Vote-and-Comment: Modeling the Coevolution of User Interactions in Social Voting Web Sites

Alceu Ferraz Costa
University of São Paulo
alceufc@icmc.usp.br

Agma Juci Machado Traina
University of São Paulo
agma@icmc.usp.br

Caetano Traina Jr.
University of São Paulo
caetano@icmc.usp.br

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

***Abstract*—In social voting Web sites, how do the user actions – up-votes, down-votes and comments – evolve over time? Are there relationships between votes and comments? What is normal and what is suspicious? These are the questions we focus on. We analyzed over 20,000 submissions corresponding to more than *100 million* user interactions from three social voting Web sites: Reddit, Imgur and Digg. Our first contribution is two discoveries: (i) the number of comments grows as a power-law on the number of votes and (ii) the time between a submission creation and a user's reaction obeys a log-logistic distribution. Based on these patterns, we propose VNC (VOTE-AND-COMMENT), a parsimonious but accurate and scalable model that models the coevolution of user activities. In our experiments on real data, VNC outperformed state-of-the-art baselines on accuracy. Additionally, we illustrate VNC usefulness for forecasting and outlier detection.**

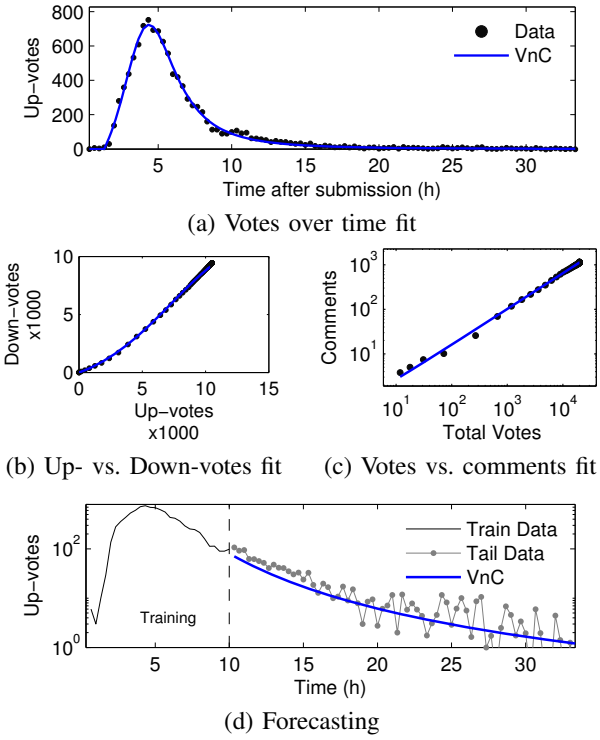
I. INTRODUCTION

Social voting Web sites [1]–[4] allow users to submit and rate content by voting and commenting on them. Voting is done using either up-votes to indicate that a submission is well-accepted or using down-votes to express disapproval.

There have been much effort on studying how the popularity of news [5], [6], memes [7]–[10] or multi-media items [11]–[13] rise and fall as they propagate over a network of users, such as bloggers. In this paper we are interested in a similar but different problem: given a social voting Web site submission, can we explain how the number of up-votes, down-votes and comments evolves over time?

To answer this question we analyzed data from three popular social voting Web sites: Reddit and Imgur, which are among the most visited Web sites in the US[1] as well as historical data from Digg [14]. Reddit, in particular, which reports more than 200 million unique visitors per month[2], has attracted the attention of the scientific community [2], [3], [15]. Based on our analysis, we propose VOTE-AND-COMMENT (VNC), a model that can describe how the up-votes, down-votes and comments change over time after a submission is created.

Figure 1 shows the up-vote, down-vote and comment data from a Reddit submission. VNC, indicated by a solid line, accurately fits three curves: up-votes time-series (Figure 1(a)), number of up-votes vs. down-votes (Figure 1(b)) and number of comments vs. number of votes (Figure 1(c)). Additionally, Figure 1(d) shows that VNC can forecast the tail part of a time-series using *only* data from the first 10 hours after the submission was created. The contributions of this paper are the following:



(a) Votes over time fit

(b) Up- vs. Down-votes fit

(c) Votes vs. comments fit

(d) Forecasting

Fig. 1. Fitting accuracy and usefulness of VNC: our proposed model accurately fits the three curves from a Reddit submission: (a) up-votes over time, (b) up-votes vs. down-votes, (c) votes vs. comments and (d) is able to forecast the number of votes received by a submission given the initial part of the votes time-series.

- **Patterns:** We show that user interactions (up-votes, down-votes and comments) in social voting Web sites are characterized by three patterns: (i) the count of comments received by a submission over time grows as a power-law on the count of votes; (ii) for popular submissions, the time for it to reach the peak of its popularity decays with respect to the number of votes received; (iii) user reaction times – i.e. the time interval between a submission creation and a user interaction – can be accurately modeled by a log-logistic distribution.

[1]Source: www.alexa.com, accessed on February 2016.
[2]Source: www.reddit.com/about, accessed on February 2016.

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.15

- **Model:** Based on the discovered patterns, we propose the VNC model and evaluate it using data from several *diverse* real datasets with over 20,000 submissions. In our experiments VNC was consistently more accurate than state-of-the-art baselines. Additionally, VNC is also able to model the relationship between votes and comments.
- **Usefulness:** We show that VNC can be used to forecast up-vote, down-vote and comment time-series. Our model provided a more accurate prediction of the final number of votes when compared to competitors. We also used VNC to spot outliers and cluster different types of content (pictures vs. discussions) based on user activity.

In order to allow *reproducibility* of our experiments, we will make available our VNC code and the datasets used in the experiments[3]. The outline of the paper starts with the background, and goes on the discoveries, model description, experiments, applications and conclusions.

TABLE I
SYMBOLS AND DEFINITIONS.

Symbol	Definition
$v_+(t)$, $v_-(t)$, $c(t)$	Up-votes, down-votes and comments received at time-tick t.
$V_+(t), V_-(t), C(t)$	Cummulative up-votes, down-votes and comments.
P_{L}	Probability of liking a submission.
P_{R}	Probability of interacting with a submission.
N	Population of potential voters.
β and ξ	VNC cascading and independent coefficients.
μ and s	Log-logistic scale and shape parameters.
α	Comments vs. votes power law exponent.
t_p	Time-series peak time.
V_{total}	Final number of up-votes.

II. DEFINITIONS AND RELATED WORK

In this section, we discuss relevant studies on social voting Web sites and on modeling user activity.

A. Definitions

We define a *submission* as any item, such as, an image or an URL to a news article that a user can upload to a social voting Web site. A *user interaction* is the act of up-voting, down-voting or commenting on a submission. For each submission, we have three time-series: $V_+(t)$, $V_-(t)$, $C(t)$ which denote, respectively, the number of up-votes, down-votes and comments accumulated by a submission over time. The derivatives of the time-series $V_+(t)$, $V_-(t)$, $C(t)$ are denoted by corresponding lower-case letters: $v_+(t)$, $v_-(t)$, $c(t)$. Table I lists the symbols and definitions used in this paper.

