Vaca Ruiz et al. *EPJ Data Science* 2014, 3:5
http://www.epjdatascience.com/content/3/1/5

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Modeling dynamics of attention in social media with user efficiency

Carmen Vaca Ruiz[1,3*], Luca Maria Aiello[2] and Alejandro Jaimes[2]

*Correspondence:
cvaca@fiec.espol.edu.ec
[1]Politecnico di Milano, Piazza Leonardo Da Vinci, 32, Milan, Italy
[3]FIEC, Escuela Superior Politecnica del Litoral, Campus Gustavo Galindo, Km 30.5 via Perimetral, Guayaquil, Ecuador
Full list of author information is available at the end of the article

Abstract

Evolution of online social networks is driven by the need of their members to share and consume content, resulting in a complex interplay between individual activity and attention received from others. In a context of increasing information overload and limited resources, discovering which are the most successful behavioral patterns to attract attention is very important. To shed light on the matter, we look into the patterns of activity and popularity of users in the Yahoo Meme microblogging service. We observe that a combination of different type of social and content-producing activity is necessary to attract attention and the efficiency of users, namely the average attention received per piece of content published, for many users has a defined trend in its temporal footprint. The analysis of the user time series of efficiency shows different classes of users whose different activity patterns give insights on the type of behavior that pays off best in terms of attention gathering. In particular, sharing content with high spreading potential and then supporting the attention raised by it with social activity emerges as a frequent pattern for users gaining efficiency over time.

Keywords: online attention; microblogging; social networks; time series

1 Introduction

Understanding users' activities in social media platforms, in terms of the actions they take and how those actions affect the attention they receive (e.g., comments, replies, re-posts of messages they post, etc.), is crucial for understanding the dynamics of social media systems as well as for designing incentives that lead to growth in terms of user activity and number of users. As expected, given the nature of such platforms, users who receive attention from their peers tend to be more engaged with the service and are less likely to churn out [1]. Insights on the kinds of actions that users take to gain more attention and become "popular" are therefore important because they can help explain how social media platforms evolve. In spite of the importance of analyzing such behavior at a large scale, the dynamics of attention are not well understood. This is largely due to two main reasons: on one hand that there are few datasets that show the evolution of a network from its very beginnings, and on the other hand, because most work has focused on the popularity of content rather than on analyzing the effects of user's behaviors on how other users react to them. For example, there have been many studies to establish the reasons behind user or item popularity in social networks (e.g., [2, 3]), but the effects that the patterns of attention received have on the activity and the engagement of the "average" users have not been thoroughly explored so far.

In this paper, we address questions that focus on social media users' behavior at different stages of their participation in social media platforms. In particular, we introduce a new way to examine attention dynamics, and from this perspective perform a deep analysis of the evolution of user activity and attention in a social network from its beginning until the service ceased to exist. Analyzing the weekly efficiency, i.e. the amount of attention received in the platform normalized by the amount of content produced, we observe that 56% of the users in the dataset exhibit a footprint of their efficiency with a clearly defined trend (i.e., sharply increasing/decreasing or peaking). We are able to extract patterns of user behavior from these temporal footprints that reveal differences in the activity behavior of users of different classes. We focus our analysis on Yahoo Meme, a microblogging service that was launched by Yahoo in 2009 and discontinued in 2012. While the mechanisms of interaction in Yahoo Meme were similar to those found in other social media platforms, to the best of our knowledge, this is the first study that examines in detail the questions we are addressing from the perspective of user efficiency, using data from a service from its initial launch.

The main contributions of this work include:

- Study of the attention dynamics in social networks from the angle of *efficiency*, namely the ratio between attention received and activity performed. The notion of efficiency in time allows to detect patterns that could not emerge using other raw popularity or activity indicators.
- Definition of a method to classify noisy time series of user-generated events. The method is successfully used to find classes of users based on the time series of their efficiency scores, with an accuracy ranging from 0.85 to 0.93, depending on the different classes.
- Extraction of insights useful to detect and prevent user churn. For instance, exploration of the efficiency time series reveals that increase in efficiency is determined by creation of high-quality content, but the acquired attention has to be sustained with additional social activity to keep the efficiency high. If such social exchange is missing, attention received drops very quickly.

2 Related work

Much effort has been spent lately in measuring the effect that the activity of content production and sharing has in influencing the actions of social media participants. Depending on whether the investigation adopts the perspective of the *user* who is sharing or of the *content* being shared, emphasis has been given to the characterization of either the influential users or the process of information spreading along social connections.

Different methods to identify influentials, namely individuals who seed viral information cascades, have been proposed recently [4], and it has been observed that simple measures such as the raw number of social connections are not good predictors of influence potential [5–7]. Instead, the ease of propagation of a piece of content is correlated with many other features, including the position of the content creator in the social network [8], demographic factors [9, 10], and the sentiment conveyed in the message [11].

For what concerns content-centered analysis, much attention has been devoted to the study of the structure and diffusion speed of information cascades in social and news media [12–14], including Yahoo Meme [14, 15]. Weng *et al.* [14] for instance have shown that

triadic closure helps to explain the link formation in early stages of the user's lifetime but later in time it is the information flow the driver for new connections. Despite the difficulty of determining whether observed cascades are generated by a real influence effect [16] (unless performing controlled experiments [17]), the role of influence in social network dynamics is widely recognized, albeit not fully understood. Factors related to influence include geolocation, visibility of the content, or exogenous factors like major geopolitical or news events for news media [18–20].

