Predicting the Success of Online Petitions Leveraging Multidimensional Time-Series

Julia Proskurnia
École Polytechnique Fédérale de Lausanne
Lausanne—Switzerland
iuliia.proskurnia@epfl.ch

Przemyslaw A Grabowicz
MPI-SWS
Germany
pms@mpi-sws.org

Ryota Kobayashi
National Institute of Informatics
Tokyo—Japan
r-koba@nii.ac.jp

Carlos Castillo
Eurecat Technology Center of Catalunya
Barcelona—Spain
chato@acm.org

Philippe Cudré-Mauroux
University of Fribourg
Fribourg—Switzerland
pcm@unifr.ch

Karl Aberer
École Polytechnique Fédérale de Lausanne
Lausanne—Switzerland
karl.aberer@epfl.ch

ABSTRACT

Applying classical time-series analysis techniques to online content is challenging, as web data tends to have data quality issues and is often incomplete, noisy, or poorly aligned. In this paper, we tackle the problem of predicting the evolution of a time series of user activity on the web in a manner that is both accurate and interpretable, using related time series to produce a more accurate prediction. We test our methods in the context of predicting signatures for online petitions using data from thousands of petitions posted on *The Petition Site*—one of the largest platforms of its kind. We observe that the success of these petitions is driven by a number of factors, including promotion through social media channels and on the front page of the petitions platform. We propose an interpretable model that incorporates seasonality, aging effects, self-excitation, and external effects. The interpretability of the model is important for understanding the elements that drives the activity of an online content. We show through an extensive empirical evaluation that our model is significantly better at predicting the outcome of a petition than state-of-the-art techniques.

Keywords

Web applications; Online petitions; time series prediction.

1. INTRODUCTION

The ability to predict user activity or engagement on the web has many applications in a wide range of domains. This includes predicting the number of people who will install an application in an app marketplace, buy a product from an online retailer, or participate in an e-government action etc. Ideally, a forecast of user involvement should be generated as *early* as possible, in a manner that is both *accurate* and *interpretable*. The quest for interpretability is due to the importance of knowing what are the elements that are driving predictions up or down as a process unfolds, in order to take corrective actions whenever possible. The problem of generating early, accurate, and interpretable predictions on the web challenges our understanding of complex interactions over time, and is further complicated by the presence of multiple confounders. Web data is almost invariably noisy and incomplete, and often comes from several heterogeneous sources. Additionally, and despite recent advances in empirical methods for predicting information dissemination [35, 39], we lack a general parametric modeling framework to predict user involvement in a *reinforced* process, for instance, a petition accompanied by an active campaign to gather signatures by mobilizing people online. For instance, the mobilization of people through an online campaign might involve several sources of reinforcement: social media, traditional news media, and word-of-mouth or viral advertising.

In this paper, we present new forecasting models for online content dissemination that are able to take into account several elements: seasonality, aging effects, self-excitation, and external influence (e.g., in the form of social media postings). Our main contribution, beyond presenting a combined parametric model that has better predictive power than the state of the art, is being able to incorporate a time series of related observations to produce a more accurate and earlier prediction, and to further enhance the interpretability of the results.

We evaluate our models by using them to predict the number of signatures an online petition will gather over time. Online petitions are a representative of a broad class of online phenomena involving active public mobilization, and thus represent a relevant scenario for testing our methods. The setting we consider might generalize to the active spread of ideas or memes, in the sense that it goes beyond passive diffusion. People promoting online petitions and people who sign petitions tend to actively encourage others to sign, instead of passively expecting that people simply learn about

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052705

these petitions through a contagion process. Often, such promoters use external platforms for dissemination; thus, it is crucial to capture the external signals.

Our contributions. In this work, we present models for user behavior with respect to online petitions. We make the following contributions:

- we analyze thousands of online petitions from one of the largest petitions sites on the web (Sections 3, 4);
- we present a model to predict user involvement in a reinforced manner combining seasonality, aging, self-excitation, and external evidence as a continuous signal; this model has easily interpretable parameters (Section 5);
- we show that our model is more accurate in both short-term and long-term predictions of user involvement, when compared with state-of-the-art methods (Section 6).

The rest of the paper is organized as follows. We start with an overview of related work in Section 2. We describe our process for collecting petition data in Section 3, and the insights we gained in Section 4. We present our new predictive model and compare it to existing models in Section 5. We experimentally evaluate the models and discuss them in Section 6. Finally, we summarize our results and outline future work in Section 7.

2. RELATED WORK

In this section, we position our paper with respect to prior work on popularity prediction for the web and for online petitions.

2.1 Popularity prediction on the web

Predicting the popularity of user generated content on the web has been studied extensively [36]. Many different settings have been considered; common content types include online videos [23], online news [4], social bookmarking sites [22], social networking services [39], and crowdfunding campaigns [9], among others. Most works on this topic tackle one of three main tasks: (i) *classify as successful/unsuccessful*, meaning trying to predict whether a particular piece of content will exceed a certain popularity threshold or not; (ii) *predict the overall popularity*, i.e., predict the final number of views or votes a piece of content will receive; and (iii) *time series forecasting,* i.e., modeling the popularity dynamics over time. Regardless of the specific task, two main types of approaches are observed: feature-based and model-based. *Feature-based* techniques rely on a set of (hand-)crafted features extracted from a single or multiple sources, for the purpose of classification or regression. *Model-based* techniques assume a specific parametric model for the process that drives the phenomenon; they are usually harder to formulate, but often produce better insight into the studied phenomenon. We summarize these approaches and include references for each one in Table 1.

This paper goes beyond analyzing "meme"-like content that spreads virally, and study a phenomenon that involves active promotion; hence, we need to consider external signals. External information is used by previous work adopting feature-based approaches that extend Szabo and Huberman [35] (such as [4]), but not in model-based methods, as we do in this work. Our approach is based on modeling the conditional mean of a Hawkes process, as Kobayashi and Lambiotte [20] suggested. However, we extend their model with a flexible aging that includes a raise and a decay, and allows both for internal dynamics (self excitation) and external factors (social media, front page). Moreover, each external factor is modeled as a continuous effect on the signature dynamics, rather than a series of individual external shocks. To the best of our knowledge, we are the first to present a model that captures interaction between multiple platforms in a model-based framework and with easily interpretable parameters.