[3]Available at: https://github.com/alceufc/vnc_model

B. Social Voting Web Sites

In social voting Web sites, users can use up-votes and down-votes to express whether they like or not other users' submissions. Popular submissions often attract the attention of even more users generating information cascades [1].

Examples of social voting Web sites studied by the scientific community include: Slashdot [16], Digg [1], [17]–[19] and Reddit [2], [3], [15]. While Web sites such as Reddit, Imgur or Slashdot allow users to up-vote or down-vote a submission, other Web sites such as Digg or Hacker News only present the up-vote option. While there are previous works on Digg that have used comment data [2], the current version of the Web site does not allow commenting.

Many authors have studied social voting Web sites to predict whether a submission will be successful based either on its early popularity [1], [4], [16], [17], comment data [20] or content [15], [21]. In this paper, instead, we focus on the problem of modeling the coevolution of different types of user interactions over time.

C. Information Diffusion

Many authors have studied how information or memes spread over a social network of users such as bloggers or friends [12], [22]–[27]. Classical approaches for modeling information diffusion include the Susceptible-Infected (SI) [28] and Bass [29] models. Two recent models for information diffusion are Spike-M [5] and Phoenix-R [30]. The VNC model that we propose in this paper is different from these contributions as it can describe the coevolution of user interactions over time. That is, VNC models the relationship between up-votes and down-votes and between comments and votes. Table II compares the capabilities of the models discussed in this Section.

TABLE II
PROPOSED VNC MODEL AND RELATED WORK CAPABILITIES.

	SI	Bass	Ph.-R	Sp.-M	**VNC**
Votes over time	✓	✓	✓	✓	✓
Up- vs. Down-votes					✓
Votes vs. Comments					✓
Heavy Tail				✓	✓
Linear Evaluation	✓	✓	✓		✓

III. USER ACTIVITY: OBSERVATIONS

In this Section, we analyze data that we have collected from more than 300,000 submissions to Reddit and Imgur and report our findings.

A. Datasets

We tracked Reddit and Imgur submissions over a period of 20 and 34 days respectively. Every ten minutes our crawler added to its tracking list the first 150 submissions appearing on Reddit news sections and the first 50 submissions appearing on Imgur news section. We used Reddit's and Imgur's API to collect every 20 minutes a snapshot of each submission data: the number of up-votes, down-votes and comments. Each

submission was tracked for 33:20h, starting from within 10 minutes from its submission. Our crawler stopped tracking a submission after 33:20h due to the APIs limitation on the number of requests per hour and because the submissions usually did not receive a significant number of votes and comments after 33:20h of their creation. Finally, we added to our datasets all submissions with at least 100 up-votes. The selected submissions attracted over 97% of the votes and comments. Table III summarizes the number of submissions and user interactions (votes and comments) for each dataset.

TABLE III
SUMMARY OF THE DATASETS.

Dataset	# Submissions	# User Interactions (Votes and Comments)
Reddit	17,205	113,331,266
Imgur	724	2,107,576
Digg	3,553	5,149,170

For each submission we used the collected data to generate time-series for the number of up-votes, down-votes and comments. Each time-series contains 100 data-points, where each data point (time-tick) corresponds to an interval of 20 minutes. We also analyze publicly available historical Digg data[4] that was originally studied in [14]. The Digg data consists of 3,553 up-votes time-series and does not contain down-votes and comment data. However, this data was employed for comparison with previous approaches from literature.

B. Main Findings

Comments and Votes Relationship: We start the analysis of social voting Web sites' data by testing whether the number of votes and comments accumulated by a submission is correlated. We selected all submissions with at least 10 comments from the Reddit and Imgur datasets. Using the selected submissions, we evaluated two hypotheses to describe how the number of comments $C(t)$ grows with respect to the number of votes $V_+(t)+V_-(t)$ accumulated by a submission:

- **H1 - Linear Relationship:** The number of comments $C(t)$ grows linearly with respect to the number of votes $V_+(t)+V_-(t)$, that is: $C(t) = a\,[V_+(t)+V_-(t)]$;
- **H2 - Power-law Relationship:** The number of comments $C(t)$ grows as a power-law on the number of votes $V_+(t)+V_-(t)$, that is: $C(t) = k\,[V_+(t)+V_-(t)]^{\alpha}$.

We fitted the two hypothesis to data from all submissions and computed the coefficient of determination R^2. Accurately fitted results are indicated by R^2 values closer to 1, while $R^2 = 1$ corresponds to a perfect fit. Table IV shows the mean R^2 obtained by each model. The power-law model obtained statistically significantly larger R^2 values than the linear model for both datasets with $p = 0.01$. Based on this result, we propose the Vote-Comment (VC) Law:

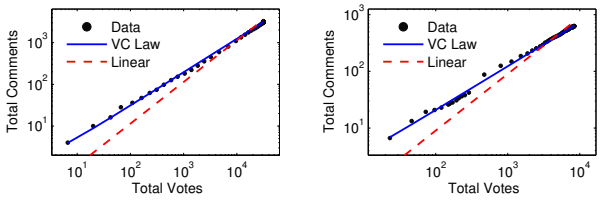
Observation 1: **VC Law**: The relationship between the number of votes and comments received by a submission can be accurately described by a power-law.

[4]http://www.isi.edu/~lerman/downloads/digg2009.html

TABLE IV
R^2 OBTAINED BY FITTING REAL DATA FROM REDDIT AND IMGUR.

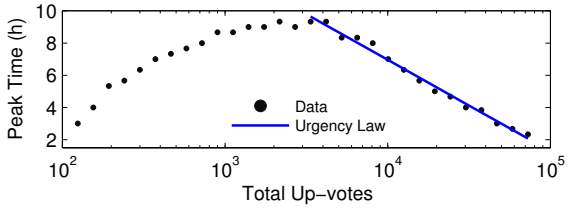
Model	Reddit	Imgur
Power-Law Relationship	**0.95 ± 0.07**	**0.93 ± 0.10**
Linear Relationship	0.87 ± 0.40	0.82 ± 0.27

In Figure 2(a) we plot the comments vs. votes trajectory for two selected submissions from the Reddit and Imgur datasets. The VC Law is indicated by the solid blue line and accurately captures the power law relationship in the growth rate of votes and comments. The α parameter, which captures the slope of the comments and votes' trajectory in a log-log scale plot is also indicated. For comparison, we also included the fit obtained by the linear relationship hypothesis, indicated by a red dashed line. The linear relationship fails to match the slope of the comments vs. votes' trajectory.



Reddit Link ($\alpha = 0.67$) Imgur Link ($\alpha = 0.77$)

(a) VC Law



(b) Urgency Law

Fig. 2. Discoveries: (a) The number of comments follows a power-law on the number of votes (VC Law). (b) Submissions with a larger number of up-votes reach their peak faster (Urgency Law).

Peak Time and Total Votes Relationship: When analyzing the interaction of the users to the Web site, the two following questions come to mind: how quickly does the popularity of a submission reaches a peak? A submission that receives a large number of up-votes will peak faster or slower than a less popular submission? To answer these questions, we analyzed the up-votes time-series of all Reddit, Imgur and Digg submissions. For each submission, we computed the peak time t_p as the time it takes for the time-series to reach its maximum value.