Patterns of temporal variation of popularity have been investigated previously, mostly focusing on the attention given to pieces of user-generated content. Previous work includes characterization of the peakness and saturation of video popularity on YouTube in relation to content visibility [18], crowd productivity dependence on the attention gathered by videos [1], the classification of bursty Twitter hashtags in relation to topic detection tasks [21], and the clustering of hashtag popularity histograms based on their shape [22]. Time series has been used to predict popularity in blogs, where the early reactions of the crowd to a piece of content is strongly correlated to the expected overall popularity [3, 23].

In this work we focus on users as opposed to content and we analyze time series of a metric combining the user activity and the attention received. We do not focus on the popularity gained at a global scale, but instead we characterize temporal patterns of activity and attention of each individual. We show that time series of individual user activity cannot be clustered accurately based on their shapes by state-of-the art methods, so we propose an algorithm to fix that. Finally, except in rare cases (e.g., [24]), previous work on network analysis has relied mostly on limited temporal snapshots. In contrast, we use the temporal data of the entire life-span of Meme, from its release date until its shutdown.

3 Dataset description

Meme was a microblogging service launched by Yahoo in April 2009 and discontinued in May 2012. Users could *post* messages, receive notifications of posts published by people they *follow* (follower ties are *directed* social connections), and *repost* messages of other users or *comment* on those messages. The overall number of registered users grew at a constant pace, up to almost $700K$. When neglecting uninvolved users (i.e., users who were registered, but stopped explicit activity), we observe a growing trend up to a maximum of $60K$ users around the end of the first year, and then a slow but steady decline. In Table 1 we report general statistics on the follower network in the last week of the service. The final network contains a well-connected core of users resulting in a greatest connected component covering almost the full network, with a high clustering coefficient. As already observed for other online social networks, the average path length is proportional to $\log\log(N)$, and similarly to other news media the level of social link reciprocity is very low [25].

Table 1 Followers network statistics

Nodes	Edges	Density	$\langle k \rangle$	$\langle k_{in} \rangle$	$GWCC_{\%}$	Reciprocity	$\langle d \rangle$	d_{max}	C
568*K*	20M	$6.2 \cdot 10^{-5}$	71	35	0.996	0.096	2.59	11	0.433

Δ = density, $GWCC_{\%}$ = relative size of the greatest weakly connected component, d = geodesic distance, C = clustering coefficient.

4 Activity vs. attention

Activity and attention are the two dimensions we aim to examine with our study. After defining the features, we look at their relationship in terms of correlations of their raw indicators and then we study them from a novel perspective by defining a metric of user efficiency. We find that very efficient users tend to write fewer posts per week but are heavily involved in social activities such as commenting.

4.1 Activity and attention metrics

We define *activity* and *attention* indicators that are computed for every user. Activity indicators are measured by the number of posts (*pd*), reposts (*rd*), and comments done (*cd*), or by the number of new followees added (*fwee*), while attention is determined by the number of reposts (*rr*) ot comments received (*cr*) from others, and by the number of new incoming follower links (*fw*). Reposts received can be *direct* or *indirect* (i.e., reposting a repost). To measure attention we consider direct reposts.

The possibility of indirect reposting originates repost *cascades* that can be modeled as trees rooted in the original post and whose descendants are the direct (depth 1) and indirect (depth 2 to the leaves) reposts. Besides being another attention indicator, the cascade size (*cs*) is a good proxy for the *perceived interestingness* of the content because, intuitively, sharing a piece of content originated by someone who is not directly linked through a social tie, and therefore is likely to be unknown to the reposter, implies a higher likelihood that the reposter was interested in that piece content. Therefore, we consider the cascade size as a measure of content interestingness.

Even though several measure of influence, authoritativeness, or more in general importance of a user in a networked system have been developed in the past (see for instance the work by Romero *et al.* [7]), here we adopt the perspective of a single user, rather than of the whole community. Therefore, we are going to interpret the system as a black box that receives input from a user (activity) and returns some output (attention), without considering the actual effect that the input causes inside the system. Although this is a simplification, it allows us to better focus on the user dimension and to cluster users with respect to the perception they get from the interaction with the system (i.e., attention in exchange for activity).

4.2 Correlations

When dealing with multidimensional behavioral data, detecting causation between events can be difficult [16], but potential mechanisms driving the interactions between the different dimensions at play can be spotted through the investigation of correlations [26]. In this case, the correlations between activity and attention metrics give a first hint about the potential payoff of some user actions in terms of attention received.

In Figure 1, visual clues of the relationship between different metrics of activity and attention are shown in the form of heatmaps. The four plots on the left display the average values of attention indicators for users whose number of posts and comments resides in given ranges. To make sure that the trends emerging from the heatmaps are significant, we count the number of users falling in each of the range buckets. In Table 2 we report the average and the median number of users in each bucket of the heatmaps. As expected from the broad distributions of the activity and attention indicators, few actors have very high values for some pairs of indicators. For instance, in the heatmap in Figure 1(E), just

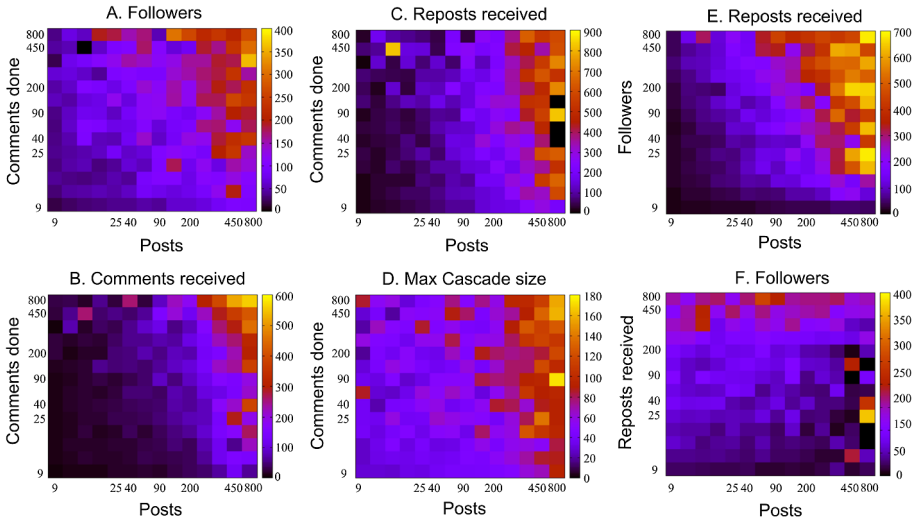

Figure 1 Correlations between activity and attention. Users were grouped according to the number of *x* and *y* values (plotted on a log scale) and, for each group, the average number of the *z*-value was calculated and mapped to a color intensity.

