Traffic in Social Media II: Modeling Bursty Popularity

Jacob Ratkiewicz*‡, Filippo Menczer*†, Santo Fortunato†, Alessandro Flammini*, Alessandro Vespignani*†
* Center for Complex Networks and Systems Research,
School of Informatics and Computing, Indiana University, Bloomington, IN
† Complex Networks and Systems Lagrange Laboratory (CNLL)
ISI Foundation, Turin, Italy
‡ Corresponding author: email jpr@cs.indiana.edu

***Abstract*—Online popularity has enormous impact on opinions, culture, policy, and profits, especially with the advent of the social Web and Web advertising. Yet the processes that drive popularity in our online world have only begun to be explored. We provide a quantitative, large scale, longitudinal analysis of the dynamics of online content popularity in two massive model systems, the Wikipedia and an entire country's Web space. In these systems, we track the change in the number of links to pages, and the number of times these pages are visited. We find that these changes occur in bursts, whose magnitude and time separation are very broadly distributed. This finding is in contrast with previous reports about news-driven content, and has profound implications for understanding collective attention phenomena in general, and Web trends in particular. To make sense of these empirical results, we offer a simple model that mimics the exogenous shifts of user attention and the ensuing non-linear perturbations in popularity rankings. While established models based on preferential attachment are insufficient to explain the observed dynamics, our stylized model is successful in recovering the key features observed in the empirical analysis of our systems.**

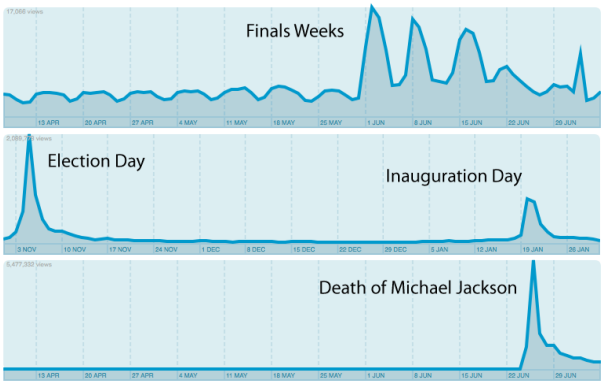


Fig. 1. Comparison between the temporal traffic patterns of three different Wikipedia topics, visualized by wikirank.com. 'Biology' (top) displays a predictable weekly cycle, as well as peaks in demand around final exam weeks. 'Barack Obama' (center) and 'Michael Jackson' (bottom) are instead dominated by exogenous news events.

I. Introduction

The advent of Web 2.0 and social media is fostering Web-mediated brokers such as blogs, wikis, folksonomies, and search engines, through which anyone can easily publish and promote content online. Popular sources have formidable power to impact opinions, culture, and policy. We know that Web traffic, one measure of popularity, is very broadly distributed [1]. Yet the dynamics that drive popularity in our online world are still unclear and largely unexplored, with a few exceptions discussed in the next section. Today, the availability of data about large Web collections with their temporal history makes it possible for the first time to study the dynamics of online popularity at the global system scale.

Here we begin to address these questions by studying popularity trends in the Web and Wikipedia. For example, different Wikipedia topics have different traffic behaviors, as illustrated in Figure 1. We would like to devise an analytical framework to look for regularity across such diverse patterns of change.

Better understanding of these trends could have broad applications. One obvious area is in improving the efficiency of the online advertisement market. Advertisers and advertisement providers, armed with knowledge of what might be popular in the coming days or weeks, would be able to bid more efficiently for advertisement keywords. Another area of interest is that of detecting Web spam. Some modern methods for this task involve analysis of the structure of the Web link graph [2]. Viewing the degree of a page as a popularity measure allows us to better analyze how this graph grows and changes, helping us determine what is normal and what is suspect. Such analysis could also aid the detection of advertisement click fraud. Here, equipped with good characterizations of normal traffic patterns for advertisement sites, the goal would be to better detect anomalous patterns of access. Finally, better understanding of traffic behavior could help methods for ranking pages that depend on traffic, such as that proposed by Liu *et al.* [3].