Figure 2(b) shows the median peak times for submissions with a different number of total up-votes. We built the plot by grouping submissions into logarithmic buckets based on the number of total up-votes. Next, for each bucket, we computed the median peak times. Figure 2(b) shows that the final number of up-votes V_{total} of a submission is related to its peak time. We name this relationship "Urgency Law":

Observation 2: **Urgency Law**: the peak time t_p of the up-votes time-series of popular submissions (with at least 3,000 up-votes) decays approximately linearly with respect to the logarithm of the total number of up-votes: $t_p \approx b - a \log V_{total}$.

Figure 2(b) shows that popular submissions reach the peak of their popularity faster. The fit obtained by our proposed Urgency Law is indicated by a solid blue line which matches the real data. The values obtained for the coefficients a and b were $a = 2.46$ and $b = 29.5$.

Reaction Times: After a submission is created, how long does it take for a user to cast a vote or make a comment? We start by defining *reaction times*:

Definition 1 (Reaction Time): The reaction time Δ_R of a given user interaction corresponds to the time interval between the instant t_o in which a submission is created and the instant t_r in which an interaction (vote or comment) occurs.

Figure 3 shows the distribution of reaction times of up-votes, down-votes and comments for all submissions from the Reddit, Imgur and Digg datasets. The solid red line corresponds to the log-logistic fit to the data whose PDF is given by:

$$f(x) = \frac{(s/\mu)(x/\mu)^{s-1}}{\left(1 + (x/\mu)^s\right)^2} \quad (1)$$

where μ and s denote the scale and shape parameters, respectively. We estimated the parameters μ and s of the log-logistic distribution using maximum likelihood estimation (MLE).

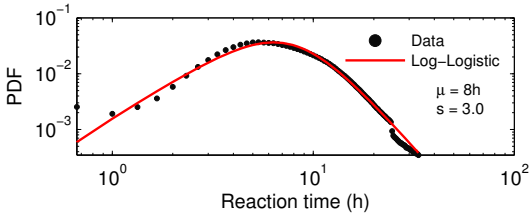


Fig. 3. Reaction times can be described by a log-logistic distribution. The parameters μ and s were estimated using MLE and the log-logistic fit is indicated by the solid red line.

Previous works have assumed that reaction times in social networks follow power-law distributions [5], [31]. While power-laws are able to match the tail part of the reaction time distribution, they do not explain the distribution for small reaction times. Figure 3 shows that the log-logistic is a more robust model as it can be used to model the distribution of small reaction times:

Observation 3: The distribution of reaction times for votes and comments in social voting Web sites can be accurately modeled by the log-logistic distribution.

IV. Proposed Method: VNC

Is it possible to find a model to describe the number of up-votes, down-votes and comments received by a submission that matches Observations 1 to 3? In this Section, we show how that it is possible by proposing VNC, a model that accurately describes how the volume of up-votes, down-votes and comments of a submission changes over time. VNC is composed of three sub-models that describe the following relationships:

1) Number of up-votes over time;
2) Up-votes vs. down-votes;
3) Comments vs. votes.

First, in Section IV-A, we describe how VNC models the number of votes received by a submission over time. In Section IV-B we describe how the number of up-votes and down-votes evolves after a submission is created. Finally, in Section IV-C, we show that the number of comments can be explained as a power law of the number of votes.

A. Modeling Up-Votes Over Time

VNC assumes that the probability $P(t)$ of a user up-voting a submission at time-tick t is given by the probability $P_{\text{L}}(t)$ that the user will like the submission with the condition that the user has already observed the submission. Denoting the probability that the user will observe (i.e. react) to the submission at the time tick t as $P_{\text{R}}(t)$ and assuming that $P_{\text{R}}(t)$ and $P_{\text{L}}(t)$ are independent, we have:

$$P(t) = P_{\text{L}}(t) \cdot P_{\text{R}}(t) \quad (2)$$

The probability $P_{\text{L}}(t)$ that a user will like a submission in VNC is governed by a *cascading mechanism*. That is, $P_{\text{L}}(t)$ increases as the submission accumulates up-votes. We model this cascading mechanism as a linear fraction of the users that have already voted on that submission:

$$P_{\text{L}}(t; \beta_+, \xi_+) = \xi_+ + \beta_+ \cdot \frac{V_+(t)}{N_+}, \quad (3)$$

where ξ_+ is the independent coefficient, β_+ is the cascading coefficient, N_+ is the population of potential voters and $V_+(t) = \sum_{t=1}^{t} v_+(t)$ is the number of votes accumulated by the submission at time-tick t.

The cascading coefficient β_+ reflects how strong is the influence exerted by users on each other. Larger values of β_+ mean that users are more likely to up-vote a submission based on the number of up-votes accumulated by that submission.

Based on Observation 3, we model the distribution P_{R} of reaction times using the log-logistic distribution with PDF given by Equation 1 and scale and shape parameters denoted by μ and s. The shape parameter μ corresponds to median reaction time while the shape parameter is related to the decay rate of P_{R}.

Since users on social voting Web sites can vote only once on a submission, the number of users that can cast a vote on a submission is given by the difference $N_+ - V_+(t)$. As s result, the number of up-votes $v_+(t+1)$ received by a submission at time-tick $t+1$ is given by the product $P(t) \cdot [N_+ - V_+(t)]$:

$$\boxed{v_+(t+1) = [N_+ - V_+(t)] \cdot P_{\text{L}}(t; \beta_+, \xi_+) \cdot P_{\text{R}}(t; \mu, s)} \quad (4)$$

where the initial conditions are $v_+(0) = 0$ and $V_+(0) = 0$.

Equation 4 assumes that a submission starts receiving votes as soon as it is created at time-tick $t = 1$. However, in certain social voting Web sites a user can create a submission at time

$t = 1$ but only share it later at time $t = t_s$. In this case, the probability of a user voting on a submission is $P(t) = 0$ if $t \leq t_s$ and $P(t) = P_{\mathrm{L}}(t; \beta_+, \xi_+) \cdot P_{\mathrm{R}}(t - t_s; \mu, s)$ for $t > t_s$.

Parameter Estimation: Given a time-series $v_+(t)$ of real data describing the number of up-votes received by a submission and the estimated values $\hat{v_+}(t; \theta)$, we learn the parameters $\theta = \{N,\ \beta_+,\ \xi_+,\ \mu,\ s,\ t_s\}$ for our model by minimizing the sum of squared errors between $v_+(t)$ and $\hat{v_+}(t; \theta)$:

$$\min_{\theta} \sum_{t=1}^{t_m} [v_+(t) - \hat{v_+}(t; \theta)]^2 \tag{5}$$

We solve the least-squares problem from Equation 5 using the *Levenberg-Marquardt* algorithm [32].