Table 2 Statistics for the number of users considered in each bucket of the heatmaps depicting the correlations between activity and popularity metrics (Figure 1)

***x*-axis**	***y*-axis**	**Average**	**Median**
Posts	Reposts received	73.1	39
Posts	Comments done	77.9	32
Posts	Followers	74.3	45

The average and median number of users per bucket in each combination of metrics is shown.

10 users are in the upper-right bucket (users with > 625 posts *and* and > 625 followers). However, in general the number of users per bucket is sufficiently high to consider the trend statistically significant, as shown by Table 2.

First, we observe that attention in terms of followers and comments (Figures 1(A)-(B)) is correlated with both number of posts and comments done, resulting in a color gradient becoming brighter when transitioning from the lower-left corner to the upper-right one. Users who gained more followers were heavier content producers and an even more evident correlation is found when considering comments received (Figure 1(B)), likely due to a comment reciprocity tendency (we calculated the comment reciprocity being around 24%, much higher than reciprocity in the follower network). We observe a partially similar effect when looking at content-centered indicators, namely the reposts received and the cascade size (Figures 1(C)-(D)). In these cases we find a positive correlation with the number of posts, but not with the amount of comments, suggesting that social interaction, such as commenting on other people posts, does not strongly characterize content propagation.

The two plots on the right of Figure 1 show the relation between pairs of attention metrics with the number of posts. From Figure 1(E) we learn that social exposure (i.e., being followed) and productivity (i.e., number of posts) are both heavily correlated with the number of reposts. However, people with moderate or heavy posting activity can reach a high level of attention even having a relatively small audience (as shown by the bright

colors extending down along the right side of the map). This intuition is confirmed by the fact that swapping the axes of the two attention measures, the correlation is disrupted (Figure 1(F)), meaning that people with high number of posts and reposts do not necessarily have a large number of followers.

4.3 User efficiency

The above findings support on one hand the intuitive principle about: "the more you give, the more you get" and, on the other hand, they reinforce the hypothesis that visibility is not enough to grant a wide diffusion of content (similarly to the "million follower fallacy" in the context of Twitter [6]). However, the user perception of the interaction with peers through an online system is not dependent just by the raw number of feedback actions received, but also by the amount of attention in relation with the effort spent to gain it. Given this perspective, we define the *efficiency* η of a user u in a given time frame $[t_i, t_j]$ as the amount of attention received over the amount of activity performed between t_i and t_j, for any pair of activity (Act) and attention (Att) metrics:

$$\eta_u^{Act,Att}(t_i, t_j) = \frac{\sum_{t_i}^{t_j} Att_u}{\sum_{t_i}^{t_j} Act_u}. \tag{1}$$

Analogous definitions have been used in different disciplines such as physics and economics [27], and in most of the cases the efficiency is upper bounded to 1, i.e., the outcome from the system cannot exceed the energy given in input. On the contrary, in a social media setting the efficiency is unbounded and it constitutes an objective function to maximize in order to increase the engagement of the user base. Even if comments can be strong indicators of involved user participation, the main focus of the online service under study is posting and reposting, similarly to Twitter. Therefore we always consider the number of posts as the metric of activity in the efficiency formula. In the above definition (Formula (1)) we assume that the attention that we take into account should be the one that is directly triggered by the activity considered, we use either the number of reposts ($\eta_u^{Post,Repost}$) or the number of comments ($\eta_u^{Post,Comm}$) as proxies for attention received, since other metrics such as number of followers are not necessarily responses to the posting activity.

The distribution of $\eta_u^{Post,Repost}$ and $\eta_u^{Post,Comm}$ for all the users during the complete lifetime of the network is drawn in Figure 2. Even if the maximum efficiency scores span up to several hundreds, the majority of users have an efficiency lower than 1, and most of them

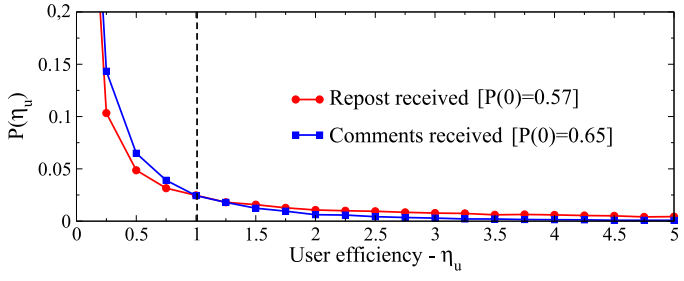

Figure 2 Efficiency scores. Distribution of efficiency scores, bucketed in 0.25-wide bins. Average scores are 0.38 for comments and 1.55 for reposts.

