Characterizing and modeling the dynamics of online popularity

Jacob Ratkiewicz,1 Santo Fortunato,2 Alessandro Flammini,1 Filippo Menczer,1,2 and Alessandro Vespignani1,2

1School of Informatics and Computing, Indiana University, Bloomington, IN, USA
2Complex Networks and Systems Lagrange Lab, Institute for Scientific Interchange, Torino, Italy

Online popularity has enormous impact on opinions, culture, policy, and profits. We provide a quantitative, large scale, temporal analysis of the dynamics of online content popularity in two massive model systems, the Wikipedia and an entire country’s Web space. We find that the dynamics of popularity are characterized by bursts, displaying characteristic features of critical systems such as fat-tailed distributions of magnitude and inter-event time. We propose a minimal model combining the classic preferential popularity increase mechanism with the occurrence of random popularity shifts due to exogenous factors. The model recovers the critical features observed in the empirical analysis of the systems analyzed here, highlighting the key factors needed in the description of popularity dynamics.

PACS numbers: 89.75.Hc, 89.20.-a

The dynamics of information and opinions have been deeply affected by the existence of Web-mediated brokers such as blogs, wikis, folksonomies, and search engines, through which anyone can easily publish and promote content online. This “second age of information” is driven by the economy of attention, first theorized by Simon [1]. Sources receiving a lot of attention become popular and have formidable power to impact opinions, culture, and policy, as well as advertising profit. The Web 2.0 and social media [2] not only modify traditional communication processes with new types of phenomena, but also generate a huge amount of time-stamped data, making it possible for the first time to study the dynamics of online popularity at the global system scale.

In this letter we focus on the dynamics of popularity of Wikipedia topics and Web pages. As popularity proxies we have chosen the traffic of a document, expressed by the number of clicks to that page generated by a specific population of users, and the number of hyperlinks pointing to a document. It is well documented that the statistical properties of these variables in the Web are very heterogeneous, with distributions characterized by fat tails roughly following power-law behavior [3–6]. Such distributions have been explained with models based on the rich-get-richer mechanism [7–9], but their validation from the point of view of the dynamical behavior is problematic, mainly due to the difficulty to gather relevant data. The data sets utilized here, however, contain temporal information that makes it possible to observe the growth in popularity of individual topics or pages, and allows us to statistically characterize the microdynamics by which online documents gather popularity.

Prior work on popularity dynamics has focused on news [10,11], videos [12,13] and music [14]. Here, we analyze three large scale data sets that we assembled about two information networks: the entire Wikipedia and the Chilean Web. Wikipedia is a large collaborative online encyclopedia with millions of articles and hundreds of thousands of registered contributors (en.wikipedia.org). By mining the full edit history of every article, we were able to reconstruct the entire Wikipedia structure at any past point in time. The raw data was available until March 2007 (download.wikimedia.org). Traffic data with hourly temporal resolution was obtained by cross-referencing with a separate data set originating from Wikipedia proxy server logs (dammit.lt/wikistats). Our third data source is a yearly sequence of crawls of the Chilean Web, made available by courtesy of the TodoCL search engine (www.todocl.com). This data consists of one complete crawl of the .cl top-level domain for each of the years 2002–2006. Basic statistics on each data set are shown in Table I. The representative graphs of these data sets have an approximately power-law distribution of indegree [15,17], like the Web graph at large.

In order to gauge quantitatively the popularity of documents we consider the number of hyperlinks pointing to a page (indegree k in the graph representation of the Web [3]), and the traffic s of the page, expressed by the number of clicks to it. Given either of these two popularity proxies xt at time t, we study its logarithmic derivative \( \Delta x / x = (x_t - x_{t-1}) / x_{t-1} \), which represents the relative variation of the measure in the time unit.