B. Modeling Up-Votes and Down-Votes

VNC assumes that the cascading and log-logistic reaction times mechanisms, that govern the evolution of up-votes, are the same for the down-votes time-series. That is, users are more likely to down-vote a submission as it accumulates down-votes (Equation 3). and the probability of a submission receiving a down-vote is modulated by the users' reaction time distribution. This results in the following equation for the down-votes time-series:

$$\boxed{v_-(t+1) = [N_- - V_-(t)] \cdot P_{\mathrm{L}}(t; \beta_-, \xi_-) \cdot P_{\mathrm{R}}(t; \mu, s)} \tag{6}$$

with initial conditions $v_-(0) = 0$ and $V_-(0) = 0$.

The parameters μ and s that control the reaction times distribution as well as the start time t_s are shared between the up-votes (Equation 4) and down-votes time-series. The cascading coefficient β_-, independent coefficient ξ_- and the population of potential down-voters N_- are allowed to be different for the time-series $v_+(t)$ and $v_-(t)$.

We fit the parameters $\theta = \{N_+,\ \beta_+,\ \xi_+,\ N_-,\ \beta_-,\ \xi_-,\ \mu,\ s,\ t_s\}$ by solving the following least-squares problem:

$$\begin{aligned} \min_{\theta} \sum_{t=1}^{t_m} & [v_+(t) - \hat{v_+}(t; N_+, \beta_+, \xi_+, \mu, s, t_s)]^2 + \\ & + [v_-(t) - \hat{v_-}(t; N_-, \beta_-, \xi_-, \mu, s, t_s)]^2 \end{aligned} \tag{7}$$

where $v_+(t)$ and $v_-(t)$ denote the real data and $\hat{v_+}(t)$ and $\hat{v_-}(t)$ are the estimated values for the up-votes and down-votes time-series.

C. Modeling comments

Based on the VC Law exhibited by real data (Observation 1), we model the number of comments $C(t)$ accumulated by a submission as a power-law on the number of votes $V_T(t)$:

$$\boxed{C(t) = k[V_T(t)]^{\alpha}} \tag{8}$$

where $V_T(t) = V_+(t) + V_-(t)$ is the sum of votes (up- and down-votes) accumulated by a submission at time-tick t. Given real data from the time-series of accumulated votes $V_T(t)$ and comments $C(t)$, we find the parameters k and α by minimizing the squared error between $C(t)$ and $k[V_T(t)]^{\alpha}$.

Definition 2: The Equations 4, 6 and 8 constitute our discrete time VNC model.

V. Experiments

Votes Time-Series: In this Section, we evaluate how well VNC fits real data from Reddit, Imgur and Digg. The Reddit and Imgur datasets were obtained by our crawlers and are described in Section III-A. The Digg dataset is publicly available[5] and was originally studied in [14]. The time interval (resolution) between two data points for all time-series is 20 minutes.

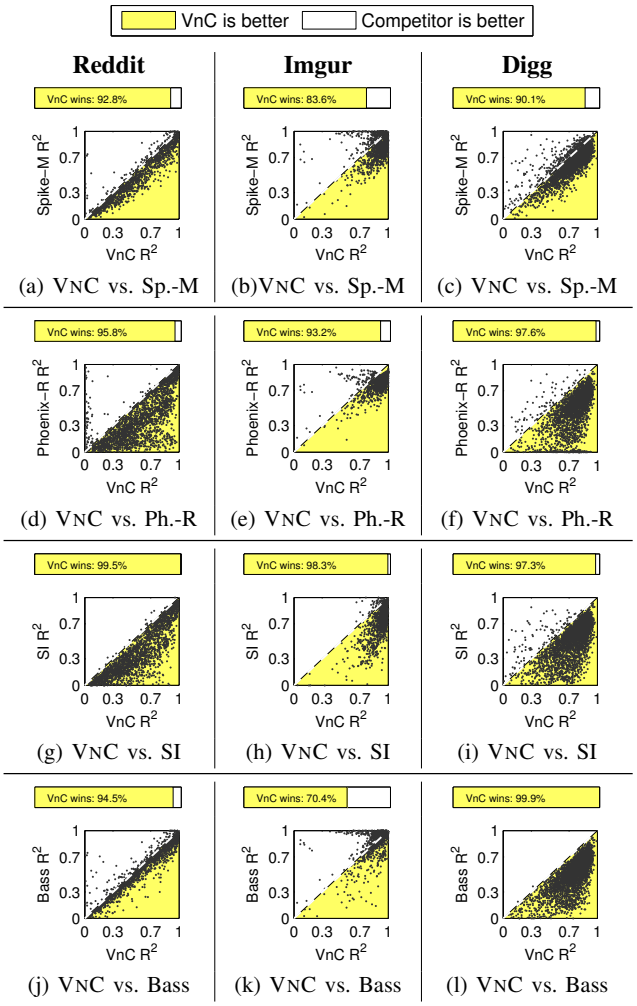


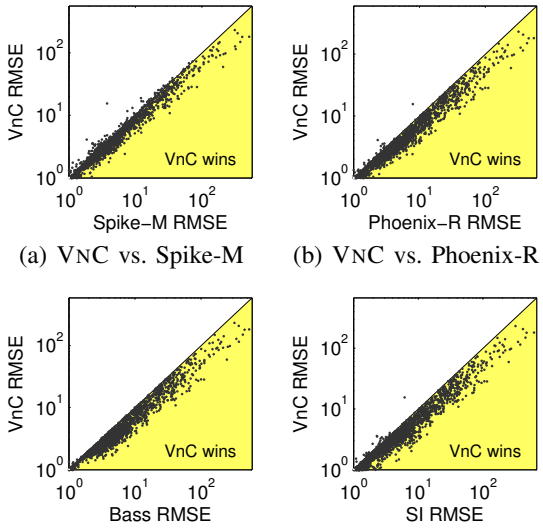
Fig. 4. VNC is more accurate than competitors. Points below the diagonal in the yellow area correspond to up-votes time-series that were best fit by VNC. The bars above each plot show the fraction of time-series that were best fit by VNC. All time-series with at least 10 up-votes were selected.

We start by comparing the VNC accuracy against the SI, Bass, Phoenix-R and Spike-M models. SI and Bass are widely employed in the literature to model information cascades while Phoenix-R and Spike-M are state-of-the-art models. Figure 4

[5]http://www.isi.edu/~lerman/downloads/digg2009.html

shows the scatter plots of the R^2 values obtained by VNC and a competitor. Accurately fitted results are indicated by R^2 values closer to 1, while $R^2 = 1$ corresponds to a perfect fit. Each point corresponds to an up-votes time-series. In each scatter plot, time-series below the diagonal, in the yellow area, correspond to time-series that were best fitted by VNC. VNC obtained best results than its competitors for all datasets. More specifically, VNC provided a better fit than the SI, Bass, Phoenix-R and Spike-M models for over 98%, 90%, 95% and 88% of the time-series, respectively. This is indicated by a larger number of time-series (black points) below the diagonal of the scatter plots of Figure 4.

Figure 5 compares the root mean squared error (RMSE) obtained by VNC, SI, Bass, Phoenix-R and Spike-M models when fitting up-votes time-series from Reddit, Imgur and Digg. A lower RMSE indicates a better fit, and points below the diagonal in the yellow area correspond to time-series that were best fit by VNC. VNC obtained a smaller RMSE for over 91%, 99%, 96% and 90% of the time-series when compared to the Bass, SI, Phoenix-R and Spike-M models, respectively.