Approach	Data source(s)	Examples
Classification		
.. Feature-based	Twitter	Hong *et al.* [17], Ma *et al.* [25], Cui *et al.* [8], Jenders *et al.* [19], Cheng *et al.* [5]
.. Social transfer	Multiple	Roy *et al.* [32]
Popularity prediction		
.. Feature-based	Online news	Castillo *et al.* [4]
.. Feature-based & regression	Twitter	Kupavskii *et al.* [21], Bao *et al.* [3], Hong *et al.* [17], He *et al.* [16]
.. Model-based	Digg, YouTube	Szabo and Huberman [35]
.. Model-based	Twitter	Zhao *et al.* [39]
.. Model-based	Earthquake, crime data	Ogata *et al.* [28], Mohler *et al.* [27]
.. Model-based	Multiple	Choi *et al.* [6]
.. Social dynamics	Digg	Lerman *et al.* [22]
Series forecasting		
.. Model-based	Twitter	Kobayashi and Lambiotte [20]
.. Model-based	Weibo, Citations	Gao *et al.* [11], Shen *et al.* [33]
.. Model-based	Multiple	Linderman *et al.* [24], Xu *et al.* [37, 38], Olshansky and Carnes [29]
.. Time series clustering	YouTube, Digg, Vimeo	Ahmed *et al.* [1]

Table 1: Selected works on popularity predictions in social media. Typical tasks in this context are to classify as successful/unsuccessful (top), to predict the overall popularity (middle), and to forecast the popularity time series (bottom).

2.2 Analyzing the dynamics of online petitions

Signature acquisition in online petitions is a complex and multi-dimensional problem. From the perspective of online activism, it is not only important to predict whether a petition will gain the required number of signatures or not, and what the final number of signatures will be, but also to start from valid assumptions about how the number of signatures evolves over time, and how external factors shape this evolution. Understanding these factors can help the organizers of these petitions to further enhance the engagement of the public with their campaigns.

Hale *et al.* [15] describe a temporal analysis of 8,000 petitions and discuss early signs of success (e.g., a large number of signatures during the first days). However, it remains unclear why some petitions become popular and others do not, or what are the factors that can lead to an increase

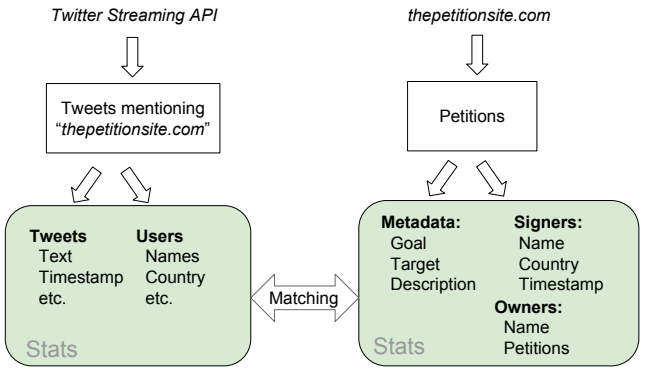

Figure 1: Collection pipeline for the petitions dataset.

in popularity. Huang *et al.* [18] analyze "power" users on petitions platforms and how user involvement changes over time on a petitions platform. Proskurnia *et al.* [30] study the effect of petition success on user involvement in public online campaigns [31, 10]. In contrast, we link social media and petitions together to model their evolution considering multiple factors, including external influence.

Online petitions can be compared to crowdfunding campaigns, as both efforts work towards obtaining a given level of support over a bounded period of time. Etter *et al.* [9] study various prediction techniques for crowdfunding campaigns on Kickstarter. An *et al.* [2] analyze investor activity on Kickstarter and make recommendations based on their activity on Twitter. Unlike these works, we focus on signature rate dynamics using co-evolving time series information, and we do not limit ourselves to signals from social media, but also utilize further available information, including the effect of being featured on the front page.

3. DATA COLLECTION

Our study is based on petitions obtained from *The Petition Site*,[1] one of the top-3 sites of its type according to Alexa.[2] *The Petition Site* allows anyone to create an online petition and to gather signatures. There are 14 categories in which petitions can be started, including Environment and Climate, Education, Health, and Human Rights. Petitions have a headline (e.g., "Help stop the Taiji dolphin slaughter"), the name of the person or entity to whom the petition is addressed (e.g., "International Marine Trainers Association"), the name of the person who creates the petition, dates of opening and closing of the signature gathering, and a description and/or letter describing the contents of the petition. Petitions also include a target number of signatures, decided by its author; we consider that petitions that reach this target are *successful*, otherwise they have *failed*.

We collect two kinds of information on those petitions: list of signatories and tweets pointing to the petitions. The entire data collection pipeline is illustrated in Figure 1. The overall characteristics of the collection are shown in Table 2.

[1] `http://thepetitionsite.com/`

[2] `http://www.alexa.com/topsites/category/Society/Activism/Petitions`

Table 2: Dataset characteristics. Each characteristic in this table shows a significant difference at $p \ll 0.001$.

	Successful	Failed
Petitions	1,219	3,505
Median signatures goal	4,319	43,838
Median signatures collected	51,986	5,687
Anonymous fraction	0.023	0.044
Fraction of signers' comments	0.031	0.045
Petitions with tweets	90%	27%
Mean number of tweets	83.3	37.1
Mean number of retweets	31.2	24.7
Mean number of unique users	62.7	26.8

Petitions data. Petitions data were obtained using a custom-made web crawler and scraper to collect petitions created after August 1st, 2016 across all the topics. The resulting petitions garnered around 85 million signatures from about 5 million unique users. While there are old petitions in the data we collected—some dating back to 2003—we decided to focus solely on petitions that started after August 1st and were active for at least 10 days. These petitions comprise 85% of the total number of signatures in the entire collection. We additionally removed five outlier petitions having unattainable goals (requiring more than 1 billion signatures).

Each petition has a web page including public information about the people who signed the petition. Each signer is authenticated on the platform by providing an e-mail address, whose ownership must be verified before the signature is recorded. Once the e-mail address is verified, signers may chose to remain anonymous (listing only the signature timestamp on the website), or to disclose more information (such as their first name and country of residence). Additionally, we collected hourly data for top 10 petitions promoted on the front page starting August 1st.

Twitter data. In addition, we used Twitter's streaming API to collect all tweets containing a link to any URL containing "thepetitionsite.com." Tweet collection was conducted from August 1st, 2016 through October 1st, 2016, collecting over 250K tweets.

4. DATA ANALYSIS

Table 2 shows that the median number of signatures collected by successful and failed petitions are significantly different ($p \ll 0.001$). We also observe that successful petitions have more modest goals than failed ones; indeed, the goals of successful petitions are 10 times smaller than the goals of failed petitions (target of 4.3K signatures in successful petitions, vs. target of 43.8K in failed ones), while the successful petitions collect about 9 times more signatures (51.9K signatures in successful petitions, vs 5.6K in failed ones). Both successful and failed petitions have similar timespans, 50 and 42 days on average, respectively.