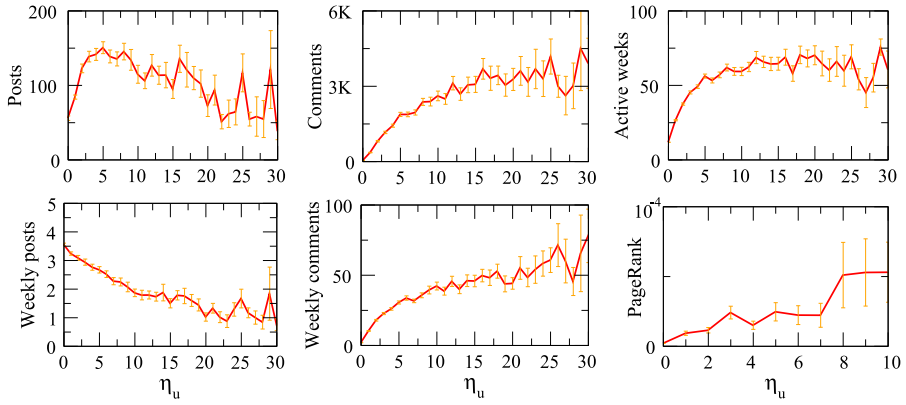

Figure 3 Activity and efficiency. Average values of activity and status indicators at fixed values of η_u.

have values close to zero. The average over the η_u values of all users is higher than 1 for reposts and much lower for comments. This is justified by the fact that Meme emphasized especially the repost feature. For this reason, next we consider only the efficiency of posts in relation to reposts, and we refer to it as η_u, for simplicity.

High activity is usually indicative of poor efficiency or, in other words, activity alone is not indicative of high potential of attention gain. To study more in depth the traits of efficient and inefficient users, we describe users with different η_u values according to several activity and status features, as shown in Figure 3.

Insightful patterns emerge. First, the higher the η_u, the lower the activity in terms of number of posts, but not in the range $0 \leq \eta_u \leq 5$ (containing most of the users), in which the number of posts grows with η_u. However, when looking at the average number of posts submitted per week instead, the trend becomes monotonic, confirming the theory about the limited attention of the audience being a barrier for attention gathering [20]. Second, the higher the η_u, the higher the amount of comments: the more efficient users are the ones who comment the most. Finally, the longevity of the profile and the prestige on the follower network (computed with standard PageRank) are also distinctive features of efficient users.

5 Evolution of efficiency in time

Attention attracted by users, and by consequence their efficiency, is not constant in time. It depends on the amount of activity, the position in the network and other factors. However next we show that, even if many users exhibit a oscillating but globally stable values of efficiency in time, more than half the users show sharp variations in their efficiency time series, that tell more about the activity behavior in different periods of the user lifetime. First, we give the definition of efficiency time series. Then, we explain the algorithm used to classify users efficiency traces according to the shape of their trend and discuss the properties of the four classes we found. We (i) find that state-of-the-art algorithms for clustering of timeseries do not perform well on the noisy traces such the ones generated by human activity, therefore, based on the observed shapes, (ii) we propose a new classification method and evaluate it against a human-curated ground truth, and (iii) we analyze the differences between user behaviors in the four main user efficiency classes around the main changepoint of the efficiency curve.

5.1 Efficiency time series definition

By adapting the efficiency formula for a discrete-time scenario, we model the temporal efficiency evolution using weekly time series for each user u measuring the efficiency η_u after each week. The elements of the series are generated as follows:

$$\eta_u(t_i) = \frac{rr(p_{t_i})}{|p_{t_i}|}, \quad t_i \in T_u = \{t_1, \ldots, t_n\},$$

where p_{t_i} represents the set of posts published by user u on week t_i, $rr(p_{t_i})$ is the total number of direct reposts received in the user's lifetime for the set of posts p_{t_i}, and T_u is the sorted list of weeks in which the user u published at least one post.

5.2 Time series type detection

Characterizing users based on the exhibited temporal behavior of their efficiency requires to extract automatically patterns out of the generated time series. There are two main families of state-of-the art methods for this task. The first one includes *feature-based* approaches that cluster series based on their kurtosis, skewness, trend, and chaos [28]. The latter one includes *area-under-the-curve* methods [29–31] that consist into dividing the time series into equally sized fragments, measure the area under the curve in each fragment, represent the time series as a vector of such quantities, and then apply a clustering algorithm over them (specifically, we used k-means). We first tried those state-of-the art methods to cluster the efficiency time series. We do not report extensively the results obtained for the sake of brevity, but both feature-based approaches area-under-the-curve methods produce clusters containing extremely heterogeneous curves, as we assessed by manual inspection. In addition to that, we tried also a separate approach, proposed few years ago, that transforms the curves through Piecewise Aggregate Approximation and Symbolic Aggregate Approximation and then clusters the resulting representations with k-means [32]. Also this method lead to very imbalanced clusters, being the 99% of curves put in one single cluster. The main issue with those approaches is that they have been tested in the past mainly on synthetic time series. When time series represent the activity of single actors they may have an extremely broad variety of length, shapes, and oscillation of the curve that the mentioned methods are not able to handle properly.