(a) VNC vs. Spike-M (b) VNC vs. Phoenix-R

(b) VNC vs. Bass (c) VNC vs. SI

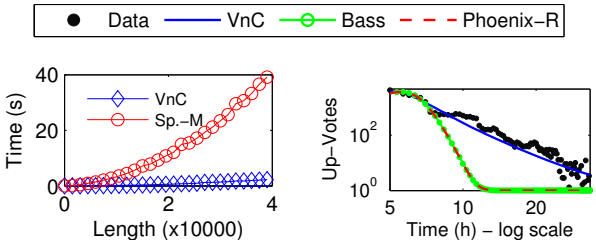
Fig. 5. VNC has the lowest fitting error (RMSE). Points below the diagonal in the yellow area correspond to time-series from Reddit, Imgur and Digg that were best fit by VNC.

Table V compares the mean R^2 values obtained by VNC, SI, Bass, Phoenix-R and Spike-M models for all up-vote time-series from the Reddit, Digg and Imgur datasets. For all datasets VNC obtained a statistically significantly higher R^2 than its competitors with $p = 0.01$.

TABLE V
FITTING ACCURACY: R^2 FOR UP-VOTES TIME-SERIES.

Model	Reddit	Imgur	Digg
VNC	**0.80 ± 0.21**	**0.88 ± 0.13**	**0.77 ± 0.16**
Spike-M	0.76 ± 0.23	0.66 ± 0.18	0.67 ± 0.15
Phoenix-R	0.65 ± 0.30	0.55 ± 0.23	0.58 ± 0.22
Bass Model	0.75 ± 0.21	0.78 ± 0.15	0.55 ± 0.17
SI	0.69 ± 0.26	0.51 ± 0.19	0.31 ± 0.39

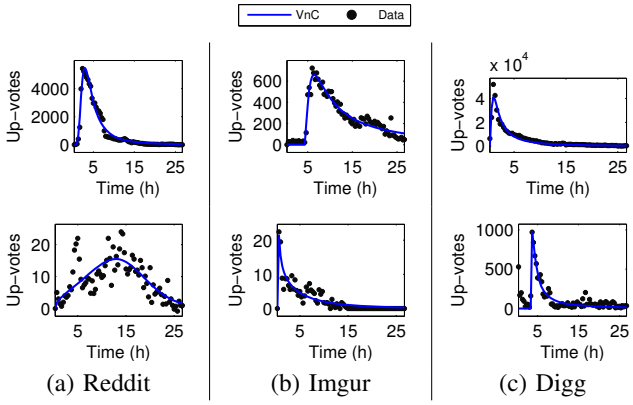
When compared to Spike-M, VNC also has the advantage of being *scalable* on the length (number of time-ticks) of the time-series. That is, VNC requires $O(n)$ operations to evaluate a time-series, while Spike-M has $O(n^2)$ runtime on the time-series length. This is shown in Figure 6(a), were we plot the time in seconds required to evaluate time-series of different length for VNC and Spike-M. Evaluation time was obtained in a computer with an Intel Core i5 3.2GHz processor, 16 GB of RAM running Mac OS X 10.9. Both methods were implemented in Matlab.



(a) VNC Scalability (b) Heavy Tail Matching

Fig. 6. Advantages of the VNC model. (a) Evaluation of VNC is linear on the length of time-series while Spike-M requires $O(n^2)$ operations. (b) VNC matches the heavy-tailed decay in the up-votes time-series while the Bass and Phoenix-R models have an unrealistic exponential decay.

Figure 7 shows the VNC fit for up-vote time-series. Due to space limitations, we show the fit for the submissions with the largest (top row) and median (bottom row) number of up-votes on each dataset. Although different submissions may have different rise and fall patterns, VNC, indicated by a solid blue line, accurately fits real data from different social voting Web sites

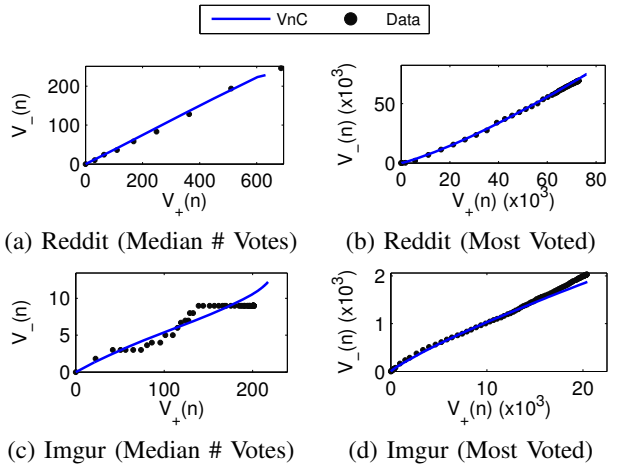


(a) Reddit (b) Imgur (c) Digg

Fig. 7. VNC fits the up-vote time-series from different social voting Web sites: (a) Reddit, (b) Imgur and (c) Digg. Submissions on the top and bottom row received, respectively, the largest and median number of up-votes on each dataset.

Heavy Tail Decay: Most of the up-votes time-series from Reddit, Imgur and Digg have a heavy-tailed decay as we show in the log-log scale time-series plots from Figure 6(b). While VNC model is able to match the heavy tail decay in the up-votes time-series, the Bass and Phoenix-R models generate a non-realistic exponential decay in the number of votes received over time. We do not show the curve for the SI model in Figure 6(b) to avoid occlusion. However, the SI model also shows an exponential decay in the tail part of the up-votes time-series.

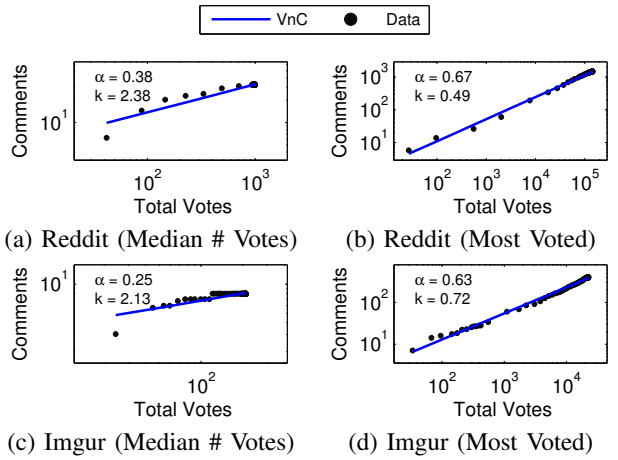
The Spike-M model is also able to generate heavy-tailed decays. However, as we show in Figure 6(a), VNC has the advantage of being scalable while Spike-M has a $O(n^2)$ run-time complexity on the length of the time-series.



(a) Reddit (Median # Votes) (b) Reddit (Most Voted)

(c) Imgur (Median # Votes) (d) Imgur (Most Voted)

Fig. 8. VNC (blue line) fits the relationship up-votes and down-votes. Submissions to the left and to the right received, respectively, the median and the largest number of up-votes on each dataset.