The majority of people include their first name and country, but signatories of failed petitions are almost twice as likely to remain anonymous (2.3% anonymous signatures in successful petitions vs 4.4% in failed ones); they might be less willing to be publicly associated to these petitions. We also observe that petitions that are successful have on average more activity on Twitter: they are three times more

likely to have tweets (90% vs 27%), and have an average number of tweets that is more than twice the number of tweets failed petitions receive (83.3 vs. 37.1).

The cumulative distribution of signatures for over 4,000 petitions is shown in Figure 2 (left). From the figure, we observe that over 70% of the failed petitions did not reach 1,000 signatures, while nearly all successful petitions obtained at least 1,000 signatures and over 20% of the successful petitions reached over 100,000 signatures.

As previous works [15, 34], we observe that the higher the number of signatures a petition receives early on, the more likely it is to gain the required number of signatures. Figure 2 (right) shows the distribution of the number of signatures for the first 3 hours of a petition. Almost all failed petitions acquire less than 10 signatures during the first 3 hours, but this does not guarantee failure: almost 60% of the successful petitions also acquire less than 10 signatures during their first 3 hours. As a result, a significant part of the successful petitions are indistinguishable from failed petitions during the first hours and, thus, it is not trivial to make an accurate prediction on whether they will succeed or not using only this data. Observations done using the first 24 hours of each petition, omitted for brevity, show a similar lack of separation between successful and failed petitions.

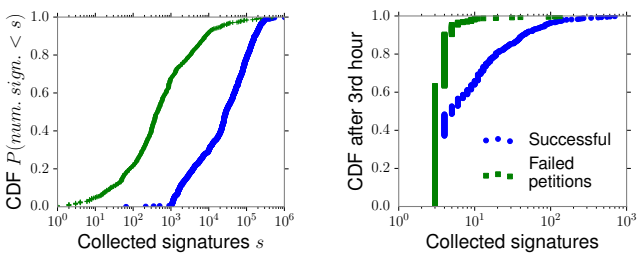

Figure 2: Cumulative Distribution Function (CDF) of the signatures collected by successful and failed petitions during their entire history (left) and during their first three hours (right).

To understand the behavior of different classes of petitions, we clustered the petitions' time series using Dynamic Time Warping [12] into four clusters (we experimented with values from 2 to 30 clusters, and found that the inter-cluster distance stabilizes at about 4 clusters). The corresponding centroids are shown in Figure 3. Each cumulative distribution function for the petition signatures has been rescaled to the unit interval and to have the same number of time bins. Again, we observe that successful petitions tend to gather a large share of their signatures early on.

4.1 Circadian Cycles and External Influence

In this section we observe two key characteristics of the time series of signatures that we subsequently use for building our prediction model.

Circadian cycles. We binned the petition signatures and corresponding tweets into 10 minute time intervals. In addition, we aligned the petition signatures and tweets with the corresponding time of the day in the users' country. Both activities clearly follow a circadian rhythm, with the signature activity showing a stronger circadian pattern than the tweets. In particular, we can observe a peak (at around 10am) in signature activity as shown in Figure 4.

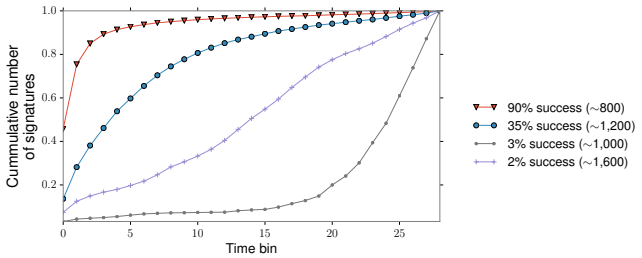

Figure 3: Average of normalized Cumulative Distribution Function (CDF) in four clusters of petitions. Clustering was performed using Dynamic Time Warping (DTW). Numbers in parenthesis represent the size of each cluster.

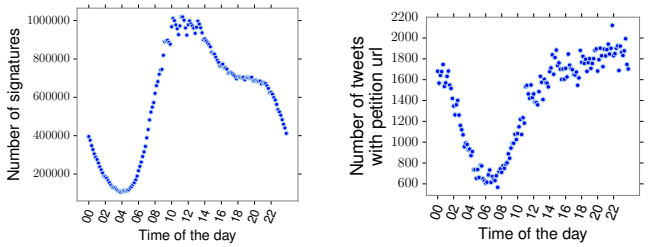

Figure 4: Daily pattern of the signature (left) and tweet activity (right) with 10 minutes time intervals. Both activities can be fitted by using a sinusoidal function: $a + b\sin(2\pi(t + t_0)/24)$.

External effects. In order to estimate whether social media and being featured on the front page affect the signatures, we performed a Granger causality [14] study between signature time series, social media and front page appearances. We examined a random sample of 30 petitions from each cluster in Figure 3 with their corresponding tweets and their presence in the front page of *The Petition Site* (as detailed in Section 5.2). Specifically, we ran the algorithm to discover the latent network structure for point processes from Linderman and Adams [24], which determines the influence of a time series on the prediction of another time series, e.g., whether signatures affect tweets or vice versa. As a result, we discovered that for the cluster containing more successful petitions, Granger causality from Twitter to the number of signatures can be observed in 90% of the cases. This fraction is lower for the remaining clusters that have less probability of success: 72%, 35%, and 20% respectively. This suggests that Twitter can accelerate the signatures early in the lifetime of a petition. We confirm this later, in Section 6.3, by showing that it mostly influences our predictive capability early in the petition lifetime. Interestingly, in the case of petitions that were promoted to the front page of *The Petition Site*, we identified cases where signatures influenced the front page time series and vice versa equally. We further study the front page effect in Section 4.3.

4.2 Matching Twitter Users and Signers

The main goal of this subsection is to establish a clearer connection between signatures and social media postings (*tweets*) beyond Granger causality. We performed a one-to-

Table 3: Characteristics of the user profiles that were unambiguously matched between *The Petition Site* and Twitter.

