure 2 shows the autocorrelation function of an example sequence, where we select the order of Markov chain to be 3, and Figure 3 shows its corresponding multi-order Markov chain model.

We now use the Markov states of our model to serve as features of a given cascade for determining its class label. Let us assume that the presence of a state in our multi-order Markov chain model is represented by a binary random variable $X_i, i = 1, 2, ..., n$, where n is the total number of states captured by the model. Given that our feature set is the same as the states of the Markov chain model, we can think of the X_is as the variables for our features. Thus, our training process proceeds as follows. For a given class Y of cascades, we evaluate the presence of a given feature (state) X_i in Y by analyzing a sufficiently large number of sample cascades that belong to the class Y. Subsequently, we are able to evaluate the a-priori conditional probability $P(X_i|Y)$ for each class $Y \in \{1, 2, ..., k\}$, where the number of classes k is usually very small. In our case, we are interested in the traditional binary classifier with $k = 2$ as discussed in what follows. However, note that this classification methodology can be extended to the cases with $k > 2$ using the well-known one-against-one (pairwise) or multiple one-against-all formulations. Given the features, we can classify cascades by deploying a machine learning classifier. In this study, we use a Bayesian classifier to jointly utilize the selected features to classify cascades.

3. EXPERIMENTAL RESULTS

We use Twitter dataset to investigate the following cascade size prediction problem using $M4$ model. This prediction problem not only serves the purpose of validating the relevance of our model but also can be used in many real-life applications. To our best knowledge, no prior work has attempted to solve exactly the same problem; however, there are several relevant attempts reported in literature [3, 4].

Using the methods described earlier, we select the appropriate order of our Markov chain model to be 8 and then use the states as features. We feed these features jointly to the Naïve Bayes classifier for automated training and testing. Using the standard 10-fold cross validation procedure, we now present the classification results. Figure 4 shows the Receiver Operating Characteristic (ROC) curves for predicting cascade sizes. We note that ROC curves approach the top-left point corresponding to 100% true positive rate and 0% false alarm rate. The average Area Under Curve (AUC) for curves in Figure 4 is 0.968. At the optimum ROC point, the accuracy is 91.2%, the average true positive rate is 94.6%, and the average false positive rate is 5.4%. The high accuracy implies that our $M4$ model indeed can capture the characteristics of cascade morphology. We have also evaluated $M4$ for varying values of τ_1 and τ_2 (detailed results are omitted due to space constraints). As expected, we observed that smaller values of τ_1 and $\tau_2 - \tau_1$ decrease the classification accuracy of $M4$.

4. REFERENCES

[1] P. Bremaud. *Markov Chains*. Springer, 2008.

[2] R. C.-N. Chuang, *et al.* Compact encodings of planar graphs via canonical orderings and multiple parentheses. *Automata, Languages and Programming*, 1443:118–129, 1998.

[3] W. Galuba *et al.* Outtweeting the twitterers - predicting information cascades in microblogs. In *Workshop on Online Social Networks*, 2010.

[4] J. Leskovec *et al.* Cascading behavior in large blog graphs. In *SIAM International Conference on Data Mining*, 2007.

[5] E. M. Rogers. *Diffusion of Innovations*. Cambridge University Press, 2003.