(a) Reddit (Median # Votes) (b) Reddit (Most Voted)

(c) Imgur (Median # Votes) (d) Imgur (Most Voted)

Fig. 9. VNC (blue line) fits the relationship votes and comments. Submissions to the left and to the right received, respectively, the median and the largest number of up-votes on each dataset.

Up-votes, Down-votes and Comments: In Figures 8 and 9 we analyze how well VNC fits the relationship between up-votes, down-votes and comments. Figure 8 shows the trajectory in the up-votes vs. down-votes plane of the most voted submission and the submission with the median number of votes in each dataset. Similarly, Figure 9 shows the trajectory of the same submissions in the comments vs. total votes (up-votes plus down-votes) plane. For both cases, VNC (indicated by a solid blue line) accurately fits the data. Additionally, VNC captures the difference in the growth rate of votes and comments. Figures 8 and 9 only show two submissions from each dataset due to space limitations, since the other submissions were also accurately fitted by VNC.

VI. VNC AT WORK

In this Section we show that VNC can be used for forecasting, detecting outliers and clustering submissions.

A. Forecasting

In this Section, we want to solve the problem of forecasting the tail part of the up-vote, down-vote and comment time-series of a given submission. The problem can be stated as follows:

Problem 1: Given the initial parts of the up-vote $v_+(t)$, down-vote $v_-(t)$ and comment time-series $c(t)$ of a submission for $t < t'$, predict the tail part of the time-series for $t > t' + 1$.

In order to forecast the tail part of the time-series we learn the VNC parameters using only data from the first six hours and 40 minutes after a submission is created (20 time-ticks).

We start by using VNC and the learned parameters to predict the final number of up-votes, down-votes and comments received by a submission. Table VI shows the median absolute percentage error (APE) given by $|A - F|/A$, where A denotes the actual final number of up-votes, down-votes or comments and F the forecasted values. The best results for each combination of dataset and time-series (including statistical ties) are indicated in bold. We used a Wilcoxon rank-sum test at 5% significance level to compare the medians.

VNC provided the most accurate forecast – or tied with the best competitor – for all combinations of time-series and datasets. The exception was the up-vote time-series from the Digg dataset for which Spike-M obtained the lowest error. VNC obtained the most striking results for the comments time-series, where the median APE (error) was considerably lower than all competitors. The volume of comments received by a submission is, on average, lower than the number of votes. As a result, forecasting the comment time-series is a challenging task. VNC solves this problem by using data from the up-vote and down-vote time-series to forecast the comment time-series (see Equation 8).

TABLE VI
COMPARISON OF VNC FORECASTING ERROR. THE TABLE SHOWS THE MEDIAN FORECASTING ABSOLUTE PERCENTAGE ERROR (APE). BEST RESULTS (INCLUDING STATISTICAL TIES) ARE SHOWN IN BOLD.

		Bass	SI	Phoenix-R	Spike-M	**VnC**
Reddit	v_+	0.57	0.57	0.82	0.42	**0.39**
	v_-	0.67	0.64	0.86	**0.55**	**0.53**
	c	0.62	0.64	0.86	0.73	**0.39**
Imgur	v_+	**0.69**	**0.65**	0.81	0.98	**0.71**
	v_-	**0.65**	0.68	0.84	0.98	**0.65**
	c	0.59	0.58	0.84	1.15	**0.47**
Digg	v_+	0.87	0.89	0.98	**0.52**	0.77

Figure 10 shows the VNC forecast results for the up-vote time-series of the most voted submissions from each dataset. We also include, for comparison purposes, the results obtained by Spike-M and Bass models. We omit forecasts from SI and Phoenix-R to avoid occlusion and because the Spike-M and Bass models obtained the best results after VNC. Only the

data points to the left of the vertical dashed line (indicated by a solid black line) were used to train the models. Both Spike-M and VNC obtained accurate forecasts while the Bass model did not predict correctly the tail, which is expected, since it yields a curve that decays exponentially while real data has a heavier tail. While Spike-M was able to match VNC forecast capability, VNC is scalable and Spike-M has a quadratic run-time complexity on the length of the time-series (see Figure 6).

B. Detecting Outliers

We have shown that VNC provides very accurate fits for real data from social voting Web sites. Now we show that VNC can also be used to detect outliers. We divide that outliers that we observed in the Reddit and Imgur datasets into three categories: (i) late-night outliers, (ii) flat down-vote outliers and (iii) "one-shot users" outliers.

Late night outliers: Figure 11(a) shows a scatter plot that compares the R^2 values obtained by VNC when fitting the up-votes and down-votes time-series of all Reddit and Imgur submissions. As expected from the results of Section V, VNC *obtained high R^2 values for most of the submissions*. However, submissions inside the dashed blue ellipsis are outliers, as they have low R^2 values for both the up-votes and down-votes time-series and represent less than 2% of the dataset. Figure 11(c) shows two of these outliers, which are characterized by having double peaked time-series. For comparison, the normal submissions shown in Figure 11(b), as well as the majority of the time-series, have a single peak.

In order to understand what causes the double peak behavior, we analyzed the time of the day of occurrence of the valleys between the two peaks. We find the time-series' two peaks by computing all local maxima in a 5.5h window and selecting the two local maxima with the largest up-vote count. Our analysis of the 35 submissions with the lowest R^2 values showed that the valleys are concentrated around 3h to 10h in US Eastern Standard Time (EST), coinciding with time interval in which Reddit users are less active. This indicates that the two peak behavior may be caused by submissions that receive a first peak of up-votes by users active at late-night and a second peak of votes by users active by the morning.

Flat down-vote outliers: Figure 11(c) shows the up-votes and down-votes time-series for submissions that are located in the red ellipsis area in Figure 11(a). These submissions represent less than 2% of the dataset. Compared to the normal submissions, the outliers located in the red ellipsis area have a flat down-vote time-series and a regular (single peaked) up-vote time-series. More than half of the flat down-vote outliers are pictures of animals such as dogs and cats, which seems to repel down-votes from Reddit and Imgur users.

"One shot users" outliers: Finally, VNC can also detect anomalous outliers by analyzing the fit accuracy of the votes vs. comments' curve (Equation 8). We selected all submissions with at least 100 comments from the Reddit and Imgur datasets. We computed VNC's R^2 values for the votes vs. comments fit. Figure 12(a) shows the distribution of R^2 values with the outliers indicated by a red ellipsis. Figure 12(c) shows the time-series for the votes (up-votes and down-votes) and comments received by the two suspicious submissions with the smallest R^2 values. Both submissions received a large number of comments just after they were created. Notice that this pattern does not occur for regular submissions such as in Figure 12(b).

We inspected each suspicious submission in order to find comments made by "one-shot users". We define "one-shot users" as users who only commented on a single submission. For the left submission in Figure 12(c), over 55% of the users who posted comments are suspicious: i.e. they only posted comments on that submission. Similarly, for the right submission of Figure 12(c), over 10% of the users who posted comments are suspicious. For comparison, we also inspected the submissions with the highest and median R^2 values, and we were not able to find a single one-shot user.

