OPEN ACCESS Freely available online

PLOS ONE

Predictability of Extreme Events in Social Media

José M. Miotto*, Eduardo G. Altmann

Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

Abstract

It is part of our daily social-media experience that seemingly ordinary items (videos, news, publications, etc.) unexpectedly gain an enormous amount of attention. Here we investigate how unexpected these extreme events are. We propose a method that, given some information on the items, quantifies the predictability of events, i.e., the potential of identifying in advance the most successful items. Applying this method to different data, ranging from views in YouTube videos to posts in Usenet discussion groups, we invariantly find that the predictability increases for the most extreme events. This indicates that, despite the inherently stochastic collective dynamics of users, efficient prediction is possible for the most successful items.

Citation: Miotto JM, Altmann EG (2014) Predictability of Extreme Events in Social Media. PLoS ONE 9(11): e111506. doi:10.1371/journal.pone.0111506

Editor: Tobias Preis, University of Warwick, United Kingdom

Received June 16, 2014; **Accepted** September 27, 2014; **Published** November 4, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All databases were retrieved from public sources available in the Internet. Compiled tables with the data necessary to reproduce the findings are available at http://dx.doi.org/10.6084/m9.figshare.1160515.

Funding: The authors have no funding or support to report.

Competing Interests: Author EGA is an academic editor for PLOS ONE. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* Email: jmiotto@pks.mpg.de

Introduction

When items produced in social media are abundant, the public attention is the scarce factor for which they compete [1–3]. Success in such *economy of attention* is very uneven: the distribution of attention across different items typically shows heavy tails which resemble Pareto's distribution of income [4] and, more generally, are an outcome of complex collective dynamics [5–12] and non-trivial maximizations of entropic functions [13,14]. Increasing availability of large databases confirm the universality of these observations and renew the interest on understanding the dynamics of attention, see Tab. 1.

Universal features of heavy-tailed distributions do not easily lead to a good forecast of specific items [5], a problem of major fundamental and practical interest [15–19]. This is illustrated in Fig. 1, which shows that the heavy-tailed distribution appears at very short times but items with the same early success have radically different future evolutions. The path of each item is sensitively dependent on idiosyncratic decisions which may be amplified through collective phenomena. An important question is how to quantify the extent into which prediction of individual items is possible (i.e., their *predictability*) [20]. Of particular interest –in social and natural systems– is the predictability of extreme events [21–26], the small number of items in the tail of the distribution that gather a substantial portion of the public attention.

Measuring predictability is difficult because it is usually impossible to disentangle how multiple factors affect the quality of predictions. For instance, predictions of the attention that individual items are going to receive rely on (i) information on properties of the item (e.g., metadata or the attention received in the first days) and (ii) a prediction strategy that converts the information into predictions. The quality of the predictions reflect the interplay between these two factors and the dynamics of attention in the system. In particular, the choice of the prediction strategy is crucial. Instead, predictability is a property of the system and is by definition independent of the prediction strategy (it is the upper bound for the quality of any prediction based on the same information on the items). A proper measure of the predictability should provide direct access to the properties of the system, enabling a quantification of the importance of different information on the items in terms of their predictive power.

In this paper we introduce a method to quantify the predictability of extreme events and apply it to data from social media. This is done by formulating a simple prediction problem which allows for the computation of the optimal prediction strategy. The problem we consider is to provide a binary (yes/no) prediction whether an item will be an extreme event or not (attention passes a given threshold). Predictability is then quantified as the quality of the optimal strategy. We apply this method to four different systems: views of YouTube videos, comments in threads of Usenet discussion groups, votes to StackOverflow questions, and number of views of papers published in the journal PLOS ONE. Our most striking empirical finding is that in all cases the predictability increases for more extreme events (increasing threshold). We show that this observation is a direct consequence of differences in (the tails of) the distributions of attention conditioned by the known property about the items.

The paper is divided as follows: Sec. Motivation motivates the problem of event prediction by showing that it is robust to data with heavy tails. Sec. Methods introduces the method to quantify predictability, which is used in the Sec. Application to Data. A summary of our findings appears in Sec. Conclusions.

Motivation

Characterization of Heavy-tails

Different systems in which competition for attention takes place share similar statistical properties. Here we quantify attention of

Table 1. Examples in which fat-tailed distributions of popularity across items have been reported.

System	Item	Attention measure	Refs.
Online Videos	video	views, likes	[18]
Discussion Groups	threads	posts, answers	[38]
Publications	papers	citations, views	[6,8,15,19]
Twitter	tweet	retweets	[9]
WWW	webpage	views	[11]
Online Petitions	petition	signers	[39]

doi:10.1371/journal.pone.0111506.t001

published items in 4 representative systems (see Appendix S1, Sec. 1 for details; all the data is available in Ref. [27]):

- views received by 16.2 million videos in YouTube.com between Jan. 2012 and Apr. 2013;
- posts written in 0.8 million threads in 9 different Usenet discussion groups between 1994 and 2008;
- votes to 4.6 million questions published in Stack-Overflow between Jul. 2008 and Mar. 2013.
- views of 72246 papers published in the journal PLOS ONE from Dec. 2006 to Aug. 2013 (see also Ref. [28]).

The tails of the distribution $P(X)$ of attention X (views, posts, etc.) received by the items (videos, threads, etc.) at a large time t after publication is characterized without loss of generality using Extreme Value Theory. It states that for large thresholds x_p the probability $P(X|X>x_p)$ follows a Generalized Pareto distribution [29]

$$P(X>x|X>x_p)\sim\left(1+\frac{x-x_p}{\sigma\alpha}\right)^{-\alpha}. \quad (1)$$

The fits of different partitions of our databases yield $\alpha\in[0.50,4.36]$ and are statistically significant already for relatively small x_p's (p-value >0.05 in 52 out of 59 fits, see Appendix S1, Sec. 2 and Fig. S1 for details). These results confirm the presence of heavy tails, an observation reported previously in a variety of cases (see Tab. 1). This suggests that our databases are representative of social media more generally (while scientific publications are usually not classified as social media items, from the point of view of their online views, they are subject to the same attention-gathering process).

Prediction of Extreme Events

Prediction in data with heavy tails is typically not robust. As an example, consider using as a predictor $\hat{X}$ of the future attention

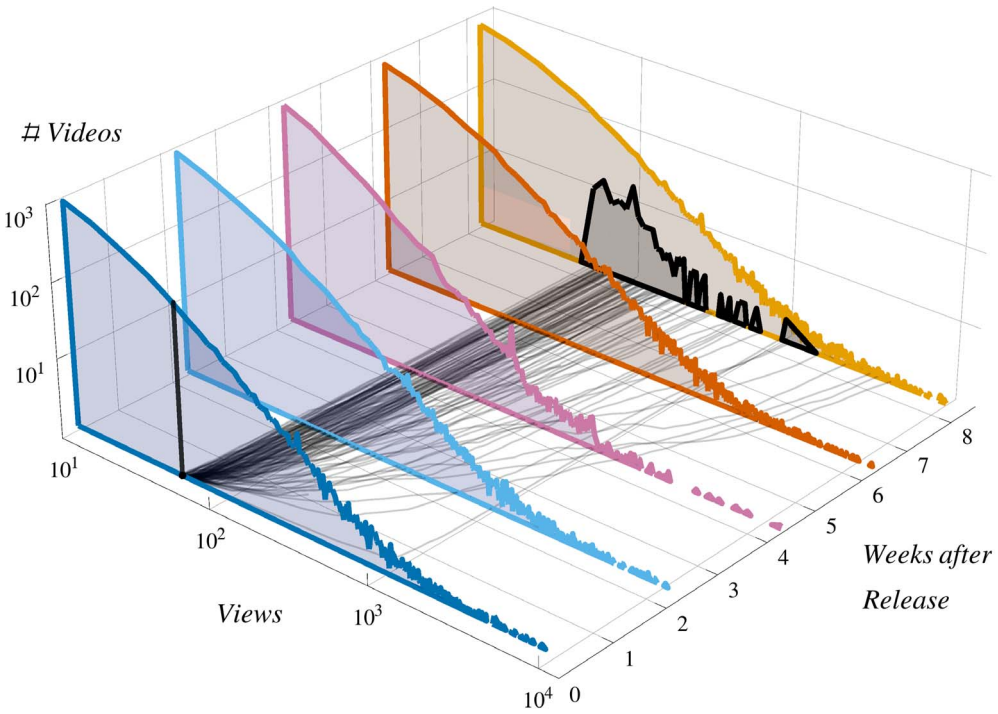

Figure 1. Dynamics of views in YouTube. Colored histograms: distributions of views at fixed times after publication (0.3 million videos from our database). **Gray lines at the bottom**: trajectories of 120 videos which had the same early success (50 views 2 days after publication). **Black histogram**: distribution of views of the 120 selected videos 2 months after publication.
doi:10.1371/journal.pone.0111506.g001

the mean $\hat{X}=\sum_{x=1}^{\infty} xP(x)$, which is the optimal predictor, if we measure the quality of prediction with the standard deviation of X. For heavy-tailed distributions, the mean and standard deviation may not be defined (for $\alpha<1$ and $\alpha<2$, respectively), making prediction not robust (i.e., it depends sensitively on the training and target datasets). This illustrates the problems heavy-tails typically appear when value predictions are issued and indicates the need for a different approach to prediction of attention.

We consider the problem of *event prediction* because, as shown below, it is robust against fat-tailed distributions. We say an event E happens at time t if the cumulative attention $X(t)$ received by the considered item until time t is within a given range of values. We are particularly interested in predicting extreme events $X(t)>x_*$, i.e., to determine whether the attention to an item passes a threshold x_* before time t. The variable to be predicted for each item is binary: E or $\bar{E}$ (not E). We consider the problem of issuing binary predictions for each item (E will occur or not), which is equivalent to a classification problem and different from a probabilistic prediction (E will occur with a given probability). Heavy tails do not affect the robusteness of the method because all items for which $X(t)>x_*$ count the same (each of them as one event), regardless of their size x. Indeed, the tails of $P(X>x_*)$ determine simply how the probability of an event $P(E)$ depends on the threshold x_* (we assume $P(X)$ exists).

Methods

In this section we introduce a method to quantify predictability based on the binary prediction of extreme events. This is done by arguing that, despite the seeming freedom to choose between different prediction strategies, it is possible to compute a single optimal strategy for this problem. We then show how the quality of prediction can be quantified and argue that the quality of the optimal strategy is a proper quantification of predictability.

Predictions are based on information on items which generally lead to a partition of the items in groups $g\in\{1,\ldots,G\}$ that have the same feature [30]. As a simple example of our general approach, consider the problem of predicting at publication time $t=0$ the YouTube videos that at $t=t_*=20$ days will have more than $x_*=1000$ views (about $P(E)\approx 6\%$ of all videos succeed). As items' information, we use the category of a video so that, e.g., videos belonging to the category *music* correspond to one group g and videos belonging to *sport* correspond to a different group g'. Since the membership to a group g is the only thing that characterizes an item, predictive strategies can only be based on the probability of having E for that group, $P(E|g)$.

In principle, one can think about different strategies on how to issue binary predictions on the items of a group g. They can be based on the likelihood (L) $P(E|g)$ or on the posterior (P) probability $P(g|E)$ [22], and they can issue predictions stochastically (S), with rates proportional to the computed probabilities, or deterministically (D), only for the groups with largest $P(g|E)$ or $P(E|g)$. These simple considerations lead to four (out of many) alternative strategies to predict events (raise alarms) for items in group g

(**LS**) stochastically based on the likelihood, i.e., with probability $\min\{1,\beta P(E|g)\}$, with $\beta\geq 0$;

(**LD**) deterministically based on the likelihood, i.e., always if $P(E|g)>p_*$, with $0\leq p_*\leq 1$;

(**PS**) stochastically based on the posterior, i.e., with probability $\min\{1,\beta' P(g|E)\}$, with $\beta'\geq 0$;

(**PD**) deterministically based on the posterior, i.e., always if $P(g|E)>p'_*$, with $0\leq p'_*\leq 1$.

In the limit of large number of predictions (items), the fraction of events that strategy (LS) predicts for each group g matches the probability of events $P(E|g)$ and therefore strategy (LS) is *reliable* [31] and can be considered a natural extension of a probabilistic predictor. Predictions of strategies (LD), (PS) and (PD) do not follow $P(E|g)$ and therefore they are not reliable.

The quality of a strategy for event prediction is assessed by computing the false alarm rate (or False Positive Rate, equal to one minus the specificity) and the hit rate (True Positive Rate, equal to the sensitivity) over all predictions (items), see Appendix S1, Sec. 3 for details. Varying the amount of desidered false alarms of the prediction strategy (β,p_*,β', and p'_* in the examples above), a curve in the hit $\times$ false-alarm space is obtained, see Fig. 2(a). The overall quality is measured by the area below this curve, known as Area Under the Curve (AUC) [32]. For convenience, we use the area between the curve and the diagonal (hits = false-alarms), $\Pi=2\mathrm{AUC}-1$ (equivalent to the Gini coefficient). In this way, $\Pi_S\in(-1,1)$ represents the improvement of strategy S against a random prediction. In absence of information $\Pi_S=0$ and perfect predictions lead to $\Pi=1$. In the YouTube example considered above, we obtain $\Pi_{PS}<\Pi_{LS}<\Pi_{PD}<\Pi_{LD}$ (17%, 18%, 29%, 32%), indicating that strategy (LD) is the best one.

We now argue that strategy (LD) is optimal (or *dominant* [33]), i.e., for any false alarm rate it leads to a larger hit rate than any other strategy based on the same set of $P(E|g)$. To see this, notice that strategy (LD) leads to a piecewise linear curve, see Fig. 2(b), and is the only ordering of the groups that enforces convexity in the hit $\times$ false-alarms rates space, see Appendix S1, Sec. 4 for a formal derivation. The ranking of the groups by $P(E|g)$ implies a ranking of the items, an implicit assumption in the measure of the performance of classification rules [32,34]. The existence of an optimal strategy implies that the freedom in choosing the prediction strategy argued above is not genuine and that we can ignore the alternative strategies. In our context, it implies that the performance of the optimal strategy measures a property of the system (or problem), and not simply the efficiency of a particular strategy. Therefore, we use the quality of prediction of the optimal strategy ($\Pi\equiv\Pi_{LD}$) to quantify the predictability (i.e., the potential prediction) of the system for the given problem and information. By geometrical arguments we obtain from Fig. 2 (b) (see Appendix S1, Sec. 5)

$$\Pi=\sum_g\sum_{h<g}\frac{P(g)P(h)(P(E|h)-P(E|g))}{P(E)(1-P(E))}, \qquad (2)$$

where $P(g)$ is the probability of group g and g is ordered by decreasing $P(E|g)$, i.e., $h<g\Rightarrow P(E|h)>P(E|g)$.

The value of Π can be interpreted as the probability of a correct classification of a pair of E and $\bar{E}$ items [32,34]. In practice, the optimality of this strategy is dependent on the estimation of the ordering of the groups according to $P(E|g)$. Wrong ordering may occur due to finite sampling on the training dataset or non-stationarities in the data. In fact, any permutation of indexes in Eq. (2) reduces Π.

Results

Application to Data

Here we apply our methodology to the four social-media data described above. We consider the problem of predicting at time $t_1\geq 0$ whether the attention x of an item at time $t_*>t_1$ will pass a threshold x_*. In practice, the calculation of Π from the data is done counting the number of items: (i) in each group g

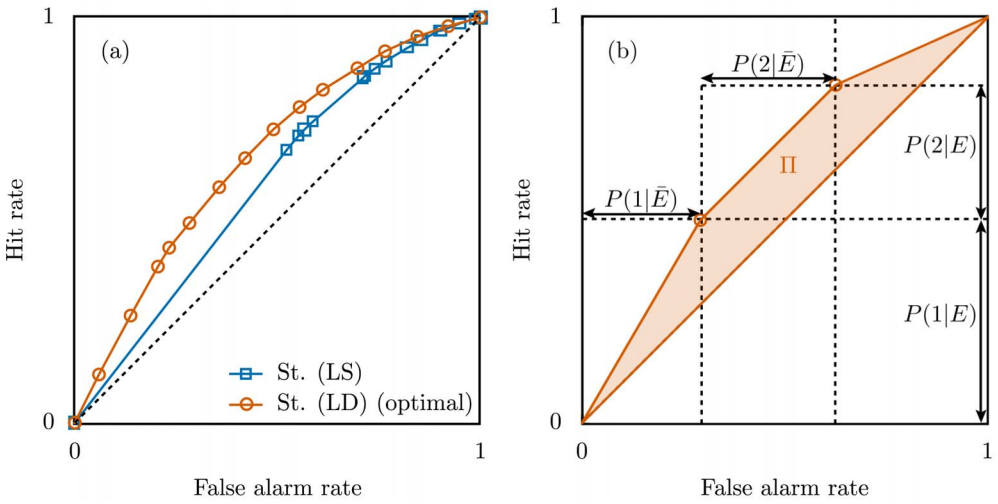

Figure 2. Quantifying the quality of event-prediction strategies requires measuring both the hit and false alarm rates. (**a**) Performance of Strategy (LS) and Strategy (LD) for the problem of predicting views of YouTube videos 20 days after publication based on their categories. The symbols indicate where the rate of issued predictions for a given group equals 1 (the straight lines between the symbols are obtained by issuing predictions randomly with a growing rate). (**b**) Illustration of the prediction curve (red line) for an optimal strategy with three groups $g=1,2,3$ with $P(1)=P(2)=P(3)=1/3$ and $P(E|1)=0.3, P(E|2)=0.2, P(E|3)=0.1$.
doi:10.1371/journal.pone.0111506.g002

$[P(g)=(\#\text{items in } g)/(\#\text{items})]$; (ii) that lead to an event $[P(E)=(\#\text{items that crossed the threshold } x_* \text{ at } t_*)/(\#\text{items})]$; and (iii) that lead to an event given that they are in group g $[P(E|g)=(\#\text{items in } g \quad \text{that crossed the threshold } x_* \text{ at } t_*)/(\#\text{items in } g)]$. Finally, the groups are numbered as $g=1,2,\ldots,G$ by decreasing $P(E|g)$ and the sum over all groups is computed as indicated in Eq. (2). In Ref. [27] we provide a python script which performs this calculation in the data.

We report the values of Π obtained from Eq. (2) considering two different informations on the items:

1) the attention at prediction time $x(t_1)$;
2) information available at publication time $t=0$ (metadata).

In case 1), a group g corresponds to items with the same $x(t_1)$. These groups are naturally ordered in terms of $P(E|g)$ by the value of $x(t_1)$ and therefore the optimal strategy is equivalent to issue positive prediction to the items with $x(t_1)$ above a certain threshold. In case 2), the groups correspond to items having the same meta-data (e.g., belonging to the same category). In this case, we order the groups according to the empirically observed $P(E|g)$ (as discussed above). Before performing a systematic exploration of parameters, we illustrate our approach in two examples:

- Consider the case of predicting whether YouTube videos at $t_*=20$ days will have more than $x_*=1{,}000$ views. For case 1), we use the views achieved by the items after $t_1=3$ days and obtain a predictability of $\Pi=90\%$. For case 2), we obtain that using the day of the week to group the items leads to $\Pi=3\%$ against $\Pi=31\%$ obtained using the categories of the videos. This observation, which is robust against variations of x_* and t_*, shows that the category but not the day of the week is a relevant information in determining the occurrence of extreme events in YouTube.
- Consider the problem of identifying in advance the papers published in the online journal PLOS ONE that received at least 7500 views 2 years after publication, i.e $X(t_*=2\text{years})>x_*=7500$ (only $P(E)=1\%$ achieve this threshold). For case 1), knowing the number of views at $t_1=2$ months after publication leads to a predictability of $\Pi=93\%$. For case 2), a predictability $\Pi=19\%$ is achieved alone by knowing the number of authors of the paper – surprisingly, the chance of achieving a large number of views decays monotonously with number of author (g increases with number of authors).

The examples above show that formula (2) allows for a quantification of the importance of different factors (e.g., number of authors, early views to the paper) to the occurrence of extreme events, beyond correlation and regression methods (see also Ref. [19]). Besides the quantification of the predictability of specific problems, by systematically varying t_1, t_*, and x_* we can quantify how the predictability changes with time and with event magnitude. Our most significant finding is that in all tested databases and grouping strategies the predictability increases with x_*, i.e., extreme events become increasingly more predictable, as shown in Fig. 3.

Discussion

We now explain why predictability increases for extreme events (increasing x_*). We first show that this is not due to the reduction of the number of events $P(E)$. Consider the case in which E is defined in the interval $[x_f-\Delta x, x_f+\Delta x)$. Assuming $P(X)$ to be smooth in X, for $\Delta x\to 0$ at fixed x_f we have that $P(E)\to P(x_f)\Delta x$ and $P(E|g)\to P(x_f|g)\Delta x$ ($P(g)$ remains unaffected), and Eq. (2) yields

$$\Pi=\frac{\sum_g \sum_{h>g} P(g)P(h)\big(P(x_f|h)-P(x_f|g)\big)}{P(E_f)[1-\Delta x P(x_f)]}, \qquad (3)$$

which decreases with $\Delta x\to 0$. This shows that the increased predictability with x_* is not a trivial consequence of the reduction of $P(E)$ ($\Delta x\to 0$), but instead is a consequence of the change in $P(E|g)$ for extreme events E.

Systematic differences in the tails of $P(X|g)$ lead to an increased predictability of extreme events. Consider the case of two groups

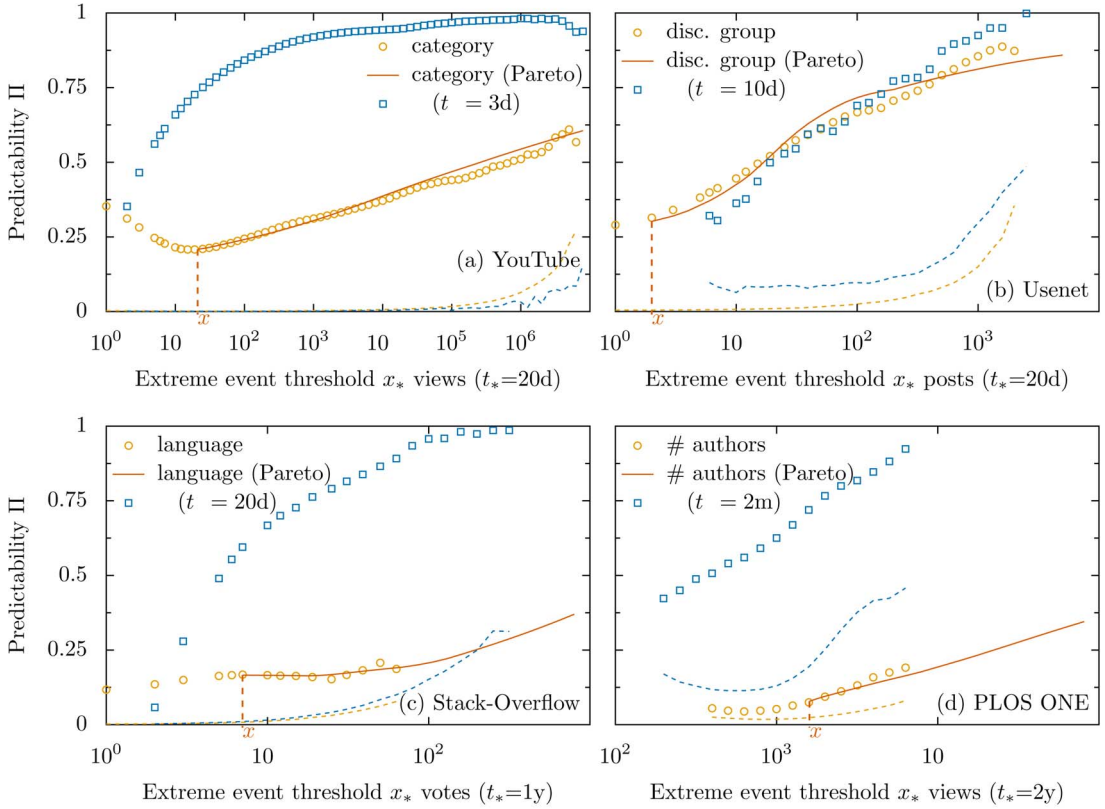

Figure 3. Predictability increases for extreme events. If the attention an item receives at time t_* is above a threshold, $X(t_*) > x_*$, an event E is triggered. The plots show how the predictability Π changes with x_* using two different informations to combine the items in groups $\{g\}$. **Black circles**: Π at time $t=0$ using metadata of the items to group them. The **red lines** are computed using as probabilities $P(E|g)$ the Extreme Value distribution fits for each group at a threshold value x_p, see Eq. (1) and SI Sec. 2. **Blue squares**: Π at time $t_1 < t_*$ using $X(t_1)$, i.e., the attention the item obtained at day t_1. The **dashed lines** are the values of the 95% percentile of the distribution generated by measuring Π in an ensemble of databases obtained shuffling the attribution of groups (g) to items (the colors match the symbols and symbols are shown only where Π is at least twice this value). Results for the four databases are shown: (**a**) YouTube (X: views of a video; metadata: video category); (**b**) Usenet discussion groups (X: posts in a thread; metadata: discussion group of the thread); (**c**) Stack-Overflow (X: votes to a question; metadata: programming language of the question, see SI Sec. 2 for details); (**d**) PLOS ONE (X: online views of a paper; metadata: number of authors of the paper).
doi:10.1371/journal.pone.0111506.g003

with cumulative distributions $P(E|g)$ that decay as a power law as in Eq. (1) with exponents α and $\alpha' = \alpha + \epsilon$, with $P(1) = P(2)$. From Eq. (2), Π for large x_* ($1 - P(E) \approx 1$) can be estimated as

$$\Pi = \frac{1}{4}\frac{P(E|1) - P(E|2)}{P(E|1) + P(E|2)} = \frac{1}{4}\frac{x_*^{-\alpha} - x_*^{-(\alpha+\epsilon)}}{x_*^{-\alpha} + x_*^{-(\alpha+\epsilon)}} \approx \frac{1}{8}\log(x_*)\epsilon, \quad (4)$$

where the approximation corresponds to the first order Taylor expansion around $\epsilon = 0$. The calculation above can be directly applied to the results we obtained issuing predictions based on metadata. The logarithmic dependency in Eq. (4) is consistent with the roughly linear behavior observed in Fig. 3(a,b). A more accurate estimation is obtained using the power-law fits of Eq. (1) for each group g and introducing the $P(E|g)$ obtained from these fits in Eq. (2). The red line in Fig. 3 shows that this estimation agrees with the observations for values $x_* \gtrapprox x_p$, the threshold used in the fit. Deviations observed for $x_* \gg x_p$ (e.g., for PLOS ONE data in panel (d)) reflect the deviations of $P(E|g)$ from the Pareto distribution obtained for small thresholds $x_p \ll x_*$. This allows for an estimation of the predictability for large thresholds x_* even in small datasets (when the sampling of E is low).

A similar behavior is expected when prediction is performed based on the attention obtained at short times t_1. Eq. (3) applies in this case too and therefore the increase in predictability is also due to change in $P(E|g)$ with x_* for different g (and not, e.g., due to the decrease of $P(E)$). For increasingly large x_* the items with significant probability of passing threshold concentrate on the large $x(t_1)$ and increase the predictability of the system. We have verified that this happens already for simple multiplicative stochastic processes, such as the geometric Brownian motion (see Fig. S2). This provides further support for the generality of our finding. The dynamics of attention in specific systems affect the shape of predictability growth with threshold.

Altogether, we conclude that the difference in (the tails of) the distribution of attention of different groups g is responsible for the increase in predictability for extreme events: for large x_*, any informative property on the items increases the relative difference among the $P(E|g)$. This corresponds to an increase of the information contained in the grouping which leads to an increase in Π.

Conclusions

In summary, we propose a method, Eq. (2), to measure the predictability of extreme events for any given available information on the items. We applied this measure to four different social media databases and quantified how predictable the attention

devoted to different items is and how informative are different properties of the items. We quantified the predictability due to metadata available at publication date and due to the early success of the items and found that usually the latter quickly becomes more relevant than the former. Our results can also be applied for combinations of different informations on the items (e.g., a group g can be composed by videos in the category *music* with a fixed $x(t_1)$). In practice, the number of groups G should be much smaller than the observations in the training dataset to ensure an accurate estimation of $P(E|g)$. Our most striking finding is that extreme events are better predictable than non-extreme events, a result previously observed in physical systems [23] and in time-series models [22,26]. For social media, this finding means that for the large attention catchers the surprise is reduced and the possibilities to discriminate success enhanced.

These results are particularly important in view of the widespread observation of fat-tailed distributions of attention, which imply that extreme events carry a significant portion of the total public attention. Similar distributions appear in financial markets, in which case our methodology can quantify the increase in predictability due to the availability of specific information (e.g., in Ref. [35] Internet activities were used as information to issue predictions). For the numerous models of collective behavior leading to fat tails [6,8–11,15,19], the predictability we estimate is a bound to the quality of binary event predictions. Furthermore, our identifications of the factors leading to an improved predictability indicate which properties should be included in the models and which ones can be safely ignored (feature selection). For instance, the relevant factors identified in our analysis should affect the growth rate of items in rich-get-richer models [11,12] or the transmission rates between agents in information-spreading models [36]. The use of Π to identify relevant factors goes beyond simple correlation tests and can be considered as a measure of causality in the sense of Granger [37].

Predictability in systems showing fat tails has been a matter of intense debate. While simple models of self-organized criticality suggest that prediction of individual events is impossible [5], the existence of predictable mechanisms for the very extreme events has been advocated in different systems [24]. In practice, predictability is not an yes/no question [7,20] and the main contribution of this paper is to provide a robust quantification of the predictability of extreme events in systems showing fat-tailed distributions.

Supporting Information

Figure S1 Distribution functions for each dataset. Dashed red line: fit of the generalized Pareto distribution (see Appendix S1 Sec. 2); Gray lines: each of the categories (see Appendix S1 Sec. 2); Blue solid line: combined data.
(PDF)

Figure S2 Predictability of simple stochastic processes. An ensemble of random walkers evolve through the dynamics $X_i(t+1)=X_i(t)(1+\varepsilon)$, where $\varepsilon\sim\mathcal{N}(\mu_i,\sqrt{\mu_i})$ (Geometric Brownian Motion with Gaussian steps. The predictability of extreme events Π was computed for $t_1=3$ steps and $t_*=15$ steps. GBM: $\mu_i=2\ \forall i$ and $X(0)\sim\mathcal{U}(0,1)$; GBM heterogeneous: $\mu_i\sim\mathcal{N}(2,0.7)$ and $X(0)\sim\mathcal{U}(0,1)$, fixed in time; GBM, init exp: $\mu_i=2\ \forall i$ and $X(0)\sim\mathcal{E}(1/6)$; GBM, $t_1=1$ the same as GBM for $t_1=1$; GBM, time decay: model proposed in Ref. [15], similar to GBM heterogeneous but with a rate that decays in time ($X_i(t+1)=X_i(t)(1+\varepsilon f_i(t)$ with $\mu_i\sim\mathcal{N}(1,0.5)$; $f_i(t)$ is a log-normal surviving probability with parameters $\mu_i^t\sim\mathcal{LN}(6.5,0.5)$ and $\sigma\sim\mathcal{LN}(1,0.2)$).
(PDF)

Appendix S1 Details on procedures, analysis and data.
(PDF)

Acknowledgments

We thank M. Gerlach, S. Hallerberg, and S. Siegert for insightful discussions and M. Gerlach and S. Bialonski for careful reading of the manuscript.

Author Contributions

Conceived and designed the experiments: JMM EGA. Performed the experiments: JMM. Analyzed the data: JMM EGA. Wrote the paper: JMM EGA.

References

1. Simon HA (1971) Designing organizations for an information rich world. In: Greenberger M, editor, Computers, communications and the public interest, John Hopkins Press. pp. 37–72.
2. Wu F, Huberman BA (2007) Novelty and collective attention. Proc Natl Acad Sci USA 104: 17599–17601.
3. Wu F, Wilkinson DM, Huberman BA (2009) Feedback loops of attention in peer production. In: International Conference on Computational Science and Engineering, 2009. CSE'09. IEEE, volume 4, pp. 409–415.
4. Pareto V (1896) La courbe de la répartition de la richesse. Ch. Viret-Genton.
5. Bak P, Paczuski M (1995) Complexity, contingency, and criticality. Proc Natl Acad Sci USA 92: 6689–96.
6. Price DJdS (1976) A general theory of bibliometric and other cumulative advantage processes. J Ame So Inf Sci Technol 27: 292–306.
7. Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. Science 854.
8. Stringer MJ, Sales-Pardo M, Amaral LAN (2010) Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. J Ame So Inf Sci Technol 61: 1377–1385.
9. Weng L, Flammini A, Vespignani A, Menczer F (2012) Competition among memes in a world with limited attention. Sci Rep 2.
10. Onnela JP, Reed-Tsochas F (2010) Spontaneous emergence of social influence in online systems. Proc Natl Acad Sci USA 107: 18375–18380.
11. Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A (2010) Characterizing and modeling the dynamics of online popularity. Phys Rev Lett 105: 158701–158705.
12. Perc M (2014) The matthew effect in empirical data. J R Soc Interface 11: 20140378.
13. Peterson J, Dixit PD, Dill KA (2013) A maximum entropy framework for nonexponential distributions. Proc Natl Acad Sci USA 110: 20380–20385.
14. Marsili M, Mastromatteo I, Roudi Y (2013) On sampling and modeling complex systems. J Stat Mech 2013: P09003.
15. Wang D, Song C, Barabási AL (2013) Quantifying long-term scientific impact. Science 342: 127–132.
16. Bandari R, Asur S, Huberman BA (2012) The pulse of news in social media: Forecasting popularity. In: Proceedings of the Sixth ICWSM.
17. Sornette D, Deschâtres F, Gilbert T, Ageon Y (2004) Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. Phys Rev Lett 93: 228701.
18. Crane R, Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. Proc Natl Acad Sci USA 105: 15649–53.
19. Penner O, Pan RK, Petersen AM, Kaski K, Fortunato S (2013) On the predictability of future impact in science. Sci Rep 3: 3052.
20. Kantz H, Altmann EG, Hallerberg S, Holstein D, Riegert A (2006) Dynamical interpretation of extreme events: predictability and predictions. In: Albeverio S, Jentsch V, Kantz H, editors, Extreme Events in Nature and Society, Springer Verlag.
21. Albeverio S, Jentsch V, Kantz H (2006) Extreme events in nature and society. Springer Verlag.
22. Hallerberg S, Altmann EG, Holstein D, Kantz H (2007) Precursors of extreme increments. Phys Rev E Stat Nonlin Soft Matter Phys 75: 016706.
23. Hallerberg S, Kantz H (2008) Influence of the event magnitude on the predictability of an extreme event. Phys Rev E Stat Nonlin Soft Matter Phys 77: 011108.

24. Sornette D (2002) Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes, and human birth. Proc Natl Acad Sci USA 99: 2522–9.
25. Ghil M, Yiou P, Hallegatte S, Naveau P, Soloviev A, et al. (2011) Extreme events: dynamics, statistics and prediction. Nonlinear Process Geophys 18: 295–350.
26. Bogachev MI, Bunde A (2011) On the predictability of extreme events in records with linear and nonlinear long-range memory: Efficiency and noise robustness. Physica A 390: 2240–2250.
27. Miotto JM, Altmann EG (2014) Time series of social media activity: Youtube, usenet, stack-overflow. Available: http://dx.doi.org/10.6084/m9.figshare.1160515.
28. Fenner M, Lin J (2013) Cumulative usage statistics for plos papers from plos website. Available: http://dx.doi.org/10.6084/m9.figshare.816962.
29. Coles S (2001) An introduction to statistical modeling of extreme values. Springer.
30. Sukhatme S, Beam CA (1994) Stratification in nonparametric roc studies. Biometrics: 149–163.
31. Bröcker J (2009) Reliability, sufficiency, and the decomposition of proper scores. Q J R Meteorol Soc 135: 1512–1519.
32. Hanley JA, McNeil B (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143: 29–36.
33. Provost FJ, Fawcett T, Kohavi R (1998) The case against accuracy estimation for comparing induction algorithms. In: ICML. volume 98, pp. 445–453.
34. Hand DJ, Till R (2001) A simple generalisation of the area under the roc curve for multiple class classification problems. Mach Learn 45: 171–186.
35. Preis T, Moat HS, Stanley HE (2013) Quantifying trading behavior in financial markets using google trends. Sci Rep 3.
36. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Rev Mod Phys 81: 591–646.
37. Granger CWJ (1980) Testing for causality: a personal viewpoint. Journal of Economic Dynamics and control 2: 329–352.
38. Altmann EG, Pierrehumbert JB, Motter A (2011) Niche as a determinant of word fate in online groups. PLoS ONE 6.
39. Yasseri T, Hale SA, Magretts H (2013) Modeling the rise in internet-based petitions. To be published.