Contributions and Outline

Here we study the dynamics of the accumulation of attention in social media and the Web at large, finding that they are characterized by wild burst events separated by extended phases in which growth is more regular. Our main contributions are summarized as follows:

- A compilation of three large data sets, with sufficient pre-processing to make them manageable. We use the logarithmic derivative to scale popularity changes in a

way that is robust with respect to different system sizes and growth patterns — §III.

- The analysis of these popularity measures and data sets, revealing bursty dynamics with broad distribution of both the size of the bursts, and the intervals between them — §IV.
- A comparison with existing growth models, which do not explain the empirical data. We therefore propose a new *rank-shift* model that fits the observed dynamics of online popularity — §V.

II. Related Work

Several studies have used crawl data to analyze the temporal evolution of the Web, focusing on creation and destruction of pages, links, and the frequency and amount of change in page content [4]. This approach, however, does not allow to track individual pages or sites longitudinally in order to accurately monitor their popularity over time. Kleinberg [5] studied the bursts associated with identifiable events in streams, such as the occurrence of a key phrase in a news feed. This approach allows to detect hot topics as temporal bursts in word usage. Kumar *et al.* [6] expanded this notion to analyze the evolution of bursty communities in blogs. They also developed the concept of time graphs, which is similar to our methodology for tracking temporal patterns of popularity.

Focusing on indegree as a popularity measure, several models have been proposed to interpret the evolution of this quantity. The best known network growth model is preferential attachment [7]. This model starts with a small random graph and iteratively adds new nodes, each linked to existing nodes with a probability that is a linear function of their indegree. An equivalent model, proposed by Kleinberg *et al.* [8], has the additional advantages of proposing a simple mechanism by which this linking behavior might come about, and not requiring page authors to have global knowledge of the network. Topological features can also be combined with content information to interpret the emergence of topical locality in the Web [9]. Other recent developments in modeling the growth of graphs are two models proposed by Leskovec *et al.*. The first is a triangle-closing model for social networks [10], and the second is the "forest fire model" which attempts to capture the changing density and diameter of growing graphs [11]. These are members of the "rich-get-richer" class of models, to which the classical preferential attachment model also belongs.

In rich-get-richer models, it is extremely rare for any node to radically change its ranking. We will see that in the systems we intend to model, pages can rapidly acquire disproportionate attention. Therefore, the model we propose in this paper builds on the *ranking model* by Fortunato *et al.* [12], which allows us to represent ranks explicitly. This simple model grows a network by iteratively adding nodes, and connecting a new node to some existing nodes i with probability $p(i) \sim r_i^{-\delta}$; here, r_i is the rank of node i as determined by some arbitrary ranking mechanism. When the rank of a node is determined by its indegree, this model is a robust generalization of preferential attachment; it is also robust with respect to other choices of the ranking mechanism and incomplete rank information.

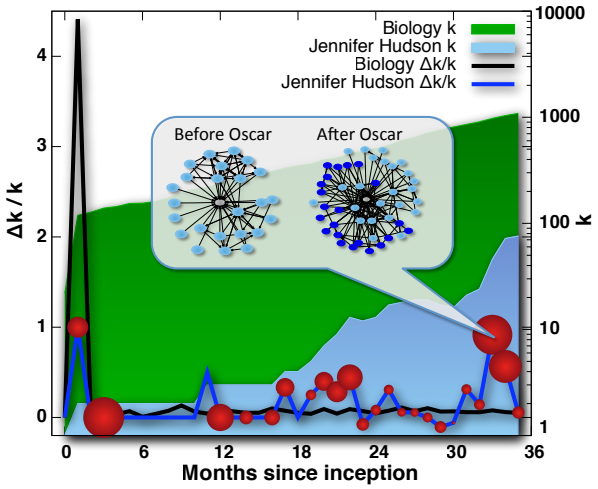


Fig. 2. Time series of indegree k and its logarithmic derivative $\Delta k/k$ for two Wikipedia topic pages. Topics typically experience a burst in their early life. The 'Biology' page then maintains a small rate of growth. The article about Jennifer Hudson, however, experiences more fluctuations later in its life. Jennifer Hudson is an artist who became popular through a television show leading to her first burst. Another burst occurred when she won an Academy Award; degree popularity doubled as many other pages linked to the article (inset). Another popularity measure is also shown for the 'Jennifer Hudson' page; the size of each circle is proportional to the logarithmic derivative of the number of times the article is revised. The article receives more edits when it attracts more links.

TABLE I
Dataset summary statistics.

	Vertices	First	Last	Temporal Resolution
Wiki k	3 293 102	Jan 2001	Mar 2007	1 sec.
Wiki s	3 490 740	Feb 2008	Current	1 hour
Chile k	3 252 779	2001	2006	1 year

Some prior work on the topic of popularity dynamics has focused on news. Wu and Huberman [13] performed a large-scale study of the news sharing site Digg.com, where users can promote links to articles they like by voting for them. The authors tracked the total number of votes that each story receives through its lifetime, finding that this quantity follows a lognormal distribution. They further examined the decay rate of incoming votes for a story, providing insight into how a story's relative popularity waxes and wanes over its lifetime. When we broaden our range in considering any online Web page or topic, the distributions of the popularity measures we study on the Web and Wikipedia — node indegree and traffic — fit a power law much better than a lognormal (as discussed in §IV). This is indicative of the distinct underlying dynamics in each of these systems. In the information networks we study, the popularity of a topic may be influenced by many news events over an indefinite timespan. Therefore the behavior of online popularity cannot in general be characterized by that of individual news-driven events.

Other recent work on human activity in the Web at large has focused on search engines [14] and Web traffic [1]. In the latter study, Meiss *et al.* find that the distribution of the traffic directed at hosts on the internet is very broad, well fit by a power law with exponent less than 2. Thus, it is not meaningful to consider the "average" popularity of a Web host. These findings were confirmed in a later study on a different dataset [15].

The popularity of videos on the YouTube video sharing site has been studied by Szabo and Huberman [16] and Crane and Sornette [17]. These dynamics are found to be similar to those of news, but with different popularity classes depending on whether a video has been featured on the front page of the site, or is the type that is likely to be spread by social networks (a so-called *viral* video).

In a companion paper, we explore various aspects of traffic patterns through social information networks [18]. We find that many bursty Wikipedia topics exhibit a strong correlation with appropriately chosen queries on Google Trends, suggesting that these bursts are often driven by external events.

All of these studies suggest that the dynamics of information access and popularity follows a bursty, intermittent behavior. It is unclear how this affects the global processes ruling the popularity accumulation and evolution in large scale information systems [19], [20]. Several of these studies show that when users have access to popularity rankings (e.g. YouTube views or presence of a book on the New York Times bestseller list), they are more likely to disproportionately favor popular items [17], [20], [21].

An initial issue facing a study of the sort presented in this paper is the identification of a suitable popularity measure. In recent years, the mapping of large, complex information networks [22]–[24] has led to identifying the number of links pointing to a node (its *indegree*) as a proxy of popularity in many domains. The evidence that many social, technological, and information networks are characterized by stable heavy-tailed distribution of indegree pointed to a strong heterogeneity in the popularity and triggered the formulation of models aimed at explaining the emergence of such broad distributions using rich-get-richer mechanisms [25] based exclusively on topology [7], [8] or combined with content information [9]. While these models have the merit of introducing irreversible growth as an important element of network generation, the dynamics characterizing these rapidly changing systems have been seldom studied because to date it has been infeasible to observe the actual growth of an online network. The datasets we utilize, however, contain longitudinal information that makes it possible to observe their growth. Further we have access to traffic data, which we consider a more direct proxy to popularity as it represents human attention more immediately.

III. Methodology

We analyze three large scale data sets about two information networks for which it is possible to gather longitudinal information: the entire Wikipedia and the Chilean Web. Wikipedia is a large collaborative online encyclopedia. Since its inception in 2001 the English version, on which we focus, has grown to contain millions of articles and hundreds of thousands of registered contributors (en.wikipedia.org). The availability of the full edit history of every article makes it possible to reconstruct the entire Wikipedia structure at any past point in time, and track it at any temporal granularity. Our data was obtained from a March 2007 dump (download.wikimedia.org). Traffic data with hourly temporal resolution is obtained by cross-referencing with a separate data set (dammit.lt/wikistats). Our third data set is a yearly sequence of crawls of the Chilean Web. This data was made available by courtesy of the TodoCL search engine (www.todocl.com), and consists of one complete crawl of the .cl top-level domain for each of the years 2002–2006.

The representative graphs of each of these data sets have been previously studied, and found to have the important properties of the Web graph at large [26]–[28] — namely, each shows a scale-free distribution of degree. Analyses of the temporal dynamics of the Wikipedia have further shown that the number of articles, their updates, visitors, and registered editors have been growing exponentially until recently, and the indegree distribution has stabilized very early to the same power law as the Web at large [29], [30].

In both data sets we track the time evolution of the indegree k of documents. In Wikipedia the high temporal resolution allows us to analyze this measure as a function of real time or age since the creation of a page, and using different timescales — e.g. months, weeks, or days — over the entire edit history. For the Chilean Web we can track the indegree with the time resolution of a year. In Wikipedia we also track the number of times s that an article is actually visited; traffic is a more direct measure of the interest generated by each topic.

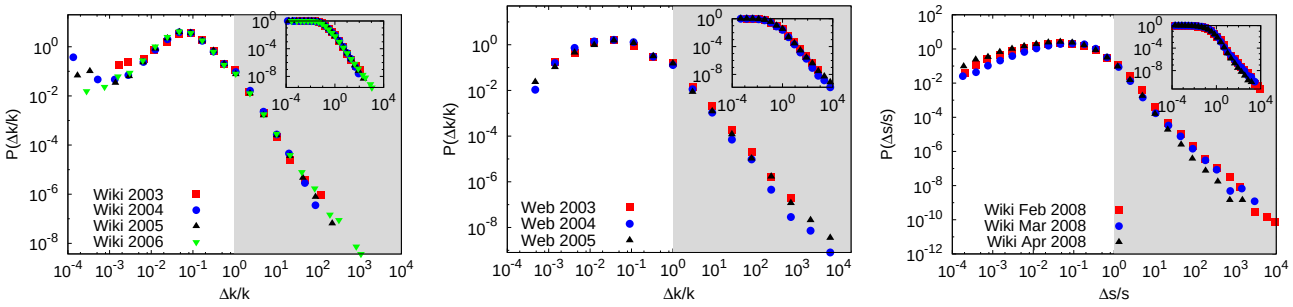
With each of the three data sources — Chilean Web, Wikipedia articles, and Wikipedia traffic counts — we produced a matrix in which the rows correspond to nodes and columns to timestamps, with each entry in the matrix referring to the value of the popularity measure for that node and timestamp. For the Chilean Web and Wikipedia articles this popularity measure is indegree (k); for Wikipedia traffic it is incoming traffic, s. Details on this derivation are below, and some vital statistics on each dataset are shown in Table I.

A. Chilean Web

This data comprises a complete crawl of the Chilean Web for each of the years 2002–2006. Its compilers have done further processing including the computation of indegree. Thus, computing the matrix was a simple matter of extracting the indegree for each page from each crawl from a proprietary format [31].

B. Wikipedia articles

This data is provided as a single compressed XML file; we obtained the English Wikipedia dump for March 2007. To our knowledge more recent dumps do not exist due to the difficulty of creating them. This file contains the full text of every revision for every page, as well as some metadata, including



(a) $\Delta k/k$ for Chilean Web indegree, with temporal resolution of one year.

(b) $\Delta k/k$ for Wikipedia indegree, with temporal resolution of one month, as measured in January over several years.

(c) $\Delta s/s$ for Wikipedia traffic, with temporal resolution of one week, as measured over a few months in 2008.

Fig. 3. Distributions of logarithmic derivative of popularity. The gray areas highlight the broad tails of the distributions. These behaviors are consistent across a wide range of temporal resolutions, from a week to a year. The inset of each figure shows the cumulative distribution, $P(X > x)$. In the remainder of the paper, for ease of exposition we show only log-binned PDFs, though cumulative distributions do not change the results.

the time and the author of the revision. Uncompressed, it is about 1.3 TB in size. We first computed the outlinks of each page at each relevant timestep, as determined by its most recent revision at that point. We then used these sets of outlinks to compute each page's indegree at each point in time. We ignored meta pages, such as discussion, user, and image pages. We further re-mapped links around redirect pages so that a link to a redirect page counts as a link to its actual target.

While it has been observed that the growth of the Wikipedia has slowed of late, our data refers to a time period in which the English Wikipedia as a whole was growing exponentially (e.g. in number of topics).

C. Wikipedia traffic

This data set comes from a person involved in the Wikipedia project who has been logging hits to the Wikipedia proxy server. The data is formatted as compressed text files, one for each hour, which record tuples of (*language code, article title, count*). Thus, computing traffic for English Wikipedia pages was a simple matter of excluding those tuples which referred to non-English pages, or to irrelevant URLs. This pruning was accomplished in the case of non-English pages by considering the language code. As for Wikipedia articles, pages with a colon in their title were removed. We further aggregated the data to obtain a temporal resolution of one day. The collection of this data began in February 2008, almost a year after the end of the Wikipedia article dataset. This makes it impossible to directly compare a page's in-strength with its indegree for the same time period.

D. Measures

To quantitatively study the dynamics of any time dependent popularity measure x_t, it is convenient to consider its *logarithmic derivative* $[\Delta x/x]_t = (x_t - x_{t-1})/x_{t-1}$, where t refers to units of time. This allows us to compare the dynamics of pages with different popularity while discounting the overall growth of the underlying system, which is not uniform across data sets. Figure 2 illustrates the logarithmic derivative of the indegree of two example pages in the English Wikipedia.

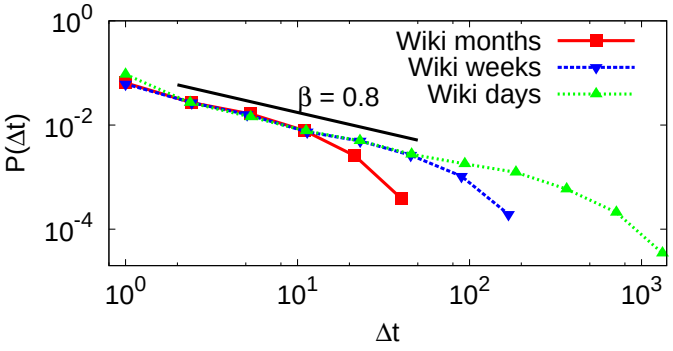


Fig. 4. Distribution of the time interval Δt between consecutive indegree bursts of Wikipedia articles. The three curves correspond to different time resolutions of months, weeks, and days, aligned on the x-axis for ease of visualization. As we increase the resolution the tail of the distribution extends further, an indication that the cutoff is a finite size effect. As a guide to the eye we show a power law $p(\Delta t) \sim (\Delta t)^{-\beta}$ for $\beta \approx 0.8 \pm 0.1$, taken from a maximum-likelihood fit to the data.

Despite a roughly exponential growth in the popularity of both topics, the logarithmic derivative provides a signature by which the two profiles can be compared on the same scale. Almost all pages experience a burst in $\Delta x/x$ near the beginning of their life,[1] and many 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. While the logarithmic derivative can be negative — indicating a decrease in popularity — we neglect these rare events and focus on the positive values.

IV. Results

As a first step, we confirmed scale-free distributions of popularity in our Wikipedia data, finding each (both indegree and traffic) to be well modeled by a power-law distribution, with the obvious exponential cutoff determined by the finite size of the network. This was done by using maximum

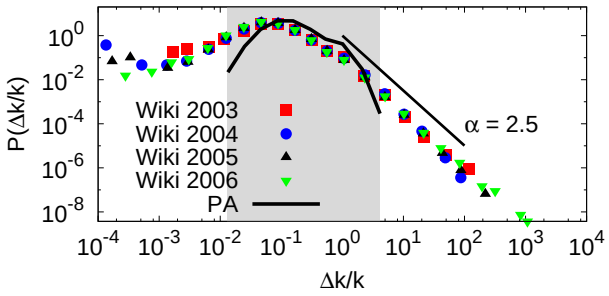
[1]We ignore the initial step of a page's life, where $x = 0$ and $\Delta x/x$ would be undefined.

likelihood methods [32], checking the Kolmogorov-Smirnoff statistic to rule out a lognormal model. This is in agreement with other studies for the Wikipedia, and with results for the Web at large. We know from Baeza-Yates and Poblete [26] that this is also true in the Chilean Web data we study. We next turn our attention to the distribution of the log derivative of popularity $\Delta x/x$.