Both suspicious submissions were posted to the Reddit's AMA (ask me anything) section, which focus on interviews. In the AMA section, a user (the interviewee) makes a submission and other users (commenters) can leave questions (comments) and vote on other questions and answers. The AMA section is often used as a marketing tool and has attracted the participation of celebrities such as Bill Gates, Barack Obama and Arnold Schwarzenegger. As a result, the outliers detected by VNC could be attributed by individuals that recruited users to boost the number of comments – and consequently, the popularity – of their submission.

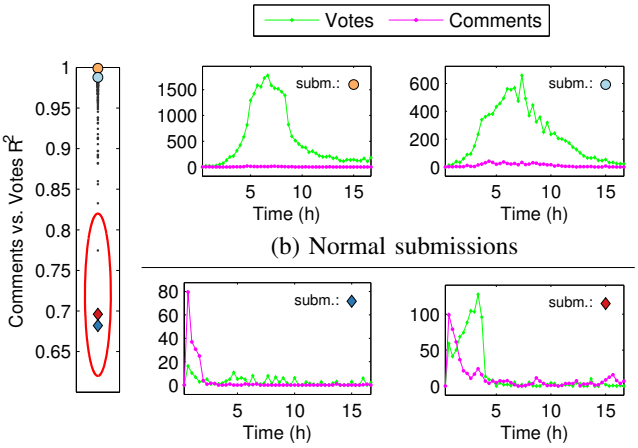
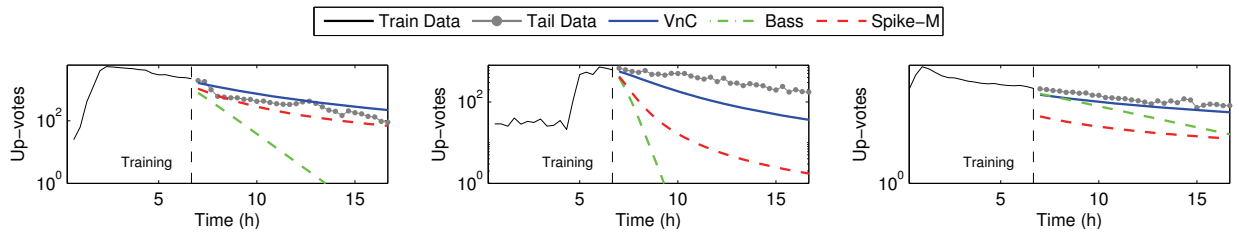


Fig. 12. VNC spotted submissions suspected of using artificial users to boost their popularity. (a) Distribution of R^2 values for the comments vs. up-votes fit. (b) Normal submissions. (c) Suspicious submissions which received a large number of comments just after their creation.

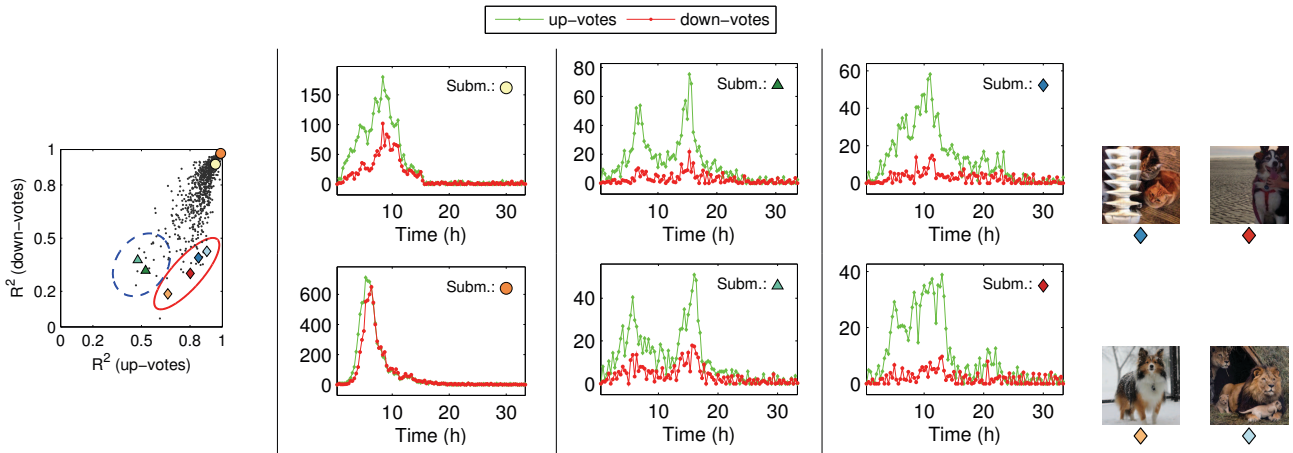
C. Clustering

While Imgur is focused only on image data, Reddit allows users to submit links to different types of content (e.g. pictures, news, movies). A question one could ask is whether the comment vs. votes behavior is the same for distinct types of links. To answer this question we make use of the fact that Reddit links are categorized into *sub-reddits*. Sub-reddits are



(a) Reddit (b) Imgur (c) Digg

Fig. 10. Forecasting using VNC. We trained the models using *only* data from the first seven hours after a submission is created. The vertical axis (up-votes) is in log scale. The interval between data points is 20 minutes.



(a) Up- vs. Down-votes R^2 (b) Normal submissions (c) Late night outliers (Blue Ellipsis) (d) Flat down-vote outliers (Red Ellipsis)

Fig. 11. Outliers detected by VNC. (a) VNC R^2 values obtained by fitting the up-votes and down-votes time-series of submissions with at least 1,000 up-votes. (b) Normal submissions for which VNC provided an accurate fit (over 95% of the dataset). (c) Late night outliers (dashed blue ellipsis, less than 2% of the dataset) have a low R^2 for the up-votes and down-votes time-series. (d) Flat down-vote outliers (solid red ellipsis, less than 2% of the dataset) have a high R^2 for the up-votes time-series but low R^2 for the down-votes time-series. More than half of the flat down-votes outliers are pictures of animals.

sub-communities focused on different interests. We used VNC to fit the votes vs. comments curve of links submitted to two sub-reddits: *pics* and *AskReddit*. The sub-reddit *pics* allows users only to post links to pictures while in *AskReddit* users submit questions asking for other users' comments. In Figure 13 we plot the parameters k and α of VNC Equation 8.

Observation 4: Links to pictures and to discussions generate two separated clusters when projected into the parameter space α and k of the VNC model (Figure 13).

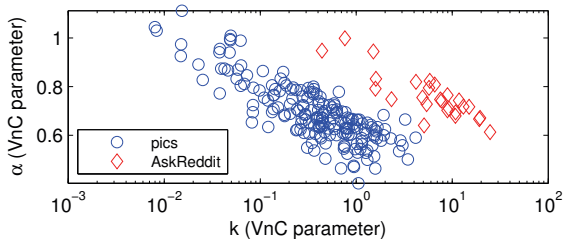


Fig. 13. Clustering submissions using the VNC model parameters. Submissions were selected from two sub-reddits: *pictures* and *AskReddit*.