Even though the produced clusters were very noisy, the area-under-the-curve method tended to group together curves in four main clusters, with a predominance of well-recognizable shapes: *increasing*, *decreasing*, *peaky* and *steady*. Some examples of time series for each class are depicted in Figure 4 (top). Driven by the qualitative insights that the clustering produced, we developed a tailored classification algorithm to obtain cleaner groups, based on a qualitative, discrete representation of the temporal data, inspired by the representation of financial time series presented by Lee *et al.* [33]. Our algorithm executes the following steps:

1. *Smoothing*. Apply the kernel regression estimator of Nadaraya and Watson [34] to the user temporal data to obtain a smoothed time series t. The smoothing process gets rid of very sharp and punctual fluctuations, which are very frequent in human activity time series. Examples of raw curves compared to their smoothed versions are shown in Figure 4 (bottom).
2. *Linguistic transform*. Generate a qualitative representation of the time series t for a user u using three states: High, Medium, Low (H, M, L). We empirically set the

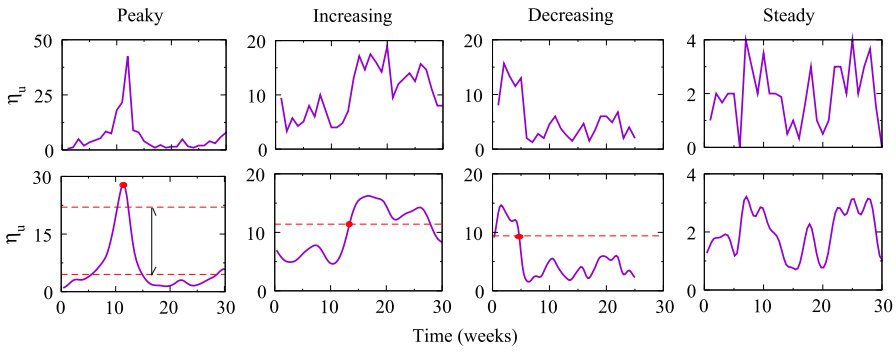

Figure 4 Time series examples for each class. Examples of efficiency time series for users of each class indicated by the clustering of time series. Top: raw time series, bottom: smoothed Time series. Threshold used to detect changepoints for the first three types are reported with dashed lines.

threshold for high values to 0.6 and for medium values to 0.3 (i.e., values greater than the 60% of the maximum efficiency reached by the user are considered High). The idea of using threshold values is supported by previous work in time-series segmentation [35].

3. *Fluctuation reduction*. Search for contiguous subsequences of a given state and drop the subsequences whose length is less than the 10% of the total length. Similarly to the smoothing procedure, this step helps to eliminate noisy fluctuations in the time series. For example, in the series *HHHMHHMMMLLL*, the fourth element, *M* is dropped.
4. *String collapsing*. Collapse the string representation of t by replacing subsequences of the same state with a single symbol of the same type. For instance, the resulting series from the previous example, *HHHHHMMMLLL*, is transformed to *HML*.
5. *Detection of Increasing/Decreasing classes*. Look for collapsed sequences with just two groups of symbols and classify as "Increasing" a sequence transitioning from L or M to the state H and as "Decreasing" those transitioning from H to L or M. The second and third columns in Figure 4 show the threshold for High values as a dotted red line.
6. *Detection of Peaky class*. For the unclassified series, find those exhibiting a peaky shape by looking at outliers in the series whose value is higher than x times the average value. This method has been successfully used before in the context of Twitter, with $x = 5$ [21]. Other methods for peak detection we tested [36] find just local peaks, which are very frequent in noisy time series.
7. *Detection of changepoint*. Accurately locating the point in which a curve transitions between different levels is important to study the behavior of users in their single activity and popularity metrics around the point in time when these changes occur [37]. For the peak type curves, the changepoint is intuitively defined by the highest peak, whereas for the increasing and decreasing types the point is identified by the time in which the linguistic representation of the series transitions from H to M or L status (decreasing) or from L or M to H status (increasing). For the sake of comparison, we match our simple technique with the statistical change point analysis recently proposed by Chen *et al.* [38]. We find that, although for most time series the values from the two methods were very close (at most 1 or 2 weeks difference in

around 80% of the cases), the statistical changepoint detection often identifies points right before or right after a change of efficiency.

8. *Detection of Steady class*. The remaining time series are classified as steady.

As in most previous work [39], in absence of an automatic way to compute the quality of the classes, two of the authors annotated a random sample of 1,000 time series per class to assess the goodness of our algorithm. Since the expected shapes of the curves for each class are very clear (see examples in Figure 4) a human evaluator can decide with certainty whether the instances from the sample match the expected template. The outcome of the labeling is very encouraging, with 93% correct instances in the Decreasing class, 86% in the Increasing, and 85% in Peak, and almost perfect agreement between evaluators (Fleiss $\kappa = 0.80$). For the Steady class, where shapes can vary much, we labeled as *misclassified* any curve belonging to the other classes. We found a low portion of *misclassifed* instances (12%). We observe that the users in the steady class are around 44%, meaning that 56% of the users exhibit a temporal footprint of the efficiency curve that has a clearly defined trend. This is a finding with important implications on the applicative side, meaning that the majority of users could be accurately profiled as having consistently increasing or decreasing efficiency patterns.

5.3 Changepoint detection

Accurately locating the point in which a curve transitions between different levels is important to characterize the user behavior when his efficiency significantly increases or drops, thus allowing to study how single activity and popularity metrics vary when these changes occur [37]. Changepoint detection refers to the problem of finding time instants where abrupt changes occur [37]. Except for the steady time series, which denote a user behavior that is quite constant in time (or for which transition to higher or lower efficiency levels are much slower), all the other three types have a changepoint in which the efficiency trend changes radically in a relatively short period of time compared to the total length of the user lifetime. For the peak type curves, the changepoint is intuitively defined by the highest peak, whereas for the increasing and decreasing types the point is identified by the time in which the linguistic representation of the series transitions from H or M to L status (decreasing) or from L or M to H status (increasing). More general methods to identify changepoints relying on the changes in mean and variance have been proposed in the past. For the sake of comparison, we match our simple technique with the statistical change point analysis recently proposed by Chen *et al.* [38]. We find that, although for most time series the values from the two methods were very close (at most 1 or 2 weeks difference in around 80% of the cases), the statistical changepoint detection sometimes identifies points right before or right after a change of efficiency. In fact, the generality of statistical methods is not a plus in cases in which the set of curves in input is quite homogeneous and for which ad-hoc methods result more reliable. For this reason, we use our definition of changepoint.