Fig. 1 shows the logarithmic derivative of the indegree vs time for an example page in the English Wikipedia. Despite a roughly exponential growth, the logarithmic derivative provides a signature by which different topics

Table I: Descriptions of the data sets constructed for our study. The two Wiki collections refer to indegree (1) and traffic (2) of Wikipedia topics, while the Chile collection refers to indegree of Chilean Web pages.

| Data Set | Vertices | Period | Temporal Resolution |
|---------|----------|--------|---------------------|
| Wiki1   | 3,293,102 | Jan 2001 – Mar 2007 | 1 sec. |
| Wiki2   | 3,490,740 | Feb 2008 – Current | 1 hour |
| Chile   | 3,252,779 | 2001 – 2006 | 1 year |

In order to gauge quantitatively the popularity of documents we consider the number of hyperlinks pointing to a page (indegree k in the graph representation of the Web [3]), and the traffic s of the page, expressed by the number of clicks to it. Given either of these two popularity proxies xt at time t, we study its logarithmic derivative \( \Delta x / x = (x_t - x_{t-1}) / x_{t-1} \), which represents the relative variation of the measure in the time unit.

Fig. 1 shows the logarithmic derivative of the indegree vs time for an example page in the English Wikipedia. Despite a roughly exponential growth, the logarithmic derivative provides a signature by which different topics
can be compared on the same scale. Almost all pages experience a burst in $\Delta x/x$ near the beginning of their life. Many pages receive little attention thereafter. While some pages maintain a nearly constant positive logarithmic derivative indicating an exponential growth, a number of pages continue to experience intermittent bursts in $\Delta x/x$ later in their life as in the example.

The distribution of magnitude $\Delta x/x$ for the two popularity measures at representative time resolutions is illustrated in Figs. 2a–c. In all cases and at all granularity we observe a heavy-tail behavior. Such heavy-tailed burst magnitude distributions suggest a dynamics lacking a characteristic scale. This is typical in a wide range of “critical” physical, economic, and social systems, such as avalanches, earthquakes, stock market crashes and human communication 19–23. Further evidence comes from the study of the distribution of the length of inter-event intervals. For each document we record the time stamp of each event for which $\Delta x/x > 1$ and measure the inter-event times $\Delta t$. The probability distributions of $\Delta t$ in the different data sets (Fig. 2d) are not distributed following a Poissonian, as expected by queueing theory in traditional systems, but in a power-law fashion with a finite size cutoff, as in Omori’s law of earthquakes 24 and other self-organized criticality phenomena 25.

The clear evidence for the bursty behavior of online popularity dynamics calls for a stylized model able to explain the observed features in terms of the already acquired popularity of each page and the shifts in collective attention triggered by exogenous events.

The rich-get-richer mechanism can be simulated with the classic linear preferential attachment model 9, in its directed version 20, or with the ranking model by Fortunato et al. 27. In the latter items are ranked according to their popularity $x$, and the probability that an existing item $i$ receives a unit (e.g., a click) is $P(i) \sim r_i^{-\delta}$, where $r_i$ is the rank of $i$ and $\delta > 0$ is a free parameter that tunes the power-law popularity distribution $P(x) \sim x^{-\gamma}$, such that $\gamma = 1 + 1/\delta$. Both preferential attachment and ranking models, however, fail to reproduce the long tails observed in the distributions of both $\Delta x/x$ and $\Delta t$ (Figs. 3a–b). Neither model accounts for the occurrence of exogenous factors that shift the attention of users and suddenly increase the popularity of specific topics because of events such as an actor winning a prize, political elections, etc. The minimal assumption in modeling exogenous perturbation consists in considering exter-
nal stochastic events interfering with the basic rich-get-richer mechanism by suddenly changing the popularity of a topic. The simplest way to implement this mechanism consists in introducing in the ranking model a reranking probability $\rho$, such that at each iteration every item is moved to a new position toward the front of the list, chosen randomly with equal probability between 1 (the top position) and the node’s current rank $j$. We call this the rank-shift model.

In Fig. 3a and 3b we show the indegree distribution of the rank-shift model for several values of $\rho$: $\delta = 1$ (a) and $\delta = 1.5$ (b). The ranking model ($\rho = 0$) yields the slope $1 + 1/\delta$ indicated by the dashed line. The reranking probability introduces an exponential cutoff in the distribution, which becomes relevant for $\rho \approx 10^{-2}$ and larger (but we used $10^{-5} < \rho < 10^{-3}$ in our simulations).