Figure 13 shows that there is a clear difference in the parameter k between links submitted to the sub-reddits pics and AskReddit, resulting in a separation of the submissions into two clusters. This indicates that the coefficient k changes significantly for different types of submissions (pictures and questions). However, the difference in the exponent α is not as pronounced. While submissions to the pics sub-reddit have a mean $\alpha = 0.68 \pm 0.12$, submissions to AskReddit have a mean $\alpha = 0.75 \pm 0.09$.

VII. Conclusions

In this paper we proposed VNC, a model for user activity on social voting Web sites. Given a submission, VNC describes: (i) how the number of votes evolves over time; (ii) how the difference between up-votes and down-votes evolves over time and (iii) how the number of comments grows with respect to the number of votes. In order to validate VNC we have tracked more than 20,000 submissions from three social voting Web sites: Reddit and Imgur, which are among the 13 most visited Web sites in the US, each attracting over 110 million unique visitors per month as well as publicly available data from Digg [14].

We compared VNC against representative models for information diffusion from the literature. VNC consistently

provided the most accurate fit and forecast to real data for all analyzed datasets. For the comment time-series, in particular, VNC performed significantly better than the baselines (Table VI).

The contributions of this paper are summarized as follows:

- **Discoveries:** We discovered the following patterns by analyzing different modalities of user interactions: (i) relationship between peak time and total number of votes (*Urgency Law*), (ii) relationship between votes and comments (VC *Law*) and (iii) log-logistic distribution of reaction times.
- **Model**: Based on the discovered patterns, we proposed VNC, a parsimonious, accurate and scalable model that describes the relationship between different types of activity on social voting Web sites. VNC consistently obtained superior results with respect to fitting accuracy when compared to competitors. More specifically, VNC provided a better fit than the closest competitor for over 92% of the more than 20,000 submissions studied. Moreover, VNC is also able to explain the relationship between up-votes vs. down-votes and votes vs. comments, while existing models such as Spike-M and SI can only be used to fit the up-votes, down-votes and comment time-series individually.
- **Usefulness**: VNC can be used to forecast the number of votes received by a submission, spot anomalies and cluster different types of content based on user activity.

In order to allow *reproducibility* of our experiments, we make our VNC code and the datasets used in the experiments available at https://github.com/alceufc/vnc_model.

ACKNOWLEDGMENT

This material is based upon work supported by FAPESP, CNPq, CAPES, the RESCUER project funded by the European Commission (Grant: 614154) and by the CNPq/MCTI (Grant: 490084/2013-3), the National Science Foundation under Grants No. CNS-1314632, IIS-1408924, by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, and by ARO/DARPA under Contract Number W911NF-11-C-0088.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] K. Lerman and T. Hogg, "Using a Model of Social Dynamics to Predict Popularity of News," in *WWW*, 2010, pp. 621–630.

[2] C. Wang, M. Ye, and B. A. Huberman, "From user comments to on-line conversations," in *KDD*. ACM Press, 2012, pp. 244–252.

[3] T. Weninger, X. Zhu, and J. Han, "An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community," in *ASONAM*, 2013, pp. 579–583.

[4] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity," in *KDD*, 2015, pp. 1513–1522.

[5] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, "Rise and fall patterns of information diffusion: model and implications," in *KDD*, 2012, pp. 6–14.

[6] A. Anagnostopoulos, A. Bessi, G. Caldarelli, M. Del Vicario, F. Petroni, A. Scala, F. Zollo, and W. Quattrociocchi, "Viral Misinformation: The Role of Homophily and Polarization," *arXiv:1411.2893*, pp. 1–12, 2014.

[7] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *KDD*. ACM Press, 2009, pp. 497–505.

[8] B. A. Prakash, D. Chakrabarti, N. C. Valler, M. Faloutsos, and C. Faloutsos, "Threshold conditions for arbitrary cascade models on arbitrary networks," *KAIS*, vol. 33, no. 3, pp. 549–575, 2012.

[9] C. Bauckhage, K. Kersting, and F. Hadiji, "Mathematical Models of Fads Explain the Temporal Dynamics of Internet Memes," in *ICWSM*, 2013, pp. 22–30.

[10] Y. Matsubara, Y. Sakurai, and C. Faloutsos, "The Web as a Jungle: Non-Linear Dynamical Systems for Co-evolving Online Activities," in *WWW*, 2015, pp. 721–731.

[11] G. Korkmaz, C. J. Kuhlman, A. Marathe, M. V. Marathe, and F. Vega-Redondo, "Collective Action Through Common Knowledge Using a Facebook Model," in *AAMAS*, 2014, pp. 253–260.

[12] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system." *PNAS*, vol. 105, no. 41, pp. 15 649–15 653, 2008.

[13] P. A. Dow, L. A. Adamic, and A. Friggeri, "The Anatomy of Large Facebook Cascades," in *ICWSM*, 2013, pp. 145–154.

[14] K. Lerman and R. Ghosh, "Information Contagion: an Empirical Study of Spread of News on Digg and Twitter Social Networks," in *ICWSM*, 2010, pp. 90–97.

[15] H. Lakkaraju, J. McAuley, and J. Leskovec, "What's in a name? Understanding the interplay between titles, content, and communities in social media," in *ICWSM*, 2013, pp. 1–10.

[16] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The Slashdot Zoo: Mining a Social Network with Negative Edges," in *WWW*, 2009, pp. 741–750.

[17] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commu. of the ACM*, pp. 80–88, 2010.

[18] C. Budak, D. Agrawal, and A. El Abbadi, "Diffusion of Information in Social Networks: Is It All Local?" in *ICDM*, 2012, pp. 121–130.

[19] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," *KAIS*, vol. 37, no. 3, pp. 555–584, 2013.

[20] J. G. Lee, S. Moon, and K. Salamatian, "Modeling and predicting the popularity of online contents with Cox proportional hazard regression model," *Neurocomputing*, vol. 76, no. 1, pp. 134–145, 2012.

[21] A. Khosla, A. Das Sarma, and R. Hamid, "What Makes an Image Popular?" in *WWW*, 2014, pp. 867–876.

[22] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading Behavior in Large Blog Graphs," in *SDM*, 2007, pp. 551–556.

[23] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, "Finding effectors in social networks," in *KDD*, 2010, pp. 1059–1068.

[24] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, "Influence Maximization in Dynamic Social Networks," in *ICDM*, 2013, pp. 1313–1318.

[25] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini, "The role of information diffusion in the evolution of social networks," in *KDD*, 2013, pp. 356–364.

[26] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can Cascades be Predicted?" in *WWW*, 2014, pp. 925–936.

[27] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec, and M. Tiwari, "Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily," in *WWW*, 2015, pp. 66–76.

[28] R. M. Anderson and R. M. May, *Infectious Diseases of Humans*. Oxford University Press, 1991.

[29] F. Bass, "A new product growth for model consumer durables," *Management Science*, vol. 15, no. 5, pp. 215–227, 1969.

[30] F. Figueiredo, J. M. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos, "Revisit Behavior in Social Media: The Phoenix-R Model and Discoveries," in *PKDD*, 2014, pp. 386–401.

[31] A. L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[32] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *SIAM*, vol. 11, no. 2, pp. 431–441, 1963.