5-2013

Retweeting: An act of viral users, susceptible users, or viral topics?

Tuan-Anh HOANG
Singapore Management University, tahoang.2011@smu.edu.sg

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1137/1.9781611972832.63

Citation

HOANG, Tuan-Anh and LIM, Ee Peng. Retweeting: An act of viral users, susceptible users, or viral topics?. (2013). *Proceedings of the 2013 SIAM International Conference on Data Mining: May 2-4, Austin, Texas*. 569-577. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/1965

Retweeting: An Act of Viral Users, Susceptible Users, or Viral Topics?

Tuan-Anh Hoang* Ee-Peng Lim*

Abstract

When a user retweets, there are three behavioral factors that cause the actions. They are the *topic virality*, *user virality* and *user susceptibility.* Topic virality captures the degree to which a topic attracts retweets by users. For each topic, user virality and susceptibility refer to the likelihood that a user attracts retweets and performs retweeting respectively. To model a set of observed retweet data as a result of these three topic specific factors, we first represent the retweets as a three-dimensional tensor of the tweet authors, their followers, and the tweets themselves. We then propose the $V2S$ model, a tensor factorization model, to *simultaneously* derive the three sets of behavioral factors. Our experiments on a real Twitter data set show that the $V2S$ model can effectively mine the behavioral factors of users and tweet topics during an election event. We also demonstrate that the $V2S$ model outperforms the other topic based models in retweet prediction.

1 Introduction

Several empirical research works have been conducted on the information diffusion phenomena in Twitter [10, 14, 21, 24]. Most of them studied diffusion at the macro level which involves aggregated diffusion patterns, e.g., persistence and shape of the adoption curve of information items. There is little research on modeling the act of retweeting which contributes to diffusion, and explaining the retweet actions using the underlying user behavioral factors and topic factors.

In this work, we attribute a user's retweet actions to three orthogonal factors, namely **topic virality**, **user virality** and **user susceptibility**. When a user retweets a message, one may attribute this to the viral topical content of the message. Different topics may demonstrate different degree of virality. Past studies have shown that death of celebrities (e.g., Michael Jackson, Steve Jobs) and political uprisings are topics that are viral. There are also many other non-viral topics. Other than topic virality, a user may retweet a message due to the user from whom she receives the message. This user, also known as the sender, may be good at generating content that attracts readership. We say that such users are viral and they are viral for some topics but not for others. Finally, a user may also retweet due to her own susceptibility, which again is topic specific.

Our research objective is therefore to develop a model for retweets based upon the above three topic-specific factors. Through the model, we would like to explain retweet actions by the appropriate factors, and to measure the underlying topic and user factors which can be used in several interesting applications including social media marketing, content recommendation, event monitoring and detection.

To the best of our knowledge, modeling retweet actions using behavioral factors is a novel research problem. In [6], Hoang and Lim introduced the concepts of message virality, user virality and susceptibility factors in viral diffusion, and proposed different ways to measure these factors considering their inter-dependencies. The above three factors are however not topic specific. In this paper, we aim to develop a new model that involves *topic-specific user virality*, *topic-specific user susceptibility*, and *topic virality.* Defined at the topic level, these factors can be more easily used to predict future retweet actions. There are nevertheless some technical challenges in this work.

- To model retweets using the three factors, we need to know if the followers are actually aware of any tweet generated from a user. A follower is clearly aware of the tweet when she retweets, but we cannot say so when she does not retweet as the browsing of incoming tweets cannot be observed.

- Each retweet action is jointly contributed by all these topic-specific factors. How to separate the effects of each factor from the other two and to measure them are therefore challenges. This scenario is analogous to the computation of hubs and authorities from a set of links between web pages, except that we now have to consider three (not two) factors simultaneously.

This paper addresses the first challenge by inferring the window size of tweets read by users when they retweet. We then address the second challenge by constructing a *retweet tensor* representing users retweeting messages posted by their followees. We then develop a

[*] Living Analytics Research Centre, Singapore Management University

factorization based model on this tensor to simultaneously measure the three topic-specific factors. Our main contributions in this work consist of the following.

- We propose a tensor factorization model which captures the relationships between retweet actions and the above three factors. The model, known as $\boldsymbol{V2S}$, represents retweets as a three-dimensional tensor which is factorized into topic virality vector, user-topic virality matrix, and user-topic susceptibility matrix simultaneously.
- We convert the above constrained factorization problem into a unconstrained optimization which can be solved using gradient descent methods.
- We apply the $V2S$ model to retweet prediction in a real Twitter dataset and show that the $V2S$ model outperforms other topic-based baseline models.
- An empirical analysis of the topic virality, user virality and user susceptibility factors for the same Twitter dataset has been conducted to demonstrate the efficacy of the $V2S$ model.

The rest of the paper is organized as follows. We cover the related works in Section 2. Section 3 provides justifications that behavioral factors should be modeled at the topic level. We define our $V2S$ model and solve the model in Section 4. We present our retweet prediction experiments in Section 5. The empirical analysis of topic and user factors is given in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

2.1 Item Virality Most of the previous works only measure item virality as the only factor related to diffusion [3, 8, 11, 14, 15, 17]. An item's virality has been simply measured by its *popularity*, i.e., the number of users adopting the item. In [8], another item virality measure called *viral coefficient* is introduced. Viral coefficient is defined by the average number of new adopters generated by each existing adopter. For Twitter data, the popularity of a tweet is therefore measured by the number of its retweets, and the viral coefficient is the same as retweet count per user. The two item virality measures are very simple and does not consider other user factors and topic level factors.