The distribution of $\Delta k/k$ shows two distinctive features, which are remarkably found in the empirical distributions: a maximum located in the range 0.01–0.1 and a fat tail. Since the reranking probability is low, to understand the existence and the location of the maximum it is convenient to consider the model in the absence of the reranking mechanism. At a large time $T$, the expected value of the degree of the node with rank $r$ is proportional to $L r^{-\delta}$, where $L$ is the number of links present in the network at time $T$. Let $\Delta L$ be the number of links added during the interval $\Delta T$ at whose extremes the ratio $\Delta k/k$ is computed. Let $\Delta L \ll L$, an assumption verified in our calculations. Therefore, one can safely assume that in the period $\Delta T$ the addition of new links does not affect significantly the degree of nodes and their relative ranking. So one can regard the growth process as a multinomial process with probabilities $p(r) \propto r^{-\delta}$. The expected number $\Delta k$ of new links acquired by a node of rank $r$ is therefore $p(r) \Delta L$. The assumption of (almost) stationarity also provides that $k(r) \sim p(r)L$. We therefore expect $\Delta k/k$ for a node to be distributed around $\Delta L/L$, regardless of the node. In Fig. 3c we compare the simulation of the ranking model with the one of the multinomial process with $p(r) \propto r^{-\delta}$, by using the parameters relative to the Wikipedia data set of January 2003, which represents an ideal tradeoff between the needs of having a sufficient number of bursts and a system size not too large for the model to run. The number of nodes/pages was $N \approx 1.3 \cdot 10^5$, the number of hyperlinks $L \approx 1.3 \cdot 10^6$ and $\Delta L \approx 8 \cdot 10^4$. Based on the above discussion we expect to observe a maximum in the distribution of $\Delta k/k$ located at $\Delta L/L \approx 0.06$. This is exactly where the maxima of the empirical distributions of popularity bursts are located (see Fig. 2a).

The ranking model cannot reproduce the fat tail observed in the real data. This is the reason why we introduced the reranking mechanism in our model. Here, it is the nodes that are suddenly promoted to a higher rank that are responsible for the high values of $\Delta k/k$ in the simulations. We consider a node that at time $T$ (the reference time at which we start measuring $\Delta k$) has rank $r_1$, and is immediately promoted to rank $r_2$, with $r_2$ chosen uniformly in $1 \leq r_2 \leq r_1$. Under the
same assumption of stationarity that we made above, the expected degree of the node before promotion is \( k(r_1) \approx Lp(r_1) \propto r_1^{-\delta} \). Let us further assume that \( \rho \ll 1 \) and that \( \Delta L \ll L \), which hold for the parameters used in our model. Since the reranking probability is small, we can safely assume that no node is reranked more than once during the observation time \( \Delta T \).

The expected number of links collected during the period \( \Delta T \) is then \( \Delta k = \Delta Lp(r_2) \propto r_2^{-\delta} \). We expect therefore \( \Delta k/k \propto (r_2/r_1)^{-\delta} \). It is straightforward to derive the distribution \( P(\Delta k/k) \) for a generic node that is promoted at the beginning of \( \Delta T \) by considering all pairs of values \( r_1, r_2 \) uniformly distributed in \( 1 \leq r_2 \leq r_1 \leq N \). We find \( P(\Delta k/k) \propto (\Delta k/k)^{-(1+1/\delta)} \). In Fig. 4 we highlight the tail of the distribution \( P(\Delta k/k) \) as produced by the rank-shift model and our expectation for its slope: the match is surprisingly good.

Simulations of the rank-shift model were performed using parameters matching those from the empirical data (e.g., \( N = 2.8 \times 10^8 \) nodes for the Wikipedia in 2003); the free model parameters were set to fit the empirical distributions: \( 1 \leq \delta \leq 1.2 \) and \( 10^{-5} \leq \rho \leq 10^{-3} \). For \( \rho = 0 \) we recover the original ranking model, which yields a lognormal distribution of \( \Delta x/x \), like the preferential attachment (Fig. 3). For \( \rho > 0 \) numerical simulations show that the tail of the popularity burst magnitude distribution shifts from a lognormal to a power law. The popularity distribution itself remains a power law; its exponent remains \( \gamma = 1 + 1/\delta \), but with an exponential cutoff depending on \( \rho \).

Such a parsimonious model is able to reproduce the most relevant features observed in the empirical data. Not only does rank-shift predict the distributions of both popularity measures in our data sets, but also the long tails of the distributions of indegree and traffic burst size (Fig. 3). Furthermore, it naturally accounts for the maxima of the empirical distributions. Remarkably the model captures the long-range distribution of interburst intervals as well (Fig. 4). The random rank-shift mechanism is therefore able to capture the way in which Web sites and pages gain and accumulate popularity: not by a gradual proportional process, but by a sequence of bursts that move them to the forefront of people’s attention. Such bursts are different from those observed in news-driven events [10], where attention fades rapidly and overall popularity is lognormal-distributed. We also found that smaller rank shifts are unable to capture the critical burst behavior observed in the data [28].

At the present stage our model is mostly descriptive and simply aims at reproducing at the coarsest level the distributions that characterize popularity changes. Possible refinements may include the effect of search engines, external events, news, word of mouth, social media, marketing campaigns, or any combination of them. The study of traffic patterns and models [11, 29, 30] may help shed empirical light on this question.

We thank R. Baeza-Yates, C. Cattuto, B. Dravid, V. Griffith, V. Loreto, M. Marchiori, M. Meiss. This work was supported in part by a Lagrange Senior Fellowship from the CRT Foundation to F.M., NSF grant IIS-0513650 to A.V., and the Lilly Endowment Foundation. S.F. gratefully acknowledges ICTeCollective, grant 238597 of the European Commission.

[1] H. A. Simon, in Computers, Communication, and the Public Interest, edited by M. Greenberger (The Johns Hopkins Press, Baltimore, 1971), pp. 37–72.
[2] D. Tapscott and A. D. Williams, Wikinomics: How Mass Collaboration Changes Everything (Portfolio Hardcover, 2006).
[3] R. Albert, H. Jeong, and A.-L. Barabási, Nature 401, 130 (1999).
[4] A. Broder et al., Computer Networks 33, 309 (2000).
[5] M. Meiss, F. Menczer, and A. Vespignani, in Proc. 14th Intl. World Wide Web Conf. (2005), pp. 510–518.
[6] M. Meiss et al., in Proc. 1st Intl. Conf. on Web Search and Data Mining (WSDM) (2008), pp. 65–76.
[7] H. A. Simon, Biometrika 42, 425 (1955).
[8] D. de Solla Price, J. Amer. Soc. Inform. Sci. 27, 292 (1976).
[9] A.-L. Barabasi and R. Albert, Science 286, 509 (1999).
[10] F. Wu and B. A. Huberman, Proc. Natl. Acad. Sci. USA 104, 17599 (2007).
[11] Z. Dezsö et al., Phys. Rev. E 73, 066132 (2006).
[12] G. Szabo and B. A. Huberman, Tech. Rep., arXiv:0811.0405v1 [cs.CY] (2008).
[13] R. Crane and D. Sornette, Proc. Natl. Acad. Sci. USA 105, 15649 (2008).
[14] M. J. Salganik, P. S. Dodds, and D. J. Watts, Science 311, 854 (2006).
[15] R. Baeza-Yates and B. Poblete, Comput. Networks 50, 1464 (2006).
[16] A. Capocci et al., Phys. Rev. E 74, 036116 (2006).
[17] V. Zlatic et al., Phys. Rev. E 74, 016115 (2006).
[18] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review 51, 661 (2009).
[19] A.-L. Barabási, Nature 435, 207 (2005).
[20] B. B. Mandelbrot, Fractals and Scaling in Finance: Discontinuity, Concentration, Risk, vol. E of Selecta (Springer, 1997).
[21] M. H. R. Stanley et al., Nature 379, 804 (1996).
[22] B. Gutenberg and C. Richter, Bull. Seismol. Soc. Am. 34, 185 (1944).
[23] D. Rybski et al., Proc. Natl. Acad. Sci. USA 106, 12640 (2009).
[24] F. Omori, J. Coll. Sci. Imp. Univ. Japan 7, 111 (1894).
[25] P. Bak, C. Tang and K. Wiesenfeld, Phys. Rev. Lett. 59, 381 (1987).
[26] S. Dorogovtsev, J. Mendes, and A. Samukhin, Phys. Rev. Lett. 85, 4633 (2000).
[27] S. Fortunato, A. Flammini, and F. Menczer, Phys. Rev. Lett. 96, 218701 (2006).
[28] See EPAPS Document No. ... for alternative reranking strategies.
[29] B. Goncalves et al., in Late-breaking results at 2nd Intl.
[30] M. Meiss et al., Proc. 21st ACM Conf. on Hypertext and Hypermedia (HT) (2010).