6-2012

# Virality and susceptibility in information diffusions

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

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

## Citation

HOANG, Tuan-Anh and LIM, Ee Peng. Virality and susceptibility in information diffusions. (2012). *Proceedings of the Sixth International Conference on Weblogs and Social Media*.
Available at: https://ink.library.smu.edu.sg/sis_research/1546

# Virality and Susceptibility in Information Diffusions

**Tuan-Anh Hoang, Ee-Peng Lim**
Singapore Management University, 80 Stamford Road, Singapore 178902
Email: tahoang.2011@phdis.smu.edu.sg, eplim@smu.edu.sg

## Abstract

Viral diffusion allows a piece of information to widely and quickly spread within the network of users through *word-of-mouth*. In this paper, we study the problem of modeling both item and user factors that contribute to viral diffusion in Twitter network. We identify three behaviorial factors, namely user virality, user susceptibility and item virality, that contribute to viral diffusion. Instead of modeling these factors independently as done in previous research, we propose a model that measures all the factors simultaneously considering their mutual dependencies. The model has been evaluated on both synthetic and real datasets. The experiments show that our model outperforms the existing ones for synthetic data with ground truth labels. Our model also performs well for predicting the hashtags that have higher retweet likelihood. We finally present case examples that illustrate how the models differ from one another.

**Keywords.** *viral diffusion, diffusion related factors, Twitter network*

# Introduction

## Motivation

Recent empirical research has shown that information diffusion occurs in Twitter[1](Kwak *et. al.* (2010), Zhou *et. al.* (2010), Romero *et. al.* (2011)). A Twitter user follows other users so as to receive the latter's messages. Other than publishing original messages or *tweets*, Twitter users may re-publish messages they received from their followees. These re-published messages are called *retweets*. This re-publishing mechanism is how messages can go through follow links from the original authors to their followers and followers of followers.

With important business applications, e.g., customer sentiment monitoring, viral marketing, etc, it is increasingly important to study viral diffusion in Twitter. There has been a number of research projects on finding user and item factors affecting the process of viral diffusion (Leskovec *et. al.* (2007; 2009), Broxton *et. al.* (2010), Szabo *et. al.* (2010), Guerini *et. al.* (2011), Li *et. al.* (2010), Weng *et. al.* (2010), Ienco *et. al.* (2010), Ratkiewicz *et. al.* (2010), Hoang *et. al.* (2011), Rowe (2011)). However, most of these works consider the effects of users and items independently. This common approach neglects effects of the underlying user network on the diffusion as well as excludes the relationship among the factors. Some other works, on the other hand, suggest that there are interrelationships among the item diffusion, the user network, and other user factors (Valente (1995), Kiecker *et. al.* (2001), Cowan (2004), Broxton *et. al* (2010), Lee *et. al* (2009), Ratkiewicz *et. al* (2011), Petrovi *et. al* (2011)).

## Research Objectives and Contributions

In this paper, we aim to identify and model factors that contribute to viral diffusion based on the interrelationships among items, users, and the user-user network. Once a user adopted an item that is introduced to her by her friends, we conjecture that this adoption may be due to influence of two set of factors. The first set includes the factors that are external to the network, e.g., advertising. The second set is called internal factors that include the virality of the item itself, the virality of the users diffusing the item, and the susceptibility of the user adopting the item. Here, susceptibility, which is newly introduced in this work, is a user factor that indicates whether a user can easily be convinced to adopt items introduced to her. We therefore propose to quantitatively measure *item virality*, *user virality*, and *user susceptibility* within a common framework. The challenges of this work are as follows.

- The effect of each of the three factors is not explicitly differentiated. They therefore cannot be measured separately.
- The information about who diffuses what items to whom is not given before hand. The diffusion process therefore cannot be clearly observed.

To deal with above challenges, we first identify item adoptions due to the effects of external and internal factors. Then, the items diffusing in the network and the users diffusing the items are extracted from the adoption logs and the user network. Lastly, we simultaneously measure the three factors using a model that is built based on the mutual dependencies among the factors.

This work improves the state-of-the-art of information diffusion. To best of our knowledge, there has not been any work modeling different factors of information items, users, and the underlying network. Our main contributions in this work are as follows.

[1] https://twitter.com

- We consider virality at both the item and user side, and introduce user susceptibility as a factor affecting item diffusion.
- We propose a novel quantitative modeling framework which utilizes the mutual dependency between item virality, user virality, and user susceptibility to measure these factors simultaneously.
- We develop an iterative computation algorithm to compute scores of item virality, user virality, and user susceptibility.
- We compare and contrast our model with other related measures, including the popularity and viral coeffecient (Penenberg (2009)), in our experiments and case studies. The experiments are conducted on both synthetic and real datasets for different purposes of evaluating the models.
- We propose a task to predict retweet likelihood order for hashtags in Twitter. The results show that the virality scores assigned to the hashtags using our proposed model can be used to predict the order more accurately than other models.

It is worthwhile to note that virality is related but not identical to influence. Firstly influence is associated with users only, while virality can be factors of users and items. Secondly, influential users are those who affect the way other users behave. This means influence is an aggregated behavior that is derived from various user actions. On the other hand, virality is about the strength of a user in propagating information items. A viral user is therefore not necessari a influential one and vice versa.

## Related Works

User and item factors that contribute to viral diffusion have been examined in both academia and industry. To measure the virality of an item or *item virality*, *popularity* (number of adopters) and *rate of popularity gain* of the item are often used (Jurvetson (2000), Broxton *et. al* (2010), Shamma *et. al* (2011)). Popularity however does not consider the *word-of-mouth* effect in viral diffusion. In our previous work (Hoang *et. al* (2011)), we introduced *retweet count* (the number of times a tweet is retweeted), and *retweet likelihood* (the fraction of followers retweeting) to identify viral tweets generated by a set of Twitter users interested in Singapore's socio-political topics. Retweet count and other similar measures capture the likelihood of spreading by *word-of-mouth* missing from the popularity measures. *Viral coefficient*, defined as the average number of new adopters generated by each existing adopter, was proposed by Penenberg to measure item virality (Penenberg (2009)). For Twitter data, viral coefficient of a tweet is the same as retweet count per user.

Calibrating item virality is different from designing items to be viral. As shown in work by Guerini *et. al.* (2011), there are fine grained factors contributing to item virality, e.g., *appreciation*, or *raising discussion*. Determining these factors and deriving the exact relationship between these factors and item virality is still an unexplored research topic. In this paper, we shall only focus on the measurement of item virality leaving the modeling of other factors for future research.

In our previous work (Hoang *et. al* (2011)), we also proposed to measure the virality of Twitter users by their efforts in tweeting or retweeting viral tweets. In another piece of work, Janghyuk *et. al.* defined the virality of a user in a marketing campaign by (a) the amount of time the user's friends take to respond to recommendations received from the user; and (b) the number of unique friends the user sends recommendation to after adopting an item (2009). This work however ignores the effects of user network of the user and their willingness to respond to the recommendations, and only considers a specific item. In this paper, we incorporate approaches of theses two works and user network in measuring user virality with respect to a large number of items.

The notion of susceptible user was used to describe a user state in the SIR and SIRS diffusion models developed in epidemiological studies (Bailey (1975), Anderson *et. al* (1992)). Users in susceptible state can be infected with some virus from their neighbors. SIR and SIRS models implicitly assume that all user have the same degree of susceptibility. We, on the other hand, assume that susceptibility may vary across different users, and measure susceptibility of each user. In the recent work by Dabeer *et. al* (2011), the authors introduced *response probability* which bears some similarity with susceptibility. The former however measures the likelihood specific to a pair of user and item, while the later is user specific only and does not depend on any item.

## Virality Modeling

### User Network Definitions and Assumptions

As shown in Table 1, we represent a set of users $U$ and their follow relationships by a directed graph $G = (U, E)$. A directed edge $(v, u) \in E$ represents the fact that $v$ *follows* $u$. The time at which $u$ follows $v$ is denoted by $t(u \rightarrow v)$. For simplicity, we assume that $u$ follows $v$ once only. Items in Twitter can be URLs, hash tags, person names, or any other identifiable information entities. We use $X$ to denote the set of such items, $X(u)$ to denote the set of items user $u$ adopts, and $t_u(x)$ to denote the time when $u$ adopts item $x$.

We say that $u$ **diffuses item** $x$ **to** $v$, or item $x$ is diffused from $u$ to $v$, if the following conditions hold:

- $u$ adopts $x$ before $v$ adopts (i.e., $t_u(x) < t_v(x)$), and $v$ adopts $x$ after $v$ follows $u$ (i.e., $t_v(x) < t(v \rightarrow u)$) ; and
- $u$ introduces $x$ to $v$ before $v$ adopts $x$ by at most some time threshold $\tau$

Here, we assume that each user may adopt an item at most one time but she may introduce the item to the same friend more than once. The ways in which a user introduces items she adopted to her friends are different in different specific networks. For example, in Twitter, $u$ may introduce a hashtag to her followers every time she posts a tweet containing the hashtag. The *time threshold* $\tau$ is introduced to determine if $v$ is diffused by $u$. Using $\tau$ to tell the social influence from one user to another has been used in several previous works, e.g., Anagnostopoulos *et. al* (2008), Dave *et. al* (2011). When $v$ adopts $x$ at some time point, $v$ may have several of her followee(s) who already adopted $x$ within $\tau$ time units ago. In this case, we say that $v$ is diffused by *multiple* followees. Our model allows a user to be diffused by multiple followees and to diffuse to multiple followers.

We use $u \xrightarrow{x} v$ to denote the fact that $u$ diffuses the item $x$ to $v$. The set of items diffused by $u$ (to her followers) is denoted by $X_d^{\rightarrow}(u)$ ($X_d^{\rightarrow}(u) \subseteq X(u)$), and the set of items diffused to $u$ (by her followees) is denoted by $X_d^{\leftarrow}(u)$ ($X_d^{\leftarrow}(u) \subseteq X(u)$).

We denote the set of users to whom $u$ diffuses $x$ to by $F_d^{\rightarrow}(u,x)$, and the set of followees of $u$ who diffuse $x$ to $u$ by $F_d^{\leftarrow}(u,x)$. Item $x$ is introduced to user $v$ (through *word-of-mouth*) if there is at least one followee of $v$, say $u$, introducing $x$ to $v$. We denote the set of items introduced to $u$ by $X_i(u)$, and the set of users whom $x$ is introduced to by $U_i(x)$.

Since not all users have chances to diffuse items to their friends, or to have items introduced to them from the friends, we may not be able to measure virality and susceptibility for every users due to the lack of historical observations. Instead, we identify the subset $V \subseteq U$ including users introducing items to their friends, and the subset $S \subseteq U$ including users having items introduced to them. We then measure virality and susceptibility for users in $V$ and in $S$ respectively.

Table 1: Notations.

| | |
|---|---|
| $U$ / $X$ | Set of users / items respectively |
| $V$ | Target users for virality |
| $S$ | Target users for susceptibility |
| $N$ / $M$ | Number of users / items respectively |
| $X(u)$ | Set of items adopted by $u$ |
| $X_d^{\rightarrow}(u)$ | Set of items diffused by $u$ |
| $X_d^{\leftarrow}(u)$ | Set of items diffused to $u$ |
| $X_i^{\leftarrow}(u)$ | Set of items introduced to $u$ |
| $U(x)$ | Set of users adopting $x$ |
| $U_d(x)$ | Set of users who diffuse $x$ to $> 0$ users |
| $U_i(x)$ | Set of users $x$ is introduced to |
| $u \overset{x}{\rightarrow} v$ | $u$ diffused the item $x$ to $v$ |
| $F_d^{\rightarrow}(u,x)$ | Set of followers whom $u$ diffuses $x$ to |
| $F_d^{\leftarrow}(v,x)$ | Set of followees who diffuse $x$ to $v$ |
| $F_i^{\leftarrow}(v,x)$ | Set of followees who introduce $x$ to $v$ |

For simplicity, we assume that all users in the network are not aware of the models to be used for measuring their properties related to virality. Hence, no users are expected to game or abuse the models to be introduced and the network is spam free. This assumption does not always hold and we shall address it in our future research.

## Virality Model: Components and Existing Models

In viral diffusion, users and items are the only entities involved. In virality modeling, we aim to model the properties of users and items that contribute to viral diffusion. Instead of trying to enumerate all properties which is still an ongoing research topic, we define three basic user and item properties that play distinct roles in diffusion. The three properties are:

- **Item virality**: This refers to the ability of an item to spread the adoptions by users through the follow links. We denote the item virality of item $x$ by $I(x)$.
- **User virality**: This refers to the ability of a user to spread the adoptions to other users through the follow links. We denote the user virality of user $u$ by $V(u)$.
- **User susceptibility**: This refers to the ability of a user to adopt items easily as other neighboring users diffuse the items to her. We denote the user susceptibility of user $u$ by $S(u)$.

For standardized interpretation across different virality models, we would like $I(x)$, $V(u)$ and $S(u)$ to be measured by some numeric score values within the range of [0,1]. In the following, we review some existing virality models that have been introduced in previous works. Most of them cover only one of the above three user and item properties. To differentiate the properties derived by different models, we adorn the notations $I_{\langle model \rangle}(x)$, $V_{\langle model \rangle}(u)$ and $S_{\langle model \rangle}(u)$ with the $\langle model \rangle$ subscript.

**Item virality.** Two widely used item virality definitions are popularity and viral coefficient[2].

- **Popularity** is defined as the number of users adopting the item.
$$I_p(x) = |U(x)|/N \tag{1}$$
- **Viral coefficient** is the average number of friends that a user diffuses the item to once she has adopted the item.
$$I_c(x) = \frac{|U_i(x) \cap U(x)|}{|U_d(x)|} \tag{2}$$

The popularity captures how widely the item is adopted but it does not tell if the adoptions are due to word-of-mouth or external influence such as media advertisements. The viral coefficient is defined purely based on the item adoptions due to word-of-mouth. When viral coefficient exceeds 1.0, it means each user adopting the item is able to get more than one other users adopt the item, making the item diffusion becomes viral. Viral coefficient however does not consider the user properties.

**User virality.** The conventional approach to measure user virality is **Fan-out**, i.e., the average number of friends she diffuses items to. That is,
$$V_f(u) = \frac{\sum_{x \in X_d(u)} |F_d^{in}(u,x)|}{|X_d(u)|} \text{ for } \forall u \in V \tag{3}$$

**User susceptibility.** To the best of our knowledge, user susceptibility has not been modeled in the previous works. For simplicity, we use **Fan-in**, i.e., the fraction of items she adopts once they are introduced to her.
$$S_f(v) = \frac{|X_d^{\leftarrow}(v)|}{|X_i(v)|} \text{ for } \forall v \in S \tag{4}$$

## Our Proposed Model

Viral diffusion in a network is caused by interactions among users as well as interactions between users and items being diffused. Given that these interactions occur in a network, to measure user and item properties from these interactions, one has to consider the mutual dependency relationships among the properties. In this section, we therefore propose a **Mutual Dependency Model** that measures item virality, user virality, and user susceptibility simultaneously based on a set of principles that help to distinguish each property from others in viral diffusion.

The three principles are:

- Viral items (with *high item virality*) can be diffused from less viral users (with *low user virality*) to less susceptible users (with *low user susceptibility*).

[2] This original definition does not have a measure between 0 and 1, but can be easily normalized.

- Viral users (with *high user virality*) can diffuse less viral items (with *low item virality*) to less susceptible users (with *low user susceptibility*).
- Susceptible users (with *high user susceptibility*) adopts less viral items (with *low item virality*) introduced to her from less viral users (with *low user virality*).

We operationalize the above three principles into the following item virality, user virality, and user susceptibility measures:

$$I_m(x) = \frac{1}{|U(x)|} \cdot \sum_{u \in U_d(x)} [(1 - V_m(u)) \cdot \sum_{v \in F_d^{\rightarrow}(u,x)} \frac{1 - S_m(v)}{|F_d^{\leftarrow}(v,x)|}] \tag{5}$$

$$V_m(u) = \frac{1}{|X(u)|} \cdot \sum_{x \in X_d^{\rightarrow}(u)} [(1 - I_m(x)) \cdot \sum_{v \in F_d^{\rightarrow}(u,x)} \frac{1 - S_m(v)}{|F_d^{\leftarrow}(v,x)|}] \text{ for } \forall u \in V \tag{6}$$

$$S_m(v) = \frac{1}{|X_i^{\leftarrow}(v)|} \cdot \sum_{x \in X_d^{\leftarrow}(v)} [(1 - I_m(x)) \cdot \frac{1}{|F_i^{\leftarrow}(v,x)|} \cdot \cdot \sum_{u \in F_d^{\leftarrow}(v,x)} (1 - V_m(u))] \text{ for } \forall v \in S \tag{7}$$

In Equations (5 - 7), the terms $(1 - I_m(x))$, $(1 - V_m(u))$, and $(1 - S_m(u))$ are inverses of item virality, user virality, and user susceptibility respectively. In Equation 5, the virality of an item is derived from the number of adoptions of the item by a set of diffusing users ($U_d(x)$) and a set of diffused users ($F_d^{\rightarrow}(u,x)$) after weighting the former by the inverse of their user virality and the latter by the inverse of their user susceptibility prorated by the number of other users who diffuse the same item to them ($|F_d^{\leftarrow}(v,x)|$). Given that $I_m(x)$ considers adoption count per diffusing user, it is an extension of viral coefficient $I_c(x)$.

In Equation 6, the virality of a user is derived from the number of adoptions of items she diffuses ($X(u)$) to a set of users ($F_d^{\rightarrow}(u,x)$) after weighting the items by their inverse item virality and the diffused users by their inverse susceptibility prorated by the number of other users who diffuse the same item to them ($|F_d^{\leftarrow}(v,x)|$). $V_m(u)$ is an extension of user virality based on fanout $V_f(u)$ as both consider the number of users diffused by $u$, i.e., $F_d^{\rightarrow}(u,x)$.

In Equation 7, the susceptibility of a user is measured by the number of adoptions of items she is introduced ($X_i(u)$) by a set of users ($F_i^{\leftarrow}(u,x)$) after weighting the items by their inverse item virality and the average inverse user virality of the introducing users who succeeded in diffusion. $S_m(u)$ also shares some similarity with $S_f(u)$ in using $X_d^{\leftarrow}(u)$.

### Model Computation

Computing the mutual dependency model is a fixed point problem (Zeidler (1995)). We employ the iterative computation method in Algorithm 1 to compute $I_m(x)$'s, $V_m(u)$'s and $S_m(u)$'s. The main idea is to initialize $V_m(u)$'s and $S_m(u)$'s with some values so as to compute $I_m(x)$'s. The computed $I_m(x)$'s and $S_m(u)$'s are then used to compute a new set of values for $V_m(u)$'s. Next, the new $S_m(u)$ values are computed from $I_m(x)$'s and $V_m(u)$'s. This process repeats until we reach a predefined maximum number of iterations or when the values converge.

We found that the iterative computation method works well for all the synthetic and real datasets (more than 50 of them) in our project and we could obtained converged measure values in less than 20 iterations. Proving the convergence of the method is however elusive and is part of our ongoing research.

**Algorithm 1** Iterative computation method for computing item virality, user virality, and user susceptibility

```
$(I_m(\cdot), V_m(\cdot), V_m(\cdot)) \leftarrow (\vec{\mathbf{1}}, \vec{\mathbf{1}}, \vec{\mathbf{1}})$    ▷ Initialization
$C \leftarrow (I_m(\cdot), V_m(\cdot), S_m(\cdot))$    ▷ Normalization
$(I_m(\cdot), V_m(\cdot), S_m(\cdot)) \leftarrow (I_m(\cdot), V_m(\cdot), S_m(\cdot))/||C||$
for $k \leftarrow 1$ to $MaxIteration$ do    ▷ Compute $I_m(\cdot)$, $V_m(\cdot)$, and $S_m(\cdot)$ using the iterative computation method
    for each $x \in X$ do
        Compute $I^{'}(x)$ using Equation 5
    end for
    for each $u \in U$ do
        Compute $V^{'}(u)$ using Equation 6
    end for
    for each $v \in S$ do
        Compute $S^{'}(v)$ using Equation 7
    end for
    $C \leftarrow (I^{'}(\cdot), V^{'}(\cdot), S^{'}(\cdot))$    ▷ Normalization
    $(I_m(\cdot), V_m(\cdot), S_m(\cdot)) \leftarrow (I^{'}(\cdot), V^{'}(\cdot), S^{'}(\cdot))/||C||$
end for
Normalize $I_m(\cdot)$, $V_m(\cdot)$, and $S_m(\cdot)$ to unit length
```

## Experiments on Synthetic Datasets

The first set of experiments is designed to evaluate and compare the different virality models including our proposed mutual dependency model. While some of them have been used in the commercial world, a systematic evaluation has not been conducted due to a lack of an existing dataset containing the ground truth labels of viral items, viral and susceptible users. We therefore create synthetic datasets with different parameter settings and corresponding ground truths and compare the models' accuracies.

### Synthetic Data Generation

We use the following steps to generate a synthetic dataset.

- **Generating the user network**. Given the number of users $N$, power law degree exponent $\alpha$, minimum degree $d_{min}$, and maximum degree $d_{max}$, we generate a undirected network of users whose degree distribution follows the power law with exponent $\alpha$ as follows.
  - Generate the degree distribution of $N$ nodes in the $[d_{min}, d_{max}]$ range following the power law distribution using the *inverse transformation method* Ross (2006)).
  - Generate the links for the $N$ nodes to follow the generated degree distribution using the *Expected Degree Model* (Chung *et. al*(2002)). The resultant network has each connected pair of users follow each other.

- **Generating the ground truth**. We designate a small number of users, let say $k_u$ of $N$, who are randomly chosen from users among the top 10 degree percentile, as viral users. This is to ensure that viral users have sufficient followers to be diffused. The susceptible users are selected the same way. These users are assigned higher virality/ susceptibility scores that are uniformly drawn from [$1 - \beta_u$, 1), while the remaining users are assigned virality/ susceptibility scores uniformly in the range [0 , $\beta_u$]. We label $k_i$ of $M$ items to be viral. These items have virality scores randomly drawn from [$1 - \beta_i$, 1), while the remaining items have scores randomly drawn from [0, $\beta_i$).
- **Generating the items adoptions**. We generate item adoptions for each item $x$ at $StepCnt$ time steps. At each time step, as suggested by the Bass Model (Bass (1969)), the probability that each non-adopter $v$ adopts $x$ is $p + q$ where $p$ is the probability attributed to external influence, and $q$ is the probability attributed to internal influence or diffusion.
$$q = \frac{1}{3} \cdot (1 - \prod_{u \in n_v^t} (1 - V(u)) + I(x) + S(v)) \qquad (8)$$
where $n_v^t$ is the set of neighbors of $v$ who adopt $x$ within $\tau$ time steps ago. In our experiments, we set $StepCnt = 10$ and $\tau = 1$.

We generated networks with different number of users ($N$), number of items ($M$), user virality/ susceptibility score width ($\beta_u$), and item virality score width ($\beta_i$) parameter settings while fixing $\alpha = 2.5$, $d_{min} = 1$, $d_{max} = 100$, $k_u = 1\%$, and $k_i = 10\%$. For each parameter setting, we generate 10 instances of item adoptions with $p$ is randomly chosen from $[0.01, 0.05]$ for each item. This range of $p$ is also suggested by experiments on a various type of items reported in Bass (1969) and Turk *et. al* (2012). We then compute the virality and susceptibility scores of each dataset instance using different models. For the mutual dependency model, the $MaxIterations$ constant in Algorithm 1 is set to 20.

## Results

**Performance Comparisons.** For each dataset instance, we rank users by their virality (susceptibility) scores produced by a virality model and select the top scored $1\%$ users as the predicted viral (susceptible ) users and denote the set by $U_v^p$ ($U_s^p$). The precision@$1\%$ of user virality (susceptibility) is then defined by $\frac{|U_v^p \cap U_v|}{|U_v|}$ ($\frac{|U_s^p \cap U_s|}{|U_s|}$) where $U_v$ and $U_s$ denote the viral users and susceptible users in the ground truth respectively. The precision@$10\%$ of item virality is similarly defined.

Figures 1(a) and 1(d) show the precision@$10\%$ of item virality and precision@$1\%$ of user virality and susceptibility for the different models as we set $N$ = 1000, 10K, 20K and 50K keeping $M = 500$ and $\beta_u = \beta_i = 0.3$. The figures show that the mutual dependency model outperforms other models, particularly for item virality and user susceptibility. The performance of mutual dependency model in user virality is only slightly better than that of fan-out. All models demonstrate decreasing precision as $N$ increases. They however still outperform the random selection significantly.

Figures 1(b) and 1(d) show the precision@$10\%$ of item virality and precision@$1\%$ of user virality and susceptibility respectively for the different models as we set $M = 100$, 200 and 500 keeping $N = 50K$ and $\beta_u = \beta_i = 0.3$. The figures show that the mutual dependency model outperforms all other models. All models demonstrate unchanged precision as $M$ increases.

Figures 1(c) and 1(f) show the precision@$10\%$ of item virality and precision@$1\%$ of user virality and susceptibility respectively for the different models as we set the score width $\beta_u = \beta_i = 0.1$ to 0.5 keeping $N = 50K$ and $M = 500$. Again, the mutual dependency model outperforms the other. The precision generally falls as we increase the score width. This is expected as larger score width creates ground truth data harder for the models.

# Experiments on a Real Dataset

In this section, we compare the different models using a real Twitter dataset containing tweets published by Singapore-based users during the Singapore's 2011 general election and presidental election. Since the elections are socially interesting events, we expect viral diffusion to exist in the data.

## Data Collection and Preprocessing

We first selected a set of of 58 Singapore-based seed users which includes user accounts of the political parties, politicians, political commentators, and bloggers. We then derived the followers and followees of the seed users creating a larger set of 32,138 users who declared themselves to be located in Singapore. We crawled tweets published by the set of users on a daily basis. We collected a set of 30,652,126 tweets published between March and September 2011 for this study. Among those tweets, we have 610,109 retweets.

**User Network Construction.** As Twitter does not provide the creation time of follow links, we had to infer the links with timestamp using tweets. That is, we created a follow link from user $u$ to user $v$ when $u$ mentions "@$v$" at least $k$ time in $u$'s tweets. The timestamp of the follow link is thus assigned the timestamp of the $k$th tweet of u mentioning "@$v$". In our experiments, we set $k = 3$.

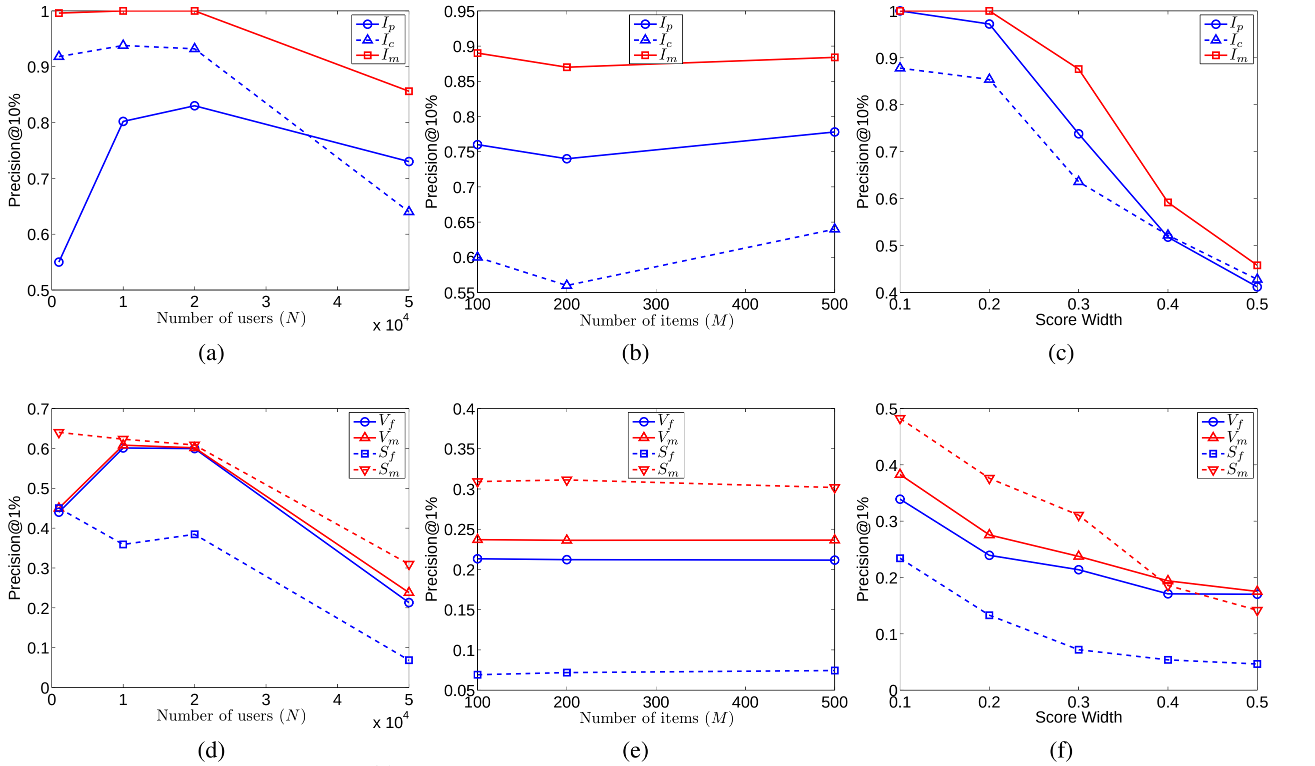
**Item Adoption** We use hashtags as items. There have been works suggesting hashtags as the topics of information diffusion in Twitter (Zhou *et. al* (2010), Romero *et. al* (2011)). In this experiment, we consider a user adopts a hashtag when she publishes a tweet containing the hashtag.

**Hashtag and User Selection.** To ensure that we have sufficient observations for each hashtag and each user, we applied the following steps to select hashtags, target users for virality $V$, and target users for susceptibility $S$.

- We selected the set of 1000 most popular hashtags
- We selected into $V$ all users adopting at least $min_a$ hashtags in 1000 selected hashtags.
- We selected into $S$ all users having at least $min_i$ selected hashtags introduced to them from users in $V$.

In our experiment, we set $min_a = min_i = 3$. This gives us $|V| = 12,978$ and $|S| = 11,069$.

**Setting the Threshold $\tau$.** The threshold $\tau$ is determined based on the time lag between retweets and their original tweets. We found that the time lag follows a long tail distribution with more than 95% of retweets having timelag within 1 day, and the maximum time lag is 205 days. We therefore set $\tau = 1$ day.



(a) (b) (c)

(d) (e) (f)

Figure 1: (a), (b), (c): Precision@10% of item virality by varying $N$, $M$ and score width respectively; (d), (e), and (f): Precision@1% of user virality and susceptibility by varying $N$, $M$ and score width respectively

## Results

**Correlation between Different Measures**. We now examine how the models rank users/ items differently. Table 2 shows that the rank correlation coefficient between the $I_p$ and $I_c$ is not high indicating that the popular hashtags are not always well diffused among the hashtag adopters. $I_m$ on the other hand is more correlated with both $I_p$ and $I_c$. $I_m$ produces rankings much more similar to $I_c$ than $I_p$, which is expected as our proposed model tends to give higher ranks to well diffused items.

Table 2: Pearson rank correlation of different item virality measures.

| | $I_p$ | $I_c$ | $I_m$ |
|---|---|---|---|
| $I_p$ | - | 0.087 | 0.145 |
| $I_c$ | - | - | 0.733 |

Similarly, we computed the Pearson rank correlations between $V_f$ and $V_m$, and between $S_f$ and $S_m$. However, since whenever $V_f(u)$ ($S_f(u)$) equals to 0, $V_m(u)$ ($S_m(u)$) equals to 0, we exclude all such users $u$ from the correlation computation. The correlation coefficient between $V_f$ and $V_m$ (respectively between $S_f$ and $S_m$) is 0.87 (respectively 0.97), which indicates that $V_m$ and $S_m$ are similar but not identical to $V_f$ and $S_f$.

**Comparison of the Top-10 Viral Hashtags.** As shown in Table 3, the top-10 viral hashtags by different models are quite different. The top-10 by $I_p$ include hashtags related to some big events (e.g., *#sgelections* and *#sgpresidents* for the two elections in Singapore in 2011), or people daily life (e.g., *#nowplaying* for what music people listen to). The top-10 by $I_c$ include mainly hashtags about funny stories and emotion (e.g., *#daveq* and *#overheard*), and those popularized by a single user (e.g., *fakemoe* or *davelimkopi*). As we expected, the top-10 by $I_m$ includes more socially and politically related hashtags (e.g., *#crappymediacorptitles* for social problems that are described by phrases similar to the names of some famous song, movie, novels, etc), and for the Singapore's 2011 Presidential Election held on August 27th, 2011 (e.g., *#asksgpresident*).

We further examine *#sgelections*, *#daveq*, and *#crappymediacorptitles*, the three hashtags top ranked for item virality by $I_p$, $I_c$, and $I_m$ respectively . For each of hashtag, we computed the number of adopters who the hashtag is diffused to, and the number of users diffusing the hashtag. As shown in Table 4, *#sgelections* has many more adopters than *#daveq* and *#crappymediacorptitles*. However, less than 50% of them adopted the hashtag due to diffusion; and less than 25% of them could diffuse the hashtag to a small number of followers. This indicates that *#sgelections* is mostly adopted due to some external factors. *#daveq* also has about 50% of the adopters adopting the hashtag due to diffusion. However, only a few of them could diffuse the hashtag. Futhermore, we found that the diffusion of *#daveq* was mostly contributed by a viral user (*fakemoe*). In contrast, more than 75% users adopting *#crappymediacorptitles* adopted the hashtag due to diffusion, and about 25% of them could diffuse the hashtag. Moreover, we also found that the diffusion of *#crappymediacorptitles* was evenly contributed by users diffusing the hashtag. It is thus reasonable to conclude that *#crappymediacorptitles* should be more viral than *#sgelections* and *#daveq*.

Table 3: Top 10 viral hashtags rank by different measures.

| **Rank** | $I_p$ | $I_c$ | $I_m$ |
|---|---|---|---|
| 1 | #sgelections | #daveq | #crappymediacorptitles |
| 2 | #nowplaying | #everysingaporeandream | #studyinginsingaporeislike |
| 3 | #sosingaporean | #ccquotes | #asksgpresident |
| 4 | #sgpresident | #overheard | #jobsforgeorgeyeo |
| 5 | #fb | #teammilo | #improvefilmtitlesbyaddinginmypants |
| 6 | #damnitstrue | #mooncakefestival | #replacesongnameswithcurry |
| 7 | #1 | #thinkaboutit | #wordspeoplebutcher |
| 8 | #justsaying | #sgreans | #chinavssgp |
| 9 | #prayforjapan | #kiasu | #yosgpresident |
| 10 | #fail | #whyifollowsosingaporean | #notsosingaporean |

Table 4: Comparison among *#sgelections*, *#daveq*, and *#crappymediacorptitles*

| **Hashtag** | **#sgelections** | **#daveq** | **#crappymediacorptitles** |
|---|---|---|---|
| $I_p$ | 6354 | 223 | 426 |
| #Users adopting due to diffusion | 2939 | 110 | 333 |
| #Users diffusing the hashtag | 1391 | 5 | 90 |
| $I_c$ | 2.11 | 22 | 3.7 |
| $I_m$ | 0.060 | 0.062 | 0.095 |

**Comparison of the Top 10 Viral Users.** The top 10 viral users by $V_f$ and $V_m$ are identical but not their ranks. They are mainly the social media accounts, portals, bloggers, and fake users.

The two users, *leticiabongnino* and *todayonline*, have significantly different ranks assigned by the two models. *leticiabongnino* is ranked 9th and 7th by $V_f$ and $V_m$ respectively, while *todayonline* is ranked 5th and 9th respectively. Although *todayonline* diffused all hashtags it had adopted and has a higher fan-out than *leticiabongnino*, the former could diffuse only a few hashtags to many followers. These are viral hashtag related to big social events. On the other hand, *leticiabongnino* could diffuse almost all hashtags she had adopted to a large number of followers. Many hashtags that *leticiabongnino* diffused were her own gossip and funny stories that are less likely to be adopted by others. The fact that she could diffuse them shows that she has high virality. Therefore, it is reasonable to assign *leticiabongnino* a virality score higher than *todayonline*.

**Comparison of the Top 10 Susceptible Users**. The top 10 susceptible users by $S_f$ and $S_m$ have 6 common users, and their ranks are different. Most of users in the two top-10 are teenages and young adults. Among them are *andyheas79* and *b2utyfulmiley*[3], the two users who have significantly different ranks by the two models. *andyheas79* is ranked 3th and 15th by $S_f$ and $S_m$ respectively, while *b2utyfulmiley* is ranked 10th and 8th respectively. We found that the hashtags that *andyheas79* adopted due to diffusion are viral, and were diffused to him by viral users. On the other hand, the hashtags diffused to *b2utyfulmiley* are not very viral, and they came from non-viral users. Therefore, although *andyheas79* has a higher fan-in than *b2utyfulmiley*, it is reasonable to assign *b2utyfulmiley* a higher susceptibility rank.

**Summary of Results**. Based on the above empirical results on the real dataset, we conclude that the different models produce results that follow our expectation. The mutual dependency model is shown to be more robust as it considers the inter-relationships of all three user and item factors.

[3] This user changed her username to *nanaphew*

## Retweet Likelihood Order Prediction for Hashtags

In this section, we hypothesize that tweets containing the higher virality hashtags are more likely to be retweeted. We therefore use the virality scores to predict, between a pair of hashtags, which one will have higher retweet likelihood in the near future. To evaluate our prediction model, we conducted the following experiment using the same Singapore-based Twitter dataset, and the same user network constructed from the dataset as described in the previous section.

We divided all tweets into weekly sets based on their published dates. For each week between May and September 2011, we used all tweets pulished within two weeks before the week as the training set, and used all the tweets pulished winthin the week as the test set. We did not examine the first 8 weeks (March and April 2011) as the tweets during this period is mainly used for user network construction. We selected 1000 most popular hashtags in the training set. Virality scores of these hashtags were computed based on diffusion information extracted from the training set. Then, we identified every tuple $(u, v, h_1, h_2)$ of two users, $u$ and $v$, and two hashtags, $h_1$ and $h_2$, that satisfies the following conditions: (a) $v$ follows $u$; (b) $h_1$ and $h_2$ are in the set of 1000 most popular hashtags in the training set (and therefore they had be assigned virality scores); and (c) $u$ posts original tweets using both $h_1$ and $h_2$ after $v$ follows $u$. For each such tuple, we computed the likelihood $l(u, v, h_1)$ (respectively $l(u, v, h_2)$) that $v$ retweets a tweet containing only $h_1$ (respectively $h_2$) that appears in the test set, and is originally posted by $u$ after $v$ follows $u$. If $h_1$ is more viral than $h_2$ (as measured by a certain model) and $l(u, v, h_1) > l(u, v, h_2)$, we say that the tuple $(u, v, h_1, h_2)$ supports the prediction model. Obviously, the virality model that gives higher fraction of supporting tuples is better.

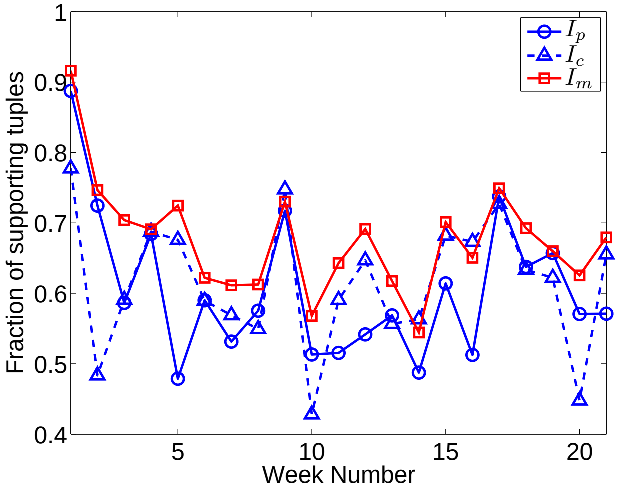


Figure 2: Fraction of tuples of users and hashtags supporting the prediction model for every week from May to September 2011

Figure 2 shows the fraction of tuples of users and hashtags that support the prediction model for every week from May to September 2011. The average fraction of tuples supporting the prediction model of $I_p$, $I_c$, and $I_m$ is 0.6, 0.61, and 0.67 respectively. The fractions of all the virality models have three common peaks at week 1, 9, and 17. This is expected as tweets are maily about big events during these weeks (the general election, the most potential predential candidates announced their candidacy, and the presidental election respectively). The prediction model based on $I_m$ achieves the highest fraction, and also has more stable performance with the faction exceeding 0.6 for almost all the weeks. This shows that our proposed model outperforms other models based on popularity and viral coefficient in this prediction task.

## Conclusions and Future work

In this paper, we propose a novel framework to model viral diffusion related user and item behaviorial factors. Considering the network effect of users and items interacting with one another in viral diffusion, we develop a mutual dependency model to measure user virality, user susceptibility and item virality simultaneously. We also propose the algorithm for implementing the model. To evaluate our proposed and other models, we have conducted extensive experiments on both synthetic and real datasets. The experiment results on synthetic datasets have shown that our proposed mutual dependency model generally outperforms the other existing ones. The results on a Twitter dataset have also shown that the mutual dependency model can better approportionate the contributions to viral diffusion by the different user and item factors properly. In the future work, we would like to further examine the detailed factors behind user virality, user susceptibility and item virality. The convergence of our proposed algorithm will also be studied. Finally, we would like to apply the virality models to important tasks such as event detection and sentiment analysis.

## 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

Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *KDD '08*.

Anderson, R. M., and May, R. M. 1992. *Infectious Diseases of Humans: Dynamics and Control.* Oxford University Press.

Bailey, N. T. J. 1975. *The mathematical theory of infectious diseases and its applications. 2nd edition.* Griffin.

Bass, F. M. 1969. A new product growth for model consumer durables. *Management Science* 15(5):pp. 215–227.

Broxton, T.; Interian, Y.; Vaver, J.; and Wattenhofer, M. 2010. Catching a viral video. *ICDM 2010 Workshops* 0:296–304.

Chung, F., and Lu, L. 2002. The average distances in random graphs with given expected degrees. *Internet Mathematics* 1:15879–15882.

Cowan, R. 2004. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control* 28(8).

Dabeer, O.; Mehendale, P.; Karnik, A.; and Saroop, A. 2011. Timing tweets to increase effectiveness of information campaigns. In *5th ICWSM*.

Dave, K. S.; Bhatt, R.; and Varma, V. 2011. Modelling action cascades in social networks. In *5th ICWSM*.

Guerini, M.; Strapparava, C.; and Ozbal, G. 2011. Exploring text virality in social networks. In *5th ICWSM*.

Hoang, T.-A.; Lim, E.-P.; Achananuparp, P.; Jiang, J.; and Zhu, F. 2011. On modeling virality of twitter content. In *13th ICADL*.

Ienco, D.; Bonchi, F.; and Castillo, C. 2010. The meme ranking problem: Maximizing microblogging virality. In *SIASP 2010 Workshop at ICDM*.

Janghyuk, L.; Jong-Ho, L.; and Dongwon, L. 2009. Impacts of Tie Characteristics on Online Viral Diffusion. *Communications of the Association for Information Systems* 24(1).

Jurvetson, S. 2000. From the ground floor: What exactly is viral marketing? *Red Herring Communications* 110.

Kiecker, P., and Cowles, D. 2001. Interpersonal communication and personal influence on the internet: A framework for examining online word-of-mouth. *Internet Applications in Euromarketing* 11(2):71–88.

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

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM Trans. Web* 1.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Memetracking and the dynamics of the news cycle. In *KDD '09*.

Li, Y.-M.; Lin, C.-H.; and Lai, C.-Y. 2010. Identifying influential reviewers for word-of-mouth marketing. *Electronic Commerce Research and Applications* 9(4):294 – 304.

Penenberg, A. L. 2009. *Viral loop : from Facebook to Twitter, how today's smartest businesses grow themselves.* Hyperion.

Petrovi, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! predicting message propagation in twitter. In *5th ICWSM*.

Ratkiewicz, J.; Fortunato, S.; Flammini, A.; Menczer, F.; and Vespignani, A. 2010. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.* 105(15):158701.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and tracking political abuse in social media. In *5th ICWSM*.

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

Ross, S. M. 2006. *Introduction to Probability Models, Ninth Edition.* Orlando, FL, USA: Academic Press, Inc.

Rowe, M. 2011. Forecasting audience increase on youtube. In *Workshop on User Profile Data on the Social Semantic Web, 8th Extended Semantic Web Conference (ESWC)*.

Shamma, D. A.; Yew, J.; Kennedy, L.; and Churchill, E. F. 2011. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *5th ICWSM*.

Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Commun. ACM* 53:80–88.

Turk, T., and Trkman, P. 2012. Bass model estimates for broadband diffusion in european countries. *Technological Forecasting and Social Change* 79(1):85 – 96.

Valente, T. W. 1995. *Network models of the diffusion of innovations.* Hampton Press (NJ).

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM '10*.

Zeidler, E. 1995. *Applied functional analysis: applications to mathematical physics.* Springer-Verlag.

Zhou, Z.; Bandari, R.; Kong, J.; Qian, H.; and Roychowdhury, V. 2010. Information resonance on twitter: watching iran. In *The First Workshop on Social Media Analytics*, SOMA '10.