Fraction of petition overlap (signed and tweeted)	12%
Mean number of distinct petitions tweeted	15.34
Mean time between first signature and tweet	26 hours
Mean number of tweets per petition	16
Mean number of petitions signed	113
Mean delay between signature and tweet	19 hours
Median delay between signature and tweet	15 minutes
Fraction of users that post a tweet after signing	74%
Fraction of users that sign after posting a tweet	26%

Table 4: Comparison of petitions that were promoted to the front page (FP) against similar petitions that were not promoted (¬FP). A significant difference at $p < 0.01$ is denoted by **.

	FP	¬FP	
Petitions	75	75	
Median signatures before t_S^*	2,146	2,038	
Mean signatures before t_S^*	9,285	9,314	
Success rate	100.0%	83.5%	
Median signatures after 2 days	14,835	8,049	**
Average signatures after 2 days	24,485	16,035	**

one matching between Twitter accounts and the names of petition signers/owners. Information about signers is represented in a structured format on the petitions platform. We adapted the method by Goga *et al.* [13] with matching parameters set according to our data. In particular, we used the following attributes to match the profiles: (1) signer full name and Twitter name/user name, (2) signer location and Twitter user location, (3) signer petitions and tweeted petitions. We tried various combinations of these three matching dimensions, and found that using all of them resulted in the maximum number of unambiguously matched users.

The main idea behind the matching is to investigate user patterns while signing the petition, specifically whether people post a tweet after signing, or sign after posting a tweet. This fine-grained matching further allows us to trace the number of followers that signed the petition and retweeted it. Overall, we were able to match 3,157 accounts (out of 37K unique Twitter users). On average, each signer was matched to 1.47 Twitter accounts (with the maximum number of matches being 45); 2,641 accounts were matched one-to-one to Twitter accounts in a non-ambiguous manner; these are the ones included in Table 3. The first observation from this table is that most people who sign a petition and post a tweet first sign the petition, and then tweet. The absolute difference in minutes between user sign/tweet behavior can be depicted with the following sparkline:  , where red line correspond to the case when a petition was signed an tweeting simultaneously and on the left of the red line we have users that first tweet and then sign. About 80% of the users perform signing and tweeting almost at the same time. In particular, 74% of users that sign and tweet almost simultaneously, tweet less than 10 minutes after signing a petition. We note that no matching scheme across websites is perfect, and this particular one might have false positives (some of the signer profiles had several identical matches on Twitter), however, we believe it provides relevant insight on the interaction between these platforms.

4.3 Front Page Effect

We identified 75 petitions that were promoted to the front page, and measured whether petitions that are promoted to the front page are already on track to be successful, and if promoting those petitions causes their success. The short answer corroborates the results of the Granger causality analysis of Section 4.1: yes to both. To arrive to this answer, we used a standard tool from observational experiments, a *matching* study, where we matched these 75 petitions featured on the front page with 75 similar petitions that were not featured on the front page. First, we computed the number of signatures that each of the 75 petitions promoted to the front page obtained before it got promoted at time t_S^*. Second, we matched each petition promoted to the front page with one that is within a 10% range of the number of signatures but was not promoted (¬FP) at time t_S^*. On average petitions appear on the front page after 27 hours (79 hours median) and remain for 14 days (6 days median). Statistics of these two samples are compared in Table 4.

Table 4 strongly suggests that the petitions that are promoted are not randomly chosen. Failed petitions constitute about 75% of our sample, and hence a petition chosen uniformly at random should have about 25% success rate. In comparison, the matched ¬FP set has a success rate above 80%. However, the same observations also confirm that being promoted on the front page has a drastic effect on these petitions. Beyond ensuring success (as the success rate of promoted petitions is 100%), it significantly increases the number of signatures received. For example, after only 2 days of being promoted on the front page, petitions gained almost twice as much signatures as ¬FP.

5. PETITIONS MODELING

In this section, we introduce new methods to model the evolution of the number of signatures. Our models take into account circadian rhythms, aging effects, self-excitation, and external signals that influence the signature rate over time. Experimentally, these signals correspond to postings related to each petition on a social media platform, and the position in which a particular petition was present on the front page of the petitions site.

First, we introduce a new deterministic model that mimics the circadian nature of the underlying phenomenon we are studying and that includes information aging and self-excitation. Next, we extend this model by incorporating the external influence of social media and front page display, describing an end-to-end prediction pipeline.

5.1 Circadian Rhythm and Aging

The engagement of users with petitions, this is, the signature rate over time, exhibits two important temporal characteristics: circadian cycles and temporal decay. Circadian cycles are visible as daily oscillations in the signature rate, as we showed in Figure 4; they affect all petitions and remain stable within a particular time zone. Decay is expected due to the aging of the petition; sometimes the signature rate starts to decrease immediately, while in other cases it increases and then decreases. Based on these observations, we propose a model called Circadian rhythm with Rise and Decay (CRD). We discretize the time using a time

step $\delta t = 1(h)$, while the signature rate (number of signatures between t and $t+1$) is described as

$$\hat{s}_p(t) = \left\{ a_p + b_p \sin\left(\frac{2\pi}{T}(t+\phi_p)\right)\right\} t^{k_p} e^{-t/\tau_p}, \quad (1)$$

where t is the time since the birth of petition p, a_p is the intensity, b_p is the amplitude of the oscillation, ϕ_p its phase (with respect to an oscillation cycle of $T = 24h$), τ_p is the decay parameter, and k_p describes the initial rise in the petition activity. Parameters are fitted by minimizing the square error $E^p = \sum_{t=1}^{T_{train}} \{\hat{s}_p(t) - s_p(t)\}^2$, using Levenberg-Marquardt's algorithm [26]. The parameter range of τ_s is restricted to $0.5 < \tau_p < 75$ hours similarly to Kobayashi and Lambiotte [20].

5.2 Self-Excitation and External Influence

The CRD model is extended to incorporate self-excitation and external influence that comes from two sources. The external influence we model comes from two sources. The first one is social media, and is expressed as $n_{\text{sm}}(t)$, the number of social media exposures at time t (the number of tweets multiplied by the average number of the authors' followers). The second one is being featured on the front page of *The Petitions Site*, expressed as the rank in the front page $n_{\text{srank}}(t)$ that contains 10 petitions at a time. An arbitrary value of $n_{\text{srank}} = 1,000$ were chosen for petitions not featured in the home page, which are the majority. The signature rate

$$\hat{s}_p(t) = \left\{ a_p + b_p \sin\left(\tfrac{2\pi}{T}(t+\phi_p)\right)\right\} t^{k_p} e^{-t/\tau_p} + \sum_{i=0}^{T_{\text{mem}}} \left(c_{\text{self}}(i) s_p(t-i) + c_{\text{sm}}(i) n_{\text{sm}}(t-i) + \frac{c_{\text{front}}(i)}{n_{\text{srank}}(t-i)} \right), \quad (2)$$

where $T_{\text{mem}} = 10h$ is the size of a memory window indicating the number of time steps to be used in the estimation, and memory kernels c_{self}, c_{sm}, c_{front} are, respectively, the relative importance of self-excitation, the external influence from social media, and the impact of being featured on the front page of *The Petitions Site* over time. The memory kernels are determined by minimizing the squared error after fitting CRD parameters a_p, b_p, k_p, τ_p and ϕ_p.