Once users with similar profiles in their temporal efficiency evolution have been grouped, time series are analyzed to identify meaningful changepoints.

6 User efficiency classes

For each detected class, we perform an analysis in aggregate over all the users first and then we characterize the evolution of the same metrics in time. We find that (i) publishing

Table 3 Activity, popularity and longevity indicators for the four user classes

Type	%users	Activity			Attention				Time	
		pd	***cd***	***fwee***	***cr***	***fw***	***rr***	***cs***	***days***	***weeks***
Decreasing	15%	6.11	2.78	10.7	4.90	3.57	25.3	34	491	53
Increasing	16%	10.3	4.74	9.69	6.14	4.82	43.4	51	690	92
Peak	25%	8.10	2.74	6.82	4.07	3.18	9.11	32	703	85
Steady	44%	8.22	3.75	10.3	5.50	4.35	29.1	40	610	72

Values are the median of the average weekly values. Abbreviations used are *pd* = posts, *cd* = commentsDone, *fwee* = followees, *fw* = followers, *rr* = repostsReceived, *cs* = cascadeSize, *days* = userLifetime, *week* = activeWeeks

interesting content helps to boost the efficiency of the subsequent posts through attention gathering and that (ii) the efficiency gained in that way should be sustained by intense social activity to avoid it to drop.

6.1 Static analysis of user classes

We aggregate different activity and attention indicator scores over users and weeks, for each of the four user classes. For all the indicators, we compute their average value per-week for every user and then we compute the median of all the results obtained for users of the same class. Median is used instead of average to account for the broad distribution of values. In addition, to get a measure of the adhesion of users to the service, we measure the median number of weeks of activity and the median number of days of duration of the user account. Values for all the metrics are shown in Table 3 and they show a first picture of the levels of activity performed and attention attracted by users of different classes. Users in the Increasing class have the highest values for almost all the metrics compared to other groups. They are able to attract high levels of attention (*fw*, *rr*), combined with the ability of conciliating the production of content of high interestingness for the community (high *cs*) with social activity (high *cd* and *fwee* values). As we will show later, the production of comments and addition of followees is a characteristic of this class through time. On the contrary, users belonging to the Peak class are the least active in terms of social activity (low *cd* and *fwee* values) but, surprisingly, they are relatively active content publishers and have the tendency to be active for long time, exhibiting a high number of active weeks and the highest account duration. They are quite involved in posting but are not much engaged in the social interactions that complements the content production and consumption process. As we will observe next, these users do some commenting activity at the beginning of their lifetime but they reduce significantly the number of followees or comments rapidly. Users in the Decreasing and Steady classes receive both a good amount of attention and establish a high number of social links, backed up by a high content-production activity in the Steady case. Given the shorter time of involvement and knowing about their sharp efficiency drop, the users in the Decreasing class are likely people with a good level of participation who, differently from the users in Steady, reduced significantly the involvement in the service at some point.

6.2 Variation around the changepoint

Here we investigate deeper how users in each class distribute the amount of activity in time. We perform an analysis around the changepoint of the efficiency curve, and see if the different temporal patterns can explain *why* their efficiency level changed over time. We decompose the timeseries into different *phases* and study the relations between them

in terms of the activity and attention indicators. Specifically, for all the users belonging to the classes where the changepoint is given (i.e., all but the Steady class).

Let us define three user-dependent time steps: the week in which the user activity started w_{start}, the week of the changepoint of the efficiency curve w_{cp}, and the week of the end of the activity w_{end}, after which no other action is performed by the user. Accordingly, we define three *phases* of the user lifespan referred as *Before*, *CP*, *After*, which represent, respectively: the weeks in the $[w_{start}, w_{cp})$ interval, the changepoint week w_{cp}, and the weeks in the $(w_{cp}, w_{end}]$ interval. We calculate the average weekly amount of activity and attention metrics during these three macro-aggregates of weeks. The three values obtained for each indicator capture the variation of activity and attention when approaching the critical point in which a consistent change of efficiency is detected.

To detect the variation of the values in the three phases we compute two ratios for each user: (a) RatioCP = activity-or-attention metric measured in w_{cp} divided by the same metric computed in $[w_{start}, w_{cp})$, and (b) RatioAfter = activity-or-attention metric measured in $(w_{cp}, w_{end}]$ divided by the same metric during $[w_{start}, w_{cp})$. Ratios are then averaged over all the users of each class. Comparison of ratios between different user classes reveals the key differences between them: values above 1 mean that the value of the indicator grew in *CP* of in *After* phases compared to the *Before* phase. Final results for different values of activity and attention are reported in Figure 5. For instance, in Figure 5(a), we observe that RatioAfter is above 1 just for users in the Peaky class. It means that the users in that class have published more posts after the changepoint than they did before it. We can summarize our findings as follows:

- *Activity and attention at CP.* Users of all classes maintain a similar trend in the number of posts done in *CP* with a slight increase in the case of the Increasing class (Figure 5(a)). For Peak and Increasing classes, the number of reposts received, cascade size and followers increases significantly in *CP* compared to *Before* (Figure 5(d), (e), (f) respectively). Since reposts received and cascade size are proxies for content interestingness, this indicates the production of content that attracts the attention of a much higher number of users. For both classes, this is the most likely cause of the rise

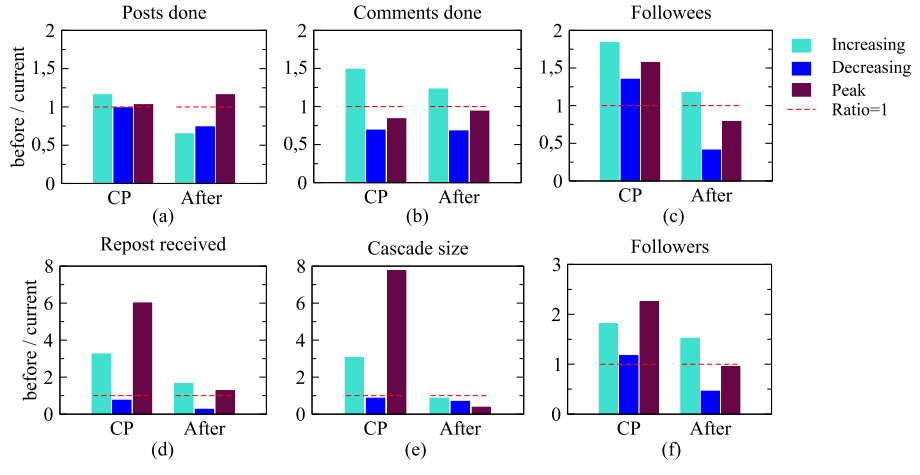

Figure 5 Ratio of activity and attention. Ratio of activity and attention metrics between the Before phase and later phases (Change Point and After), for the 3 user classes.

of their efficiency at *CP*. For the Decreasing class the attention values start dropping instead. Finally, differently from other classes, users in the Increasing class produce a higher number of comments in *CP* (Figure 5(b)).

- *Social activity after CP*. In the *After* phase, social interaction such as the number of comments and the addition of new followees considerably increase compared to *Before* for the Increasing class (Figures 5(b), (c)), while they remain stable or in slight decrease for the Peak class. Decreasing class values drop also in this case.
- *Content production activity after CP*. The reverse scenario is found when looking at the posting activity. In the *After* phase, Peak post messages at a higher rate than Before (Figure 5(a)), while *Increasing* posting activity drops in favor of a higher attention to social interaction.

The main lesson learned from the above findings is that the submission of pieces of "interesting" content, namely posts that attract the attention of a wider audience than usual, is the trigger to transition to higher efficiency levels. However, efficiency cannot be maintained without cost. Increasing engagement in social activity and expanding the potential audience turns out to be an effective strategy not to lose efficiency. Conversely, producing more content without reinforcing the social relationships with the potential consumers of the content results in a rapid drop of efficiency to the original levels. The difference between the Increasing and Peaky classes is particularly striking, having the Increasing-type users fully exploiting social activity with 17% more followees, 23% more comments and 61% reposts after their changepoint, while Peaky-type users keep their activity approximately stable (except for an increase of reposts done). Moreover, as expected, when a status of equilibrium between attention received and activity is disrupted by an arbitrary reduction of productivity and social interactions, the efficiency is destined to fade quickly.

7 Conclusions

We explored the interplay between activity and attention in Yahoo Meme by defining the notion of user *efficiency*, namely the amount of attention received in relation to the content produced. We find that, unlike the raw attention measures, efficiency has strong negative correlation with the amount of user activity and users who are involved in social activities such as commenting, have higher centrality in the social network than average, but are not necessarily heavy content producers.

However, if we consider commenting as a form of content creation, we observe that comment takes less effort than creating a post but, frequently, it can be more effective. It is so because the reciprocity plays a role and the comments network exhibits a higher reciprocity than that of the follower network. Users can, thus, benefit from the visibility of a post whenever they comments on it.

We classify into four main classes (sharp increasing/decreasing steps, peaks or stable trend) the time series of user efficiency with a novel algorithm that overcomes limitations of previous approaches and we find four main clusters. By analyzing the variation of activity and attention around the changepoints of the timeseries, we find evidences that user efficiency is boosted by a particular combination of production of interesting content and constant social interactions (e.g., comments). In these cases, users gather the attention from a wider audience by publishing content with higher spreading potential and then they manage to keep the attention high through regular and intensified social activity. These insights find direct application on the detection and prevention of user churn: being able

to detect users who increase their efficiency but that are frustrated by not being able to keep it high can be helped either by recommending them social activities or pushing their contacts to interact with them. The task of churn prediction is a natural continuation of the present work that we plan to address in the future.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
CVR and LMA designed the methodology and conceived the experiments. CVR performed the data processing, the time series clustering and calculated the ratios. LMA calculated the overall network statistics. All authors wrote and revised the manuscript. This work was carried out while CVR was an intern at Yahoo Labs, Barcelona.

Author details
[1]Politecnico di Milano, Piazza Leonardo Da Vinci, 32, Milan, Italy. [2]Yahoo Labs, Av. Diagonal 177, 08018, Barcelona, Spain. [3]FIEC, Escuela Superior Politecnica del Litoral, Campus Gustavo Galindo, Km 30.5 via Perimetral, Guayaquil, Ecuador.

Acknowledgements
This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975. Carmen Vaca research work has been funded by ESPOL and the Ecuadorian agency SENESCYT. We would like to thank Amin Mantrach, Neil O'Hare, Daniele Quercia, and Rossano Schifanella for the useful discussions.

Received: 19 December 2013 Accepted: 13 February 2014 Published: 04 Mar 2014

References

1. Huberman BA, Romero DM, Wu F (2009) Crowdsourcing, attention and productivity. J Inf Sci 35(6):758-765
2. Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A (2010) Characterizing and modeling the dynamics of online popularity. Phys Rev Lett 105(15):158701
3. Szabo G, Huberman BA (2010) Predicting the popularity of online content. Commun ACM 53(8):80-88
4. Pal A, Counts S (2011) Identifying topical authorities in microblogs. In: Proceedings of the fourth ACM international conference on web search and data mining (WSDM), pp 45-54. ACM, New York
5. Asur S, Huberman BA, Szabo G, Wang C (2011) Trends in social media: persistence and decay. In: Proceedings of the 5th AAAI conference on weblogs and social media (ICWSM)
6. Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in Twitter: the million follower fallacy. In: AAAI conference on weblogs and social media (ICWSM), vol 10, pp 10-17
7. Romero DM, Galuba W, Asur S, Huberman BA (2011) Influence and passivity in social media. In: WWW'11: proceedings of the 20th international conference companion on world wide web. ACM, New York, pp 113-114
8. Hong L, Dan O, Davison BD (2011) Predicting popular messages in Twitter. In: WWW. ACM, New York
9. Strufe T (2010) Profile popularity in a business-oriented online social network. In: Proceedings of the 3rd workshop on social network systems (SNS). ACM, New York, p 2
10. Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: 2010 IEEE second international conference on social computing (SocialCom). IEEE Press, New York, pp 177-184
11. Quercia D, Ellis J, Capra L, Crowcroft J (2011) In the mood for being influential on Twitter. In: 2011 IEEE third international conference on privacy, security, risk and trust (PASSAT) and 2011 IEEE third international conference on social computing (SocialCom). IEEE Press, New York, pp 307-314
12. Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of the 18th international conference on world wide web (WWW). ACM, New York, pp 721-730
13. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on Twitter. In: Proceedings of the fourth ACM international conference on web search and data mining (WSDM). ACM, New York, pp 65-74
14. Weng L, Ratkiewicz J, Perra N, Gonçalves B, Castillo C, Bonchi F, Schifanella R, Menczer F, Flammini A (2013) The role of information diffusion in the evolution of social networks. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'13, pp 356-364
15. Ienco D, Bonchi F, Castillo C (2010) The meme ranking problem: maximizing microblogging virality. In: 2010 IEEE international conference on data mining workshops (ICDMW). IEEE Press, New York, pp 328-335
16. Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. Sociol Methods Res 40(2):211-239
17. Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of the 21st international conference on world wide web (WWW). ACM, New York, pp 519-528
18. Figueiredo F, Benevenuto F, Almeida JM (2011) The tube over time: characterizing popularity growth of YouTube videos. In: Proceedings of the fourth ACM international conference on web search and data mining (WSDM). ACM, New York, pp 745-754
19. Brodersen A, Scellato S, Wattenhofer M (2012) YouTube around the world: geographic popularity of videos. In: Proceedings of the 21st conference on world wide web (WWW). ACM, New York, pp 241-250
20. Weng L, Flammini A, Vespignani A, Menczer F (2012) Competition among memes in a world with limited attention. Sci Rep 2:335

21. Lehmann J, Gonçalves B, Ramasco JJ, Cattuto C (2012) Dynamical classes of collective attention in Twitter. In: Proceedings of the 21st international conference on world wide web (WWW). ACM, New York, pp 251-260
22. Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on web search and data mining (WSDM). ACM, New York, pp 177-186
23. Mathioudakis M, Koudas N, Marbach P (2010) Early online identification of attention gathering items in social media. In: Proceedings of the third ACM international conference on web search and data mining (WSDM). ACM, New York, pp 301-310
24. Kooti F, Yang H, Cha M, Gummadi KP, Mason WA (2012) The emergence of conventions in online social networks. In: AAAI conference on weblogs and social media (ICWSM)
25. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web (WWW). ACM, New York, pp 591-600
26. Schifanella R, Barrat A, Cattuto C, Markines B, Menczer F (2010) Folks in folksonomies: social link prediction from shared metadata. In: Proceedings of the third ACM international conference on web search and data mining. ACM, New York, pp 271-280
27. Arthur S, Sheffrin SM (2003) Economics: principles in action. Prentice Hall, New York
28. Wang X, Smith K, Hyndman R (2006) Characteristic-based clustering for time series data. Data Min Knowl Discov 13(3):335-364
29. Fu T-C (2011) A review on time series data mining. Eng Appl Artif Intell 24(1):164-181
30. Geurts P (2001) Pattern extraction for time series classification. In: Principles of data mining and knowledge discovery. Springer, Berlin, pp 115-127
31. Warren Liao T (2005) Clustering of time series data - a survey. Pattern Recognit 38(11):1857-1874
32. Lin J, Keogh E, Wei L, Lonardi S (2007) Experiencing sax: a novel symbolic representation of time series. Data Min Knowl Discov 15(2):107-144
33. Lee CHL, Liu A, Chen WS (2006) Pattern discovery of fuzzy time series for financial prediction. IEEE Trans Knowl Data Eng 18(5):613-625
34. Härdle W, Vieu P (1992) Kernel regression smoothing of time series. J Time Ser Anal 13(3):209-232
35. Assfalg J, Kriegel HP, Kroger P, Kunath P, Pryakhin A, Renz M (2006) Similarity search on time series based on threshold queries. In: Advances in database technology - EDBT, pp 276-294
36. Palshikar G (2009) Simple algorithms for peak detection in time-series. In: Proceedings of the 1st international conference on advanced data analysis, business analytics and intelligence (ADABAI)
37. Basseville M, Nikiforov IV (1993) Detection of abrupt changes: theory and applications. Prentice Hall, New York
38. Chen J, Gupta AK (2011) Parametric statistical change point analysis: with applications to genetics, medicine, and finance. Birkhäuser, Basel
39. Lin J, Li Y (2009) Finding structural similarity in time series data using bag-of-patterns representation. In: Scientific and statistical database management. Springer, Berlin, pp 461-477

10.1186/epjds30
Cite this article as: Vaca Ruiz et al.: **Modeling dynamics of attention in social media with user efficiency.** *EPJ Data Science* 2014, 3:5