2.2 Retweet Modeling There have been some empirical research on understanding the correlation between retweet likelihood and network metrics (e.g., number of followers and followees), as well as content characteristics (e.g., the presence of URLs and hashtags) [16, 18, 20]. However, to the best of our knowledge, there are only few works on modeling retweet actions. Zi Yang *et. al.* proposed a factor graph based method using the retweet traces as input data [22]. However, as Twitter now does not include screen names of intermediate users in retweet content, the trace of second or subsequent hop of retweets cannot be fully observed by a user who receive the retweet. Hence, this method is no longer appropriate. Luo *et. al.* proposed a log-linear model to explain retweets by social tie based features [23], which are independent from user behaviors and topics. Recently Peng *et. al.* [12] and Chen *et. al.* [4] proposed methods that predict retweets based on author's profile, content elements, network and temporal characteristics. These features require a large dataset covering user activities over a long time period. In contrast, our model only requires the retweet data and considers new topic and user factors.

3 Topic and Retweet

In this section, we demonstrate our research motivation. By performing topic analysis on a real dataset collected from Twitter, we show that user behavioral factors and item factor contributing to retweet actions should be studied at topic level.

3.1 Dataset We first select a set of 58 Singapore-based seed users which includes accounts of political parties (e.g., *PAPSingapore* and *wpsg*), politicians (e.g., *georgeyeo*), political commentators (e.g, *temasekreview*), and bloggers (e.g., *mrbrown*). We then derived the followers and followees of the seed users. This allows us to create a larger set of 32,138 users who declare themselves to be located in Singapore. We then crawled tweets published by the set of users on a daily basis using Twitter REST API[1]. We collected 2,091,906 tweets published between April and August 2011 for this study. We apply the following steps for data preprocessing.

Constructing the user network. As Twitter REST API does not provide the creation time of follow links, we have to infer the follow links and their timestamp based on the mention as suggested in [14]. That is, we create for users u and v a follow link from u to v at the time when u mentions "@v" the kth time in u's tweets. In our experiments, we set $k = 3$.

User and tweet selection. We first remove users who neither retweet nor get retweeted. The tweets posted by these users are also removed as they do not have any effect on other users's retweets and getting retweeted. Furthermore, in Twitter, topics of tweet content change rapidly over time [7, 10], and so do the user behaviors. We therefore only use a subset of data collected within a short duration of time to

[1] Twitter API: http://dev.twitter.com/doc

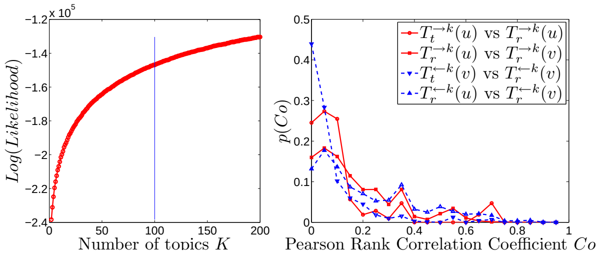
Table 1: Statistics of One-Week Dataset

#Users	32,138
#Tweets	86,729
#Original tweets	29,838
#Retweets	56,891
#Retweets from original tweets	26,202

study user behavior in that time. For this study, we use tweets published within one week up to Singapore General Election Day (May 8th, 2011). Since the election is a socially interesting event, we expect that tweets generated by the event to be well read, and highly retweeted by the election voters. The statistics of the final dataset are given in Table 1. The statistics suggest that the selected original tweets are highly retweeted as we have more retweets than tweets, and the number of retweets from original tweets is approximately equals to the number of original tweets.

Determining user awareness of receiving tweets. In Twitter, the latest tweets posted by a user's followees always appear at the top of her timeline. Hence, when a user has many followees and receives many tweets, some earlier received tweets may not be read by the user as they are hidden by newer incoming tweets. In that case, the earlier tweets may have been missed by the user and never been retweeted. As Twitter REST API does not provide user click-through information that reveals the tweets that have been seen by users, we define a time window in which the received tweets will be read. This time window size is determined as follows. We know that every retweet by a user u comes with a corresponding tweet m that u must have read. We first count the number of other tweets u received within the time window from the time u received m to the time u retweeted m. Based on this count we estimate N_r^w the number of tweets a user may read on her timeline whenever she performs a retweet. We found that N_r^w follows a long tail distribution. For more than 90% of the times, N_r^w is not larger than 50. We therefore determine that a user u received and actually read through the tweet m if and only if m is among last 50 tweets posted by u's followees up to the time u makes a retweet. Otherwise, the tweet m is considered not read by the user u.

Topic discovery. We applied LDA model [2] to automatically identify the topics of every original tweet. We first remove all retweets and non-informative tweets, e.g., tweets generated by third party applications like FourSqure[2] or Instagram[3]. We then remove from remaining tweets all stop words, slang words[4], and non-English phrases. Finally, we use the LDA implementation from Stanford's Topic Modeling Toolbox[5] to discover topics.



(a) Likelihood of the LDA model (b) Distribution of Pearson rank correlation between relative popularities of topics

Figure 1: Topic discovery and analysis

Figure 1(a) shows the likelihood of the LDA model with respect to the number of topics K. As expected, larger K gives larger likelihood. The amount of improvement decreases as K increases. In consideration of time and space computational complexity, we set the number of latent topics to 100. Moreover, as each tweet is a short document, we are not interested in tweets that cover many topics. We therefore only consider tweets having some dominating topics. To do this, we filter out tweets having sum of K_{dom} highest topic probabilities (given the tweet) less than 0.95. Then, for each of the remaining tweets, we *normalize* topic distribution of the tweet such that sum of K_{dom} highest topic probabilities equals to 1, and all other topics have probability 0. In this study, we set $K_{dom} = 3$.

3.2 Observations

We now show observations that how different topics get retweeted. In particular, we aim to answer the following questions: (a) *Topic and retweet at network level*: Do all topics get equally retweeted? (b) *Topic and retweet at author side*: Does a user get same amount of retweets for every topic? and (c) *Topic and retweet at receiver side*: Does a user performs same amount of retweets for every topic?

Topic and retweet at network level. In order to compare the likelihood of being retweeted across topics, for each topic k, we derive the relative popularities of k among the set of all original tweets and the bag of retweets. The former is called *global popularity* of the topic k, denoted by $T_t^{\rightarrow}(k)$, and the later is called *retweet popularity*, denoted by $T_r^{\rightarrow}(k)$. The two popularities are defined as follows.

$$T_t^{\rightarrow}(k) = \frac{1}{|M|} \sum_{m \in M} T_k(m) \tag{3.1}$$

$$T_r^{\rightarrow}(k) = \frac{1}{\sum_{m \in R} |RT(m)|} \sum_{m \in M_R} |RT(m)| \cdot T_k(m) \tag{3.2}$$

where M is the set of all original tweets which excluded the retweets, M_R is the set of original tweets that have been retweeted, $T_k(m)$ is the k-th component of the topic distribution vector of the tweet $m \in M$, and $RT(m)$ is the bag-of-retweets of m. The two

[2] https://foursquare.com/

[3] http://instagram.com/

[4] http://en.wikipedia.org/wiki/Slang

[5] http://nlp.stanford.edu/software/tmt/tmt-0.4/

popularities has Pearson rank correlation coefficient 0.536. The low coefficient clearly shows that the relative popularity of a topic in the bag of retweets does not strongly correlate the topic's global popularity. This implies that different topics have different likelihood of being retweeted.

Topic and retweet at author side. To compare the likelihood of user u getting retweeted for different topics, we compare the relative popularities of each topic k in the set of tweets that u posted, and in the bag-of-retweets that u got. The former popularity is called *sender-specific popularity* of u for topic k, while the latter one is called *sender-specific retweet popularity* of u for topic k. The two popularities are denoted by $T_t^{\rightarrow k}(u)$ and $T_r^{\rightarrow k}(u)$ respectively, and are defined below.

$$T_t^{\rightarrow k}(u) = \frac{1}{|M^{\rightarrow}(u)|} \sum_{m \in M^{\rightarrow}(u)} T_k(m) \tag{3.3}$$

$$T_r^{\rightarrow k}(u) = \frac{1}{|R^{\rightarrow}(u)|} \sum_{m \in M^{\rightarrow}(u)} [|RT(m)| \cdot T_k(m)] \tag{3.4}$$

where $M^{\rightarrow}(u)$ and $R^{\rightarrow}(u)$ are the set of tweets posted by u, and the bag-of-retweets u got respectively.

We compute Pearson rank correlation coefficient between $T_t^{\rightarrow k}(u)$ and $T_r^{\rightarrow k}(u)$ for each user u, and between $T_r^{\rightarrow k}(u)$ and $T_r^{\rightarrow k}(v)$ for each pair of different users u and v. The distributions of the coefficients are shown in Figure 1(b). The figure clearly shows that, for each user, the relative popularities of topics in her bag-of-retweets are different from that popularities in her tweets, and are also different from the popularities in the bag-of-retweets of other users. This implies that **(1)** the same user has different likelihoods of getting retweeted for different topics, and **(2)** the same topic has different likelihoods of being retweeted when the topic is mentioned in the tweets generated by different users.

Topic and retweet at receiver side. To compare the likelihood of retweeting by user v for different topics, we compute the relative popularities of each topic k in the set of tweets v received and read, and in the set of tweets v retweeted. The former popularity is called *receiver-specific popularity* of user v for the topic k, and the latter is called *receiver-specific retweet popularity* of user v for topic k. The two popularities are denoted by $T_t^{\leftarrow k}(v)$ and $T_r^{\leftarrow k}(v)$ respectively, and are defined below.

$$T_t^{\leftarrow k}(v) = \frac{1}{|M^{\leftarrow}(v)|} \sum_{m \in M^{\leftarrow}(v)} T_k(m) \tag{3.5}$$

$$T_r^{\leftarrow k}(v) = \frac{1}{|R^{\leftarrow}(v)|} \sum_{m \in R^{\leftarrow}(u)} T_k(m) \tag{3.6}$$

where $M^{\leftarrow}(v)$ and $R^{\leftarrow}(v)$ is the set of tweets v has received and read, and the set of tweets v has retweeted respectively.

Similar to the case of author side, we compute Pearson rank coefficient between $T_t^{\leftarrow k}(v)$ and $T_r^{\leftarrow k}(v)$ for each user v, and between $T_r^{\leftarrow k}(u)$ and $T_r^{\leftarrow k}(v)$ for each pair of different users u and v. The distributions of the coefficients are shown in Figure 1(b). Again, the figure clearly shows that, for each user, the relative popularities of topics in the set of tweets she retweeted are different from that popularities in the set of tweets she received and read, and are also different from the popularities in the set of tweets that other users retweeted. This implies that **(1)** the same user has different likelihoods of performing retweet for different topics, and **(2)** the same topic has different likelihoods of being retweeted when the topic is mentioned in tweets received by different users.

4 Topic-specific Virality and Susceptibility Modeling

Motivated by the observations in Section 3, we now propose a tensor factorization based model to incorporate all user behavioral factors and content factors contributing to retweet actions at topic level.

4.1 Behavioral factors

We consider each tweet is an information item, and each retweet action is an instance of information diffusion caused by following three factors.

- **Topic virality**: This refers to the ability of a topic to attract retweets. Each topic k is associated to a virality score $v_T^k \in [0, 1]$ indicating how viral the topic is, i.e. how likely a tweet about the topic gets retweeted. We use $\mathcal{V}_T$ to denote the vector $(v_T^1, \cdots, v_T^K)$ of virality score of all K topics.
- **Topic-specific user virality**: This refers to the ability of a user to get retweeted for a specific topic. Each user u is associated to a topic-specific user virality vector $\mathcal{V}_U(u) = (v_U^1(u), \cdots, v_U^K(u))$ where $v_U^k(u) \in [0, 1]$ for $\forall k = 1, \cdots, K$, $v_U^k(u)$ indicates how viral user u is for topic k, i.e., how likely u gets retweeted for her tweets about the topic k.
- **Topic-specific user susceptibility**: This refers to the ability of a user to retweet for a specific topic. Each user v is associated to a topic-specific user susceptibility vector $\mathcal{S}(v) = (s^1(v), \cdots, s^K(v))$ where $s^k(v) \in [0, 1]$ for $\forall k = 1, \cdots, K$, $s^k(v)$ indicates how susceptible user v is to topic k, i.e., how likely v retweets a tweet about the topic k after reading the tweet.

Note that not all users have chances to tweet about a given particular topic and then get their tweets retweeted, or to read tweets about the topic that are diffused to them from their followees. We therefore may not be able to measure virality and susceptibility for every user for every topic due to the lack of historical observations. Instead, we identify, for each topic k, the subset V_k which includes all users tweeting about the topic (with regards to the topic discovery step in Section 3), and the subset S_k which includes all users receving and reading tweets about the topic. We then measure virality and susceptibility to topic k for users in V_k and in S_k respectively.

4.2 The $V2S$ Model We use a tuple (u, v, m) to denote a retweet observation that m is a tweet posted by user u, and received and read by v. We use a R_{uvm} to denote whether v retweets m ($R_{uvm} = 1$) or not ($R_{uvm} = 0$). A retweet observation is *positive* or *negative* when $R(u, v, m) = 1$ and 0 respectively. Our proposed model measures the likelihood of R_{uvm} based on topic-specific virality of u, topic-specific susceptibility of v, overall topic virality, and the topics of m as follows.

We assume that the likelihood that v retweets m is determined by **(a)** how m's topic distribution $T(m) = (T_1(m), \cdots, T_K(m))$ correlates with u's topic-specific user virality $\mathcal{V}_U(u)$; **(b)** how $T(m)$ correlates with topic virality $\mathcal{V}_T$; and **(c)** how $T(m)$ correlates with v's topic-specific user susceptibility $\mathcal{S}(v)$. Under this assumption, we therefore may estimate R_{uvm} using the dot product of $T(m)$, $\mathcal{V}_U(u)$, $\mathcal{V}_T$, and $\mathcal{S}(v)$. That is,

$$(4.7) \qquad R_{uvm} \approx \sum_{k=1}^{K} [T_k(m) \cdot v_U^k(u) \cdot v_T^k \cdot s^k(v)]$$

Given the approximation in Equation 4.7, topic-specific user virality and susceptibility, and topic virality can be learnt by solving the following regularized tensor factorization problem.

$$(4.8) \qquad (\mathcal{V}_T^*, \mathcal{V}_U^*, \mathcal{S}^*) = \underset{\mathcal{V}_T, \mathcal{V}_U, \mathcal{S}}{arg.min} \mathcal{L}(\mathcal{V}_T, \mathcal{V}_U, \mathcal{S})$$

subject to

$$(4.9) \qquad \begin{cases} v_T^k \in [0,1] & \forall k = 1, \cdots, K \\ v_U^k(u) \in [0,1] & \forall u \in V_k, \forall k = 1, \cdots, K \\ s^k(v) \in [0,1] & \forall v \in S_k, \forall k = 1, \cdots, K \end{cases}$$

where $\mathcal{L}$ is the regularized sum-of-squares error function which is defined as follows.

$$\begin{aligned} (4.10) \quad \mathcal{L}(\mathcal{V}_T, \mathcal{V}_U, \mathcal{S}) \quad &= \\ = \sum_{(u,v,m) \in \mathcal{K}} &\Big[R_{uvm} - \sum_{k=1}^{K} (T_k(m) \cdot v_U^k(u) \cdot v_T^k \cdot s^k(v)) \Big]^2 \\ &+ \alpha_t \cdot ||\mathcal{V}_T - T_r^{\rightarrow} \cdot \sum_{k=1}^{K} v_T^k||^2 + \beta_t \cdot ||\mathcal{V}_T||^2 + \\ &+ \alpha_u \cdot \Big[\textstyle\sum_{u \in V} ||\mathcal{V}_U(u) - T_r^{\rightarrow}(u) \cdot \displaystyle\sum_{k=1}^{K} v_U^k(u)||^2 + \\ &+ \textstyle\sum_{v \in S} ||\mathcal{S}(v) - T_r^{\leftarrow}(v) \cdot \displaystyle\sum_{k=1}^{K} s^k(v)||^2 \Big] + \\ &+ \beta_u \cdot \Big[\textstyle\sum_{u \in V} ||\mathcal{V}_U(u)||^2 + \sum_{v \in S} ||\mathcal{S}(v)||^2 \Big] \end{aligned}$$

where $\mathcal{K}$ is the set of all retweet observations, $V = \bigcup_k V_k$, $S = \bigcup_k S_k$, $T_r^{\rightarrow} = (T_r^{\rightarrow 1}, \cdots, T_r^{\rightarrow K}))$, $T_r^{\rightarrow}(u) = (T_r^{\rightarrow 1}(u), \cdots, T_r^{\rightarrow K}(u))$, and $T_r^{\leftarrow}(v) = (T_r^{\leftarrow 1}(v), \cdots, T_r^{\leftarrow K}(v))$

In Equation 4.10, the term $||\mathcal{V}_U(u) - T_r^{\rightarrow}(u) \cdot \sum_{k=1}^{K} v_U^k(u)||^2$ is the distance between $\mathcal{V}_U(u)$ and $T_r^{\rightarrow}(u)$ after weighting the latter by sum of all components of the former. This term ensures that $V_U(u)$ follows a distribution that is close to $T_r^{\rightarrow}(u)$ as we do expect that users should be more viral for topics where they are more likely get retweeted. Similarly, the terms $\sum_{v \in S} ||\mathcal{S}(v) - T_r^{\leftarrow}(v) \cdot \sum_{k=1}^{K} s^k(v)||^2$ and $||\mathcal{V}_T - T_r^{\rightarrow} \cdot \sum_{k=1}^{K} v_T^k||^2$ ensure that $\mathcal{S}(v)$ and $\mathcal{V}_T$ follow distributions that are respectively close to $T_r^{\leftarrow}(v)$ and $T_r^{\rightarrow}$. Lastly, the terms $||\mathcal{V}_T||^2$ and $\Big[\sum_{u \in V} ||\mathcal{V}_U(u)||^2 + \sum_{v \in S} ||\mathcal{S}(v)||^2 \Big]$ are the Tikhonov regularization to avoid overfitting problem [5].

Minimizing $\mathcal{L}$ as in Equation 4.10 is a non-convex problem which could only be solved locally, e.g., by gradient based methods [9]. However, due to the conditions in Equation 4.9, we cannot directly apply the gradient descent methods as they are used for unconstrained problems. To deal with the conditions, we employ the following transformation to transform Problem 4.8 into a unconstrained problem.

$$(4.11) \qquad z = h^{-1}(x) \text{ for } \forall x \in [0,1]$$

where $h(z) = \frac{1}{2} \cdot \frac{exp(z) - exp(-z)}{exp(z) + exp(-z)} + \frac{1}{2}$ is a S-shape continuous monotone map from $\mathcal{R}$ to $[0,1]$.

Denote $\mathcal{Z}_T = \{h^{-1}(v_T^k), \text{ for } \forall k = 1, \cdots, K\}$, $\mathcal{Z}_U = \{h^{-1}(v_U^k(u)), \text{ for } \forall u \in V_k, \forall k = 1, \cdots, K\}$, and $\mathcal{Z}_S = \{h^{-1}(s^k(v)), \text{ for } \forall v \in S_k, \forall k = 1, \cdots, K)\}$, then Problem 4.8 becomes a unconstrained optimization problem with respect to $\mathcal{Z}_T$, $\mathcal{Z}_U$, $\mathcal{Z}_S$ which now can be solved using gradient descent based methods. Hence, the main computational cost in solving Problem 4.8 is at evaluating $\mathcal{L}$. From Equation 4.10, this includes (1) cost of computing the sum of squared errors, which is $\mathcal{O}(K_{dom} \cdot |\mathcal{K}|)$ since we normalized topic distribution of tweets so that each tweet has at most K_{dom} topics; and (2) cost of computing the regularization terms, which is $\mathcal{O}(K \cdot (2 + |V| + |S|))$. Finally, the cost of evaluating $\mathcal{L}$ is $\mathcal{O}(K_{dom} \cdot |\mathcal{K}| + K \cdot (2 + |V| + |S|))$, which is linear in all variables, i.e., number of retweet observations $|\mathcal{K}|$, number of topics K, and number of users $|V| + |S|$. Our method is therefore scalable to large datasets.

In our implementation, we employ the alternating gradient descent method. The main idea is to perform gradient descent steps by $\{\mathcal{Z}_U, \mathcal{Z}_S\}$ directions while keeping $\mathcal{Z}_T$ unchanged, and later perform gradient descent steps by $\mathcal{Z}_T$ directions, while keeping $\mathcal{Z}_U, \mathcal{Z}_S$ unchanged. In each gradient descent step, the step size is determined by the line search method [9]. This process repeats until we reach a predefined maximum number of iterations or when the values converge. We found that the converged measure values could be obtained within 50 alternating iterations, each iteration includes 20 gradient descent steps. The control parameters α_u, α_t, β_u, and β_t are also set through empirical evaluation on a large set of tuples of values. We found that parameter set $\alpha_u = \beta_u = 0.0001$ and $\alpha_t = \beta_t = 1$ gives the best performance. This parameter setting is reasonable as we have many more variables $v_U^k(u)$ and $s_k(v)$ that affects to only a subset of retweet observations where u and v are involved respectively; but in contrast, we have much fewer variables v_T^k that affect a much larger set of retweet observations (where the tweets are about topic k). Hence, the variables v_T^k should be regularized with larger weights.

5 Experimental Evaluation

In this section, we evaluate and compare our proposed method with some baseline methods in retweet prediction task using the One-Week dataset described in Section 3. This dataset includes tweets published by Singapore based users within one week up to Singapore General Election 2011.

5.1 Data Preprocessing Recall that for each topic k, we only can measure user virality for the topic for a subset V_k of users tweeting about the topic, and measure topic-specific user susceptibility to the topic for a subset of users S_k who receive and read tweets about the topic. We first initialize V_k and S_k to be the set of all users in our dataset. To ensure that we have sufficient observations for each user and each topic, we iteratively remove from V_k users who generate less than 5 tweets about the topic k, and remove from S_k users who read less than 5 tweets about the topic. Table 2 shows the statistics of the final dataset ExpDB which has much fewer users than the original dataset due to the filtering steps. Nevertheless we still have a large number of retweet obervations. The table also shows that ExpDB is highly imbalanced with only 4.1% positive observations. This makes the prediction task much more difficult.

5.2 Comparative methods Other than $V2S$, we employed the following methods for comparison.

LDA-based methods.

Table 2: Statistics of the experimental dataset (ExpDB)

Average $\lvert V_k \rvert$	50
#Unique users in $\bigcup_k V_k$	266
Average $\lvert S_k \rvert$	382
#Unique users in $\bigcup_k S_k$	1,677
#Original tweets	8,173
#Retweet Observations	138,721
#Positive Retweet Observations	5,731

- B_2: The likelihood of $R(u, v, m)$ depends on how much the message m matches with topics where u is more retweetable, and topics where v is more likely to retweet.

$$B_1(u, v, m) = \sum_{k=1}^{K} \left[T_k(m) \cdot T_t^{\rightarrow}(u)^k \cdot T_r^{\leftarrow}(v)^k \right]$$

- B_2: The likelihood of $R(u, v, m)$ depends on how much the message m matches with global retweetable topics, and topics where v is more likely to retweet.

$$B_2(u, v, m) = \sum_{k=1}^{K} \left[T_k(m) \cdot T_t^{\rightarrow k} \cdot T_r^{\leftarrow}(v)^k \right]$$

- B_3: A combination of B_1 and B_2. That is

$$B_3(u, v, m) = \tau \cdot B_1(u, v, m) + (1 - \tau) \cdot B_2(u, v, m)$$

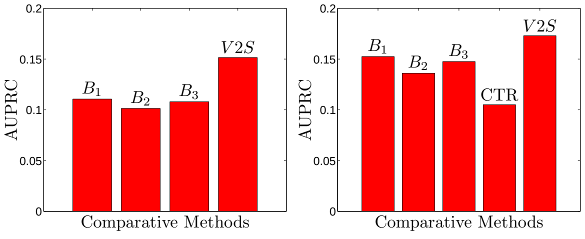
where $\tau \in (0, 1)$ is a parameter. In this experiment, we set $\tau = 0.5$.

In all of LDA-based methods, topic distribution of each tweet and each user is computed as in Section 3.

Collaborative Topic Regression (CTR). Proposed by Wang *et. al.* [19] that combines collaborative filtering data with content-based features to perform recommendation tasks. Similar to our proposed method, CTR is purely based on hidden user and content characteristics, and therefore is a suitable baseline. We used the authors' implementation with number of topics is also set to 100 and all other parameters are kept as default.

5.3 Metrics. We randomly divided both positive and negative retweet observations into training set and test set according to ratio 80% and 20%. This ensures that we have the same fraction of positive observations in both training and test sets. Then for each comparative method, we generate a ranking of observations in the test dataset based on the likelihood of retweet returned by the method. We then construct a Precision-Recall (PR) curve from the test set and the ranking, and measure the area under the PR curve (AUPRC). Methods with the higher AUPRC are the better.

Note that $V2S$ and LDA-based methods can be applied to the whole ExpDB, but CTR only works on users having at least one retweet. We therefore trained and tested CTR with subsets of training and test datasets that include including observations where



(a) LDA-based methods vs $V2S$ (using the whole ExpDB)

(b) All comparative methods (using the subset of ExpDB where the receiver has at least one retweet)

Figure 2: Prediction performance comparisons

the receiver has at least one retweet. We compare CTR against LDA-based methods and $V2S$ using these subsets, and compare $V2S$ against LDA-based methods using the whole dataset.

5.4 Performance comparision Figure 2(a) shows the performance of LDA-based method and $V2S$ method in the prediction task when the whole dataset is used. Figure 2(b) shows the performance of all comparative methods in the task using the subset of retweet observations where each receiver has at least one retweet. Among LDA-based methods, B_1 outperforms B_2 and B_3 in both cases. This suggests that user specific retweetable topics give a stronger retweet prediction than global retweetable topics. The fact that LDA-based methods outperform CTR can be explained that CTR suffers from noise as the model infers tweet topics and user preference simultaneously, while LDA-based methods does not since we employ the topic normalization step (see Section 3). Finally, in both Figures 2(a) and (b), the results indicate that our $V2S$ model is the best among all the comparative methods.

6 Analysis of Virality and Susceptibility

In this section, we analyze the user and topic behavioral factors, and the topic popularities obtained from the experiment in Section 5. We make empirical comparisons among the measures to show their differences, and provide example cases of highly viral topics/ users and highly susceptible users.

6.1 Topic Virality Figure 3(a) shows distribution of topic virality score. The figure clearly shows that most of the topics have some degree of virality; some topics are extremely viral with the virality score close to 1. The Pearson rank correlation coefficient between the topic global/ retweet popularity and topic virality is 0.686 and 0.912 respectively. This indicates that topic virality somewhat correlates but does not identical to global popularity or retweet popularity. Moreover, we found that the most viral topics do not necessarily have high global or retweet popularity; and neither the highest global popularity topics nor the highest retweet popularity topics are the most viral ones.

Table 3: Top topics by global popularity, topic retweet popularity, and virality

ID	Label	Rank		
		Global Popularity	Retweet Popularity	Virality
74	Campaign Events	**1**	7	42
64	SDP[6] Campaign	**2**	12	35
36	Election Results	**3**	**1**	**7**
2	$PoliticianA$[7]'s Speech	**4**	14	39
5	PAP's Victories	**5**	**3**	31
99	Job Advertisement	**7**	92	93
17	Spoke out against Opposition Parties	**8**	22	45
12	Foreign Policy	**9**	17	5
15	Polling Day	**10**	11	**2**
10	WP[8]'s Victories	11	**2**	25
50	Election result at Aljunied[9]	16	**4**	**9**
75	Election result at Jo Chiat[10]	24	**5**	18
58	$PoliticianB$'s Lost	19	**6**	
16	Supporters at Rally Speeches	15	**8**	27
55	$PoliticianC$'s Lost	29	**9**	11
80	$PoliticianD$	27	10	**1**
70	Emotional responses to Election Results	61	45	**3**
4	Waiting for Election Results	49	29	**4**
14	Media Channels Reporting Election Results	28	25	**6**
86	Oppositions in Parliament	20	13	**8**

Table 3 shows the top ten topics by global popularity, retweet popularity, and topic virality. Note that the topic labels are manually assigned based on the topic representative words (which are excluded from the table due to space limitation), and further insights from the top tweets of each topic. While most globally popular topics are related to election campaign, the most retweeted ones are about results of the election. This is reasonable as the dataset covers tweets generated in the week before the election, including the day when the election results were announced. Finally, the most viral topics are about episodes around the election (topic 80) and emotional responses to the election results (topic 70). The virality of topic 80 is expected as the election candidate $PoliticianD$ suffered from online flaming throughout her election campaign. The virality of the topic 70 agrees with the prior works by Berger *et. al.* [1] and Pierce *et. al.* [13] that people are more likely to share information that evokes high-arousal emotions.

We further examine "Job Advertisement" (topic 99), "WP's Victories" (topic 10), and "$PoliticianD$" (topic 80), the three topics that are ranked highly by global popularity, retweet popularity and topic virality respectively but not by others. For each topic, we select a set of tweets with the normalized probability of the topic (see Section 3) is not smaller than $\theta = 0.5$, and call them the *on-topic tweets*.

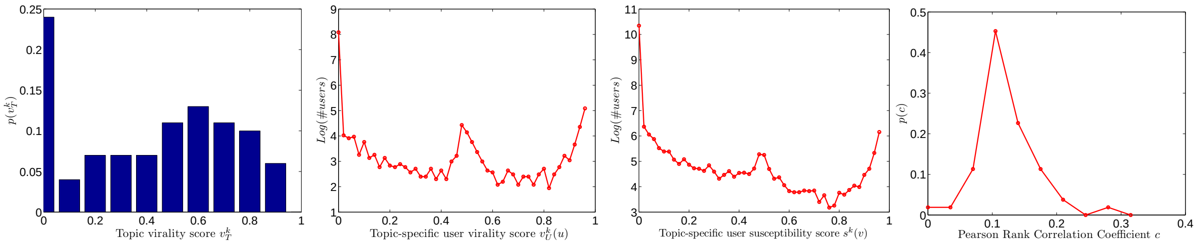
[6] http://en.wikipedia.org/wiki/Singapore_Democratic_Party

[7] We do not show the politician names as they are not the focus of this paper

[8] http://en.wikipedia.org/wiki/Workers'_Party_of_Singapore

[9] http://en.wikipedia.org/wiki/Aljunied_Group_Representation_Constituency

[10] http://en.wikipedia.org/wiki/Joo_Chiat_Single_Member_Constituency



(a) Distribution of Pearson rank correlation between topic-specific user virality and susceptibility (b) Distribution of topic-specific user virality score (c) Distribution of topic-specific user susceptibility score (d) Distribution of Pearson rank correlation between topic-specific user virality and susceptibility

Figure 3: Topic virality and topic-specific user virality and susceptibility

Table 5: Top-10 viral users across the topics

Rank	User name	User type	Topic ID	Label
1	todayonline	Media	36	Election Results
2	stcom	Media	5	PAP's Victories
3	temasekreview	Blogger	10	WP's Victories
4	todayonline	Media	10	WP's Victories
5	stcom	Media	55	$PoliticianC$'s Lost
6	ge2011	Portal	36	Election Results
7	todayonline	Media	58	$PoliticianB$'s Lost
8	temasekreview	Blogger	80	$PoliticianD$
9	todayonline	Media	75	Election result at Jo Chiat
10	wpsg	Party Portal	74	Campaign Events

Table 6: Top-10 susceptible users across the topics

Rank	User name	User type	Topic ID	Label
1	mdamerhamzah	Young adult	74	Campaign Events
2	flippers1452	Teenage	75	Election result at Jo Chiat
3	automathic	Social Activists	72	$PoliticianE$'s speech
4	boonkhiang	Artist	36	Election Results
5	nujnewyohc	Teenage	36	Election Results
6	ekeked	Young adult	10	WP's Victories
7	rnlni7	Young adult	36	Election Results
8	nicocakes	Young adult	36	Election Results
9	cherylquincy	Young adult	36	Election Results
10	energywen	Blogger	10	WP's Victories

Table 4(a) shows that all tweets about "Job Advertisement" (topic 99) did not get any retweet even though the topic is globally popular. We found that the on-topic tweets of topic 99 are mostly generated by *sg_job_adm* and *sg_job_marketin*, the two users who follow each other and promote their advertising tweets[11].

Table 4(b) shows that most of retweets on the topic "WP's Victories" (topic 10) are due to the top three users: *stcom*, *temasekreview*, and *todayonline*. These users are highly viral on the topic and followed by many users. Hence, these users can enjoy lots of retweets on the topic. On the other hand, the top three retweeted users on the topic "*PoliticianD*" (topic 80) do not contribute a major fraction of retweets on the topic, most of user got retweeted on the topic 80 have similar fraction of retweets. Furthermore, these most retweeted user on the topic 80 are not the most viral ones. This suggests that most retweets on the topic 80 are due to the virality of the topic. Therefore, although topic 10 has higher retweet popularity than topic 80, it is reasonable to assign topic 80 a higher topic virality rank.

6.2 User Virality and User Susceptibility

Figure 3(b) shows the distribution of user virality score across topics. The figure shows that most of users are not viral for most of the topics, while a very small group of users are highly viral for some topics. Table 5 lists the top ten viral users across topics. The table shows that most of the extremely viral users are mass media streams who tweet about news they broadcast. The table also shows that topics having extremely viral users are all related to the election.

Figure 3(c) shows the distribution of user susceptibility score across topics. Similar to user virality, the figure shows that most of users are not susceptible to most of the topics, while a very small group of users are highly susceptible to some topics. Table 6 lists the top ten susceptible users across topics. Again, topics having extremely susceptible users are all related to the election. The table also shows that most of extremely susceptible users are at the young ages. This suggests that young people tend to be more susceptible to online events.

Finally, we examine the correlation between topic-specific user virality and susceptibility. For user u, we count the number of topics $K_m(u)$ where u's topic-specific virality and susceptibility are both measurable. Then, for each of user u with $K_m(u) \geq 5$, we compute Pearson rank correlation coefficient between the u's topic-specific user virality and user susceptibility over the topics where the u's behavioral factors are mesurable. The distribution of the coefficients is shown in Figure 3(d). The figure clearly shows that across topics, user virality has quite a low correlation with user susceptibility. This indicates that the likelihood that a user gets retweets for a topic does not depend on whether she retweets for the topic.

7 Conclusion

In this paper, we propose a novel framework to model retweet actions based on user and content behavioral factors. Motivated by differences between topic distribution of the set of tweets and the bag-of-retweets at both network and user levels, we model retweets using user and item behavioral factors to the topic level. Our framework takes into account topic-specific user virality

[11] Both *sg_job_adm* and *sg_job_marketin* have been suspended by Twitter

Table 4: Comparion of “Job Advertisement” (topic 99), “WP’s Victories” (topic 10), and “*PoliticianD*” (topic 80).

(a) On-topic tweets and retweets

ID	Users generating on-topic tweets			On-topic tweets			Users receiving on-topic tweets		
	#Users	#Users with retweeted on-topic tweets	#Retweets per user	#Tweets	#Tweets with retweeted	#Retweets per tweet	#Users receiving	#Users retweeting	#Retweets per user
99	6	0	0	258	0	0	28	0	0
10	56	23	9.16	184	96	2.79	1085	240	2.14
80	56	26	1.57	104	29	1371	428	66	1.38

(b) On-topic most retweeted users

ID	Top Retweeted Users								
	User name		#On-topic tweets (% in all)	#On-topic tweets with retweeted (% in all)	#Retweets (% in all)	Virality rank	#Receivers	Average of #Retweeters #Receivers (%)	Average Retweeter’s Susceptibility
10	stcom	Media	17 (**9.23**)	16(**16.7**)	145 (**28.26**)	24	2697	10.2	0.5
	temasekreview	Blogger	24 (**13.04**)	21 (**21.88**)	140 (**27.29**)	1	938	16.3	0.63
	todayonline	Media	18 (**9.78**)	17 (**17.71**)	132 (**25.73**)	2	1767	7.03	0.46
80	thenooselite[12]	Comedy	4 (**3.5**)	4 (**10**)	16 (**17.6**)	2	96	15.4	0.56
	stcom	Media	2 (**1.75**)	2 (**5**)	12 (**13.2**)	34	404	2.91	0.54
	fakemoe	Parody	2 (**1.75**)	2 (**5**)	4 (**4.4**)	22	74	11.01	0.79

and susceptibility, and topic virality as predictive factors for retweet actions. We develop $V2S$, a tensor factorization based model, to measure these behavioral factors based on users’ tweeting and retweeting historical data. Our experiments on a Twitter dataset shows that the proposed $V2S$ model outperforms baseline models.

In the future, we would like to calibrate more fine-grained factors contributing to retweeting. These factors include users’ positions in the network, linguistic features in content, and psychological factors of users.

8 Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

[1] J. A. Berger and K. L. Milkman. What makes online content viral? *Social Science Research Network*, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[3] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. *ICDM Workshops*, 2010.

[4] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In *SIGIR*, 2012.

[5] G. H. Golub, P. C. Hansen, and D. P. OLeary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 1999.

[6] T.-A. Hoang and E.-P. Lim. Virality and susceptibility in information diffusions. In *ICWSM*, 2012.

[7] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *Int. J. Emergency Management*, 2010.

[8] S. Jurvetson. From the ground floor: What exactly is viral marketing? *Red Herring Communications*, 2000.

[9] C. T. Kelley. Iterative Methods for Optimization. *Frontiers in Applied Mathematics*, SIAM, 1999.

[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW’10*, 2010.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD ’09*, 2009.

[12] H.-K. Peng, J. Zhu, D. Piao, R. Yan, and Y. Zhang. Retweet modeling using conditional random fields. In *Data Mining Workshops, ICDM*, 2011.

[13] D. Pierce, D. P. Redlawsk, W. W. Cohen, T. Yano, and R. Balasubramanyan. Social and affective responses to political information. *APSA*, 2012.

[14] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW ’11*, 2011.

[15] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *ICWSM*, 2011.

[16] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom*, 2010.

[17] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Comm. ACM*, August 2010.

[18] A. Wang, T. Chen, and M.-Y. Kan. Re-tweeting from a linguistic perspective. In *NAACL-HLT 2012 Workshop on Language in Social Media*, 2012.

[19] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD ’11*, 2011.

[20] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM*, 2010.

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

[22] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *CIKM*, 2010.

[23] L. Zhilin, W. Xintao, C. Wandong, and D. Peng. Examining multi-factor interactions in microbloging based on log-linear modeling. In *ASONAM*, 2012.

[24] Z. Zhou, R. Bandari, J. Kong, H. Qian, and V. Roychowdhury. Information resonance on twitter: watching iran. SOMA, 2010.