6. EXPERIMENTS

In our experiments, we consider two main prediction tasks: short-term $T_{\text{tot}} = 72$ (3 days) and long-term $T_{\text{tot}} = 168$ (1 week) prediction. We vary the size of the input that is available to each model T_{train}(from 12 hours to 71 or 167 hours respectively).

6.1 Metrics

Two metrics were used for calculating prediction performance of different prediction models.

Symmetric Median Absolute Percentage Error (SMAPE) measures the median hourly deviation between the predicted and actual time series signature counts for a predicted period over N petitions:

$$\text{SMAPE} = \underset{p}{\text{median}} \frac{1}{T_{\text{tot}} - T_{\text{train}}} \sum_{t=T_{\text{train}}}^{T_{\text{tot}}} \left| \frac{\hat{s}_p(t) - s_p(t)}{\hat{s}_p(t) + s_p(t)} \right|$$

where, $\hat{s}_p(t)$ and $s_p(t)$ are the predicted and actual numbers of signatures of the p-th petition between t and t+1. We use median to reduce the effect of outliers, similarly to previous works on web predictions [20, 39].

Cumulative Symmetric Median Absolute Percentage Error (CSMAPE) measures the median deviation between the predicted and actual cumulative signature counts for a predicted period over N petitions:

$$\text{CSMAPE} = \underset{p}{\text{median}} \left| \frac{\hat{S}_p(T_{\text{train}}, T_{\text{tot}}) - S_p(T_{\text{train}}, T_{\text{tot}})}{\hat{S}_p(T_{\text{train}}, T_{\text{tot}}) + S_p(T_{\text{train}}, T_{\text{tot}})} \right|,$$

where $\hat{S}_p(T_{\text{train}}, T_{\text{tot}})$ and $S_p(T_{\text{train}}, T_{\text{tot}})$ are the predicted and actual number of signatures of the p-th petition in the prediction period $(T_{\text{train}}, T_{\text{tot}}]$, respectively.

6.2 Baselines

We compared our methods against three state-of-the-art baselines.

Linear Regression. We trained the linear regression model proposed by Szabo *et al.* [35], which is a standard method for popularity prediction. The logarithm of the cumulative number of signatures $S(T)$ at time T is fitted by a linear function $\log S(T) = \alpha_T + \log S(T_{\text{train}}) + \sigma_T \epsilon_T$. Parameter α_T, σ_T are obtained by minimizing the squared error of the prediction on a training set, and ϵ_T is a Gaussian random variable with zero mean and unit variance.

SVM with self-excitation and SVM with social media. A strong and simple baseline to predict complex time series is SVM regression with the Gaussian radial basis function (RBF) [7]. Similarly to our model, SVM with self-excitation and SVM with social media are given time series $s_p(t-i)$ and $n_{sm}(t-i)$ for a time window $T_{\text{mem}} = 10$ respectively. The best performing parameters for the model determined experimentally for our case are $C = 1000$ and $\gamma = 0.1$, where C is the soft margin penalty parameter and γ is the kernel coefficient.

Reinforced Poisson Process (RPP) The RPP model has been used for modeling the cumulative number of citations to journal papers published by the American Physical Society [33]. The signature rate λ_t is expressed as $\lambda_t = c f_\gamma(t) r_\alpha(R_t)$, where c represents the attractiveness, $f_\gamma(t) \propto t^{-\gamma} (\gamma > 0)$ describes the aging, and the reinforcement function $r_\alpha(R_t) (\alpha > 0)$ models the "rich gets richer" phenomenon. The parameters c, γ, α are determined by maximizing the likelihood function [11, 20].

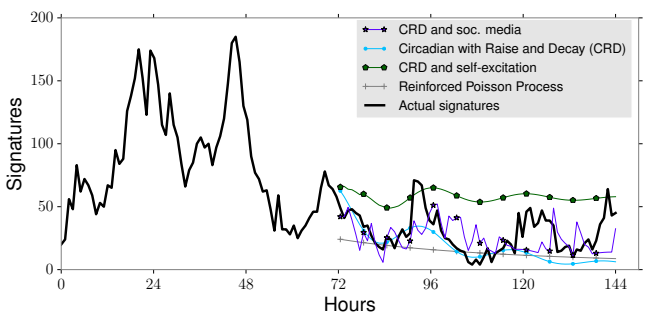

Figure 5: Example showing the prediction of the number of signatures of one petition, after 3 days of observation.

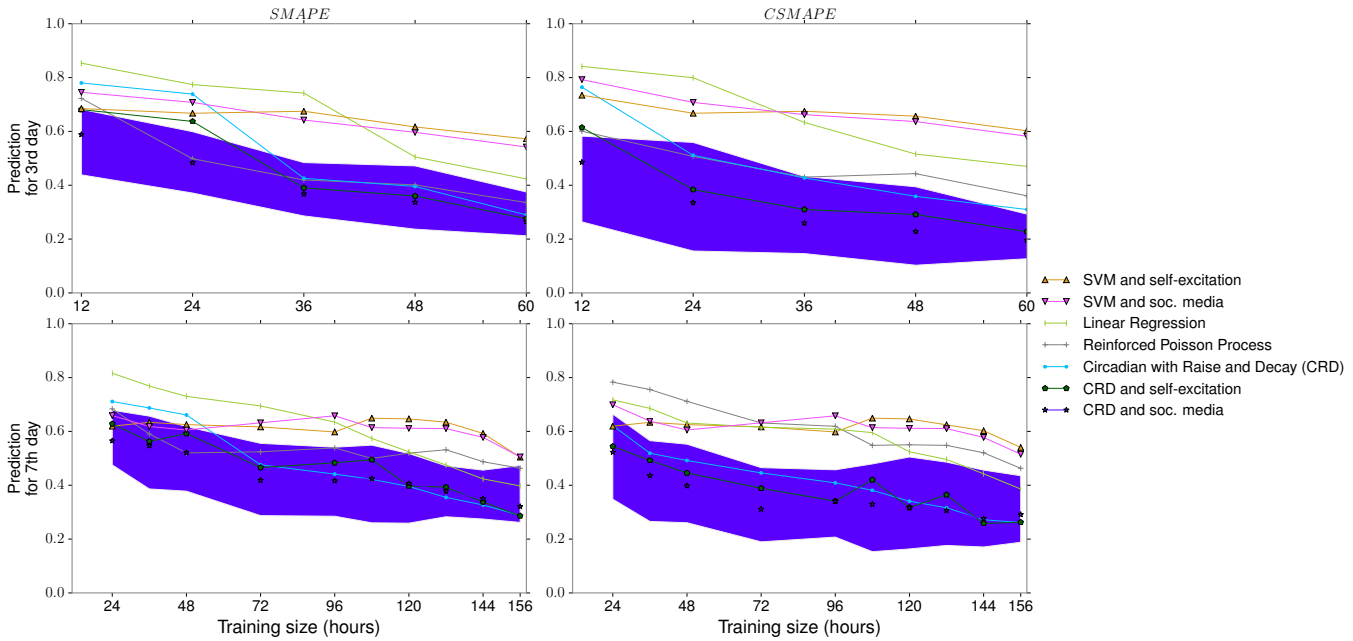

Figure 6: Error in the prediction of signatures for up to 3 days (top) and 7 days (bottom), in terms of SMAPE (left) and cumulative SMAPE (right). For each timestamp t_s (x-axis) a predictor was trained using $\{s_p(i)\}, i < t_s$ for all petitions p in the data set. The shaded area depicts the 20th and 80th percentile of the performance of the best model (CRD and social media).

6.3 Prediction

We train linear regression and SVM models for each input size T_{train} and prediction length $T_{\text{tot}} - T_{\text{train}}$. As training data, we use 70% of the petitions selected uniformly at random. We train the model to predict signature rate occurring at an arbitrary hour in the future, as well as the cumulative number of signatures up to that point, using hourly signature $s_p(t)$ and tweet $n_{sm}(t)$ rates from the training dataset. We then test the prediction on the rest of the petitions. These experiments are performed 10 times, and we report their average performance. Estimation of the parameters of our model is performed in two steps. First, we estimate the parameters of seasonality and aging using the plain CRD model for each petition. Second, we train a linear regression model either with self-excitation $c_{self}(i)$ or social media $c_{sm}(i)$ component separately, using the results of the previous step. The latter we estimate it on the training set using Eq. 1, since the information about future postings on the social media is not known. Figure 7 shows the hourly average of social media exposures as well as its estimation by CRD model. Upon prediction we reestimate parameters a and b of Eq. 1 based on the actual social media exposures. Further, we utilize the predicted values as $n_{\text{sm}}(t)$ in Eq. 2.

Prediction accuracy. Figure 5 shows an example of an actual time series for signatures and the result of predictions with our models and the baselines. We show the advantage of incorporating information from social media in terms of generating a prediction that follows more closely the actual evolution of the number of signatures. Note that our models significantly outperform the baselines. We systematically evaluate all models using introduced metrics in Figure 6, which shows the results of predicting the number of signatures for up to 3 (upper plots) or 7 days (bottom plots). The

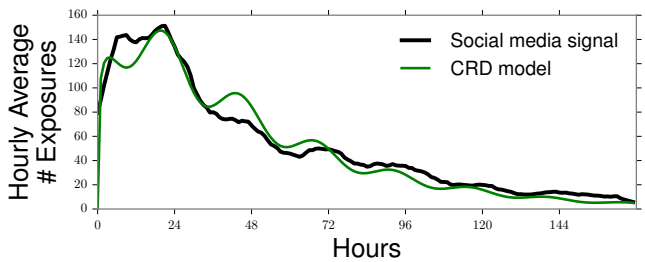

Figure 7: Hourly average social media exposure for a petition during its first week. CRD model has the following parameters: $a = 102.14, b = 16.67, \phi = 8.90, k = 0.25, \tau = 37.86$

x-axis corresponds to the amount (in hours) of training data each method receives. We observe that the performance of the SVM-based methods is the lowest, linear regression and reinforced Poisson process have intermediate performance, and the performance of CRD, CRD with social media and CRD with self excitation are the highest. The latter two behave similarly, except when little training data is available, at the very beginning of a petition. In that case, CRD with social media is better than CRD with self excitation.

Given the size of the entire collection, the average improvement of considering front page information for 75 petitions is relatively small. However, among 150 petitions described in Section 4.3, the front page effect brings an improvement of about 4% on average in terms of SMAPE for the prediction of up to 3 days, with respect to CRD with social media and front page effect in which c_{front} is forced to be 0.

6.4 Analysis of Estimated Parameters

This subsection describes the analysis of the estimated parameters of the CRD models as well as its external influence functions.

Circadian Rhythm and Aging. As a by-product of modeling each petition using the Circadian with Rise and Decay (CRD) model given in Eq. 1, we obtain a distribution for each parameter across all petitions. These distributions are shown in Figure 9, where we separate failed petitions from successful ones, as well as a special case of successful petitions, which are the ones promoted on the front page.

As expected, we observe that the intensity parameter a, which corresponds to the offset in the signature rate, is higher for successful petitions that for unsuccessful ones. Interestingly, the amplitude parameter b shows that the oscillations of the series are larger for failed petitions, perhaps because failed petitions are more localized within a single time zone. The growth parameter k, which influences the day at which a petition reaches its peak, shows that successful petitions tend to be more popular early on in comparison with failed petitions, and that the peak of the petitions that are promoted on the front page happens later in time—likely at the moment when the petition ranks the highest on the front page. The decay parameter τ can be much larger for successful petitions, meaning that they sustain interest for a longer period of time (in the model this appears as $e^{-t/\tau}$). Finally, most of the petitions have a similar shift of the circadian rhythm, given by phase parameter ϕ, since most of them are created in the USA and signed by people in the same country, in time zones that are close to each other (the distributions are almost equal so they are omitted from the figure).

Self-Excitation vs External Influence. Our model uses a time window of size T_{mem} hours, which allows to incorporate information from the recent past in its estimation of the future. Each of the coefficients for the influence of self-excitation $c_{\text{self}}(i)$, social media $c_{\text{sm}}(i)$, and front-page effect $c_{\text{front}}(i)$ can be seen as a time-indexed vector reflecting the importance of different moments of the recent past for each specific influence across successful petitions. If we are predicting the popularity on $t+1$ hour, the influence function corresponds to the vector of size T_{mem} that contains the impact of each prior hour $t-i$ of past signatures, social media exposures, or front-page rank, where $i = 0, 1, \ldots, T_{\text{mem}}$. The centroids of these vectors are shown in Figure 8.

Several interesting observations can be made from Figure 8. First, self-excitation seems to be largely memoryless, with the immediately preceding step being the most influential element. Second, social media (Twitter in this case) has an influence that can last up to four hours for the successful petitions, and peaks about 2 hours after posting; this means that posting at time t mostly affects the signature rate between times $t+1h$ and $t+3h$. Failed petitions are less affected by social media and only within 1 hour after the posting. Third, being featured on the front-page significantly boosts signature rate for up to 3 hours, in agreement with our observations from Section 4.3. In relative terms, a post on social media has a stronger (external) impact on future number of signatures than adding one signature (self-excitation), and being featured on the front page has a stronger effect than social media activity.

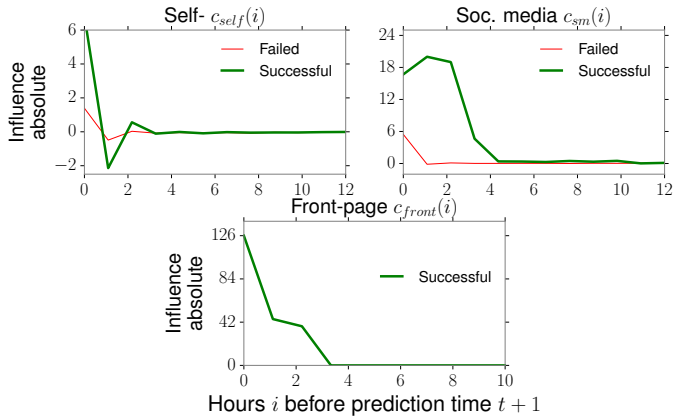

Figure 8: Influence functions estimation for self-excitation, social media, and front-page effect. A value of i on the x-axis refers to the median influence of this aspect i hours in the past. On each plot, the y-axis presents an absolute scale for successful and failed petitions and shows the multiplicative effect on the number of signatures. Petitions promoted on the home page are all successful.

7. CONCLUSIONS

Online user engagement is a complex phenomenon, challenging us to understand interdependent activities across websites that are less studied than those happening on a particular website. In this paper, we studied an important form of engagement, signing an e-petition, and modeled two external influences: activity on social media, and promotion to front page. We demonstrated significant improvement in modeling and predicting engagement when those influences are taken into account. In addition, we showed that the circadian rhythm of human activity, and the fact that interest decays over time, also need to be considered. We analyzed the effect of social media and found it to be impactful in two ways. First at a micro level, as demonstrated by the matching of people signing a petition and then posting about it shortly afterwards. Second at a macro level, where we analyzed the effect of Twitter on the signature rate using a Granger causality test, and showed significant improvement in prediction accuracy when using social media—improvements that are particularly important to reduce the amount of time/data needed to perform an accurate prediction. We were also able to determine that the effect of Twitter posts lasts for about 5 hours and peaks at about 1-3 hours since posting. These findings are relevant beyond online petitions, as many campaigners in social media (e.g., promoting brands, causes, or candidates) also perform similar activities in order to boost user engagement.

Specifically for online petitions, we showed that successful petitions tend to peak early and continue receiving attention for longer time. In other words, it is not just about having a "strong start," but about being able to sustain this engagement day after day. Petitions can be boosted by activity on social media, and by featuring them prominently to a large audience of potential signatories, as demonstrated by the front page effect that we have modeled and measured. These findings are probably relevant for people running other types of campaigns, and may be particularly important for crowd-

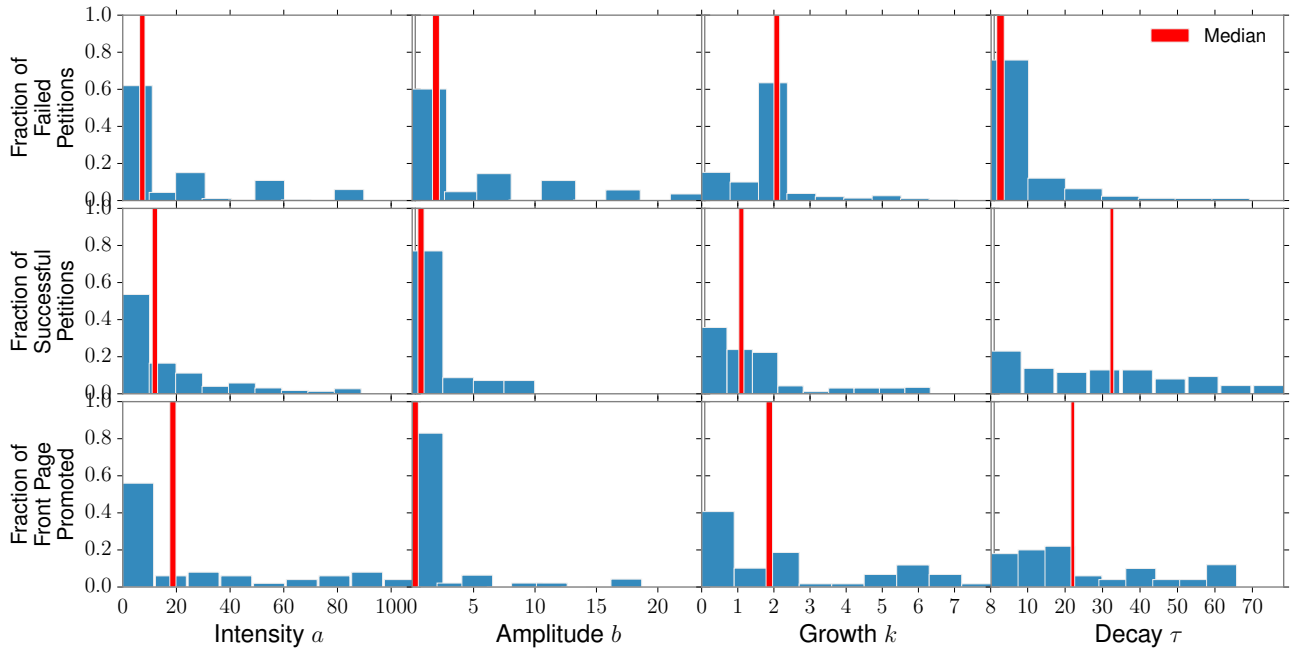

Figure 9: Distributions of parameter estimations for failed petitions (top) and successful petitions (middle). We also consider separately petitions promoted on the front page, all of them successful (bottom). Details about each parameter are provided in Section 5.1.

funding campaigns. In general, running a successful online campaign requires sustained attention and punctual interventions. In that context, interpretable models that can provide actionable insight about how a campaign is evolving are vastly more useful than opaque models, even if the latter were to provide small advantages in terms of prediction accuracy.

Future Work. We believe that this paper is an important step towards better modeling and predicting how reinforced information spreads online. It can be extended in a number of ways. In terms of new methods, it would be interesting to explore how the effects of several petitions on each other could be modeled, and how social media communities and influencers, defined both topically and through network structures, could be incorporated into our models. Moreover, impact functions could be represented through parametric distribution functions. In terms of enhancing the prediction accuracy, further sources of social media, and new features, could easily be incorporated into our model. Since we are modeling the petitions at an individual level, it might also be interesting to build and compare our model to a batch model and apply it over specific clusters of petitions. Finally, a prediction using a stochastic Hawkes process might be compared to the deterministic one presented in this paper.

Code and anonymized data are available at `https://github.com/toluolll/www2017petitions`.

Acknowledgments. We would like to thank Prof. Daniel Gatica-Perez for the fruitful discussions, advices and comments during the initial stage of the project. The first author was supported by a Sinergia Grant by the Swiss National Science Foundation (SNF 147609). This work was supported by the Catalonia Trade and Investment Agency (*Agència per la competitivitat de l'empresa*, ACCIÓ); ACT-I, JST, JSPS KAKENHI Grant Number 25870915, and the Okawa Foundation for Information and Telecommunications.

8. REFERENCES

[1] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proc. of WSDM*, pages 607–616. ACM, 2013.

[2] J. An, D. Quercia, and J. Crowcroft. Recommending investors for crowdfunding projects. In *Proc. of WWW*, pages 261–270. ACM, 2014.

[3] P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng. Popularity prediction in microblogging network: A case study on Sina Weibo. In *Proc. of WWW (companion volume)*, pages 177–178, 2013.

[4] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proc. of CSCW*, pages 211–223. ACM, 2014.

[5] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proc. of WWW*, pages 925–936. ACM, 2014.

[6] J. Choi and W. B. Croft. Temporal models for microblogs. In *Proc. of CIKM*, pages 2491–2494. ACM, 2012.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[8] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang. Discover breaking events with popular hashtags in Twitter. In *Proc. of CIKM*, pages 1794–1798. ACM, 2012.

[9] V. Etter, M. Grossglauser, and P. Thiran. Launch hard or go home!: Predicting the success of kickstarter campaigns. In *Proc. of COSN*, pages 177–182. ACM, 2013.

[10] E. Ferrara, O. Varol, F. Menczer, and A. Flammini. Detection of promoted social media campaigns. In *Proc. of ICWSM*, 2016.

[11] S. Gao, J. Ma, and Z. Chen. Modeling and predicting retweeting dynamics on microblogging platforms. In *Proc. of WSDM*, pages 107–116. ACM, 2015.

[12] T. Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(1):1–24, 2009.

[13] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi. On the reliability of profile matching across large online social networks. In *Proc. of KDD*, pages 1799–1808. ACM, 2015.

[14] C. Granger. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1):199 – 211, 1988.

[15] S. A. Hale, H. Margetts, and T. Yasseri. Petition growth and success rates on the UK no. 10 Downing street website. In *Proc. of WebSci*, pages 132–138. ACM, 2013.

[16] X. He, M. Gao, M.-Y. Kan, Y. Liu, and K. Sugiyama. Predicting the popularity of web 2.0 items based on user comments. In *Proc. of SIGIR*, pages 233–242. ACM, 2014.

[17] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proc. of WWW (companion volume)*, pages 57–58. ACM, 2011.

[18] S.-W. Huang, M. M. Suh, B. M. Hill, and G. Hsieh. How activists are both born and made: An analysis of users on change.org. In *Proc. of CHI*, pages 211–220. ACM, 2015.

[19] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In *Proc. of WWW (companion volume)*, pages 657–664, 2013.

[20] R. Kobayashi and R. Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Proc. of ICWSM*, 2016.

[21] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *Proc. of CIKM*, pages 2335–2338. ACM, 2012.

[22] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. of WWW*, pages 621–630. ACM, 2010.

[23] H. Li, X. Ma, F. Wang, J. Liu, and K. Xu. On popularity prediction of videos shared in online social networks. In *Proc. of CIKM*, pages 169–178. ACM, 2013.

[24] S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. In *Proc. of ICML*, pages 1413–1421, 2014.

[25] Z. Ma, A. Sun, and G. Cong. Will this #hashtag be popular tomorrow? In *Proc. of SIGIR*, pages 1173–1174. ACM, 2012.

[26] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[27] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[28] Y. Ogata and K. Katsura. Immediate and updated forecasting of aftershock hazard. *Geophysical Research Letters*, 33(10), 2006. L10305.

[29] S. J. Olshansky and B. A. Carnes. Ever since gompertz. *Demography*, 34(1):1–15, 1997.

[30] J. Proskurnia, K. Aberer, and P. Cudré-Mauroux. Please sign to save... : How online environmental petitions succeed. In *Proc. of Workshop on Social Web for Environmental and Ecological Monitoring*, 5 2016.

[31] J. Proskurnia, R. Mavlyutov, R. Prokofyev, K. Aberer, and P. Cudre-Mauroux. Analyzing large-scale public campaigns on twitter. In *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA*, pages 225–243. Springer International Publishing, 2016.

[32] S. D. Roy, T. Mei, W. Zeng, and S. Li. Towards cross-domain learning for social video popularity prediction. *IEEE Transactions on Multimedia*, 15(6):1255–1267, Oct 2013.

[33] H.-W. Shen, D. Wang, C. Song, and A.-L. Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *Proc. of AAAI*, pages 291–297. AI Access Foundation, 2014.

[34] B. Shulman, A. Sharma, and D. Cosley. Predictability of popularity: Gaps between prediction and understanding. In *Proc. of ICWSM*, 2016.

[35] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, Aug. 2010.

[36] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1):1–20, 2014.

[37] H. Xu, M. Farajtabar, and H. Zha. Learning granger causality for hawkes processes. In *Proc. of ICML*, pages 1717–1726, 2016.

[38] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang. Modeling user posting behavior on social media. In *Proc. of SIGIR*, pages 545–554. ACM, 2012.

[39] 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 *Proc. of KDD*, pages 1513–1522. ACM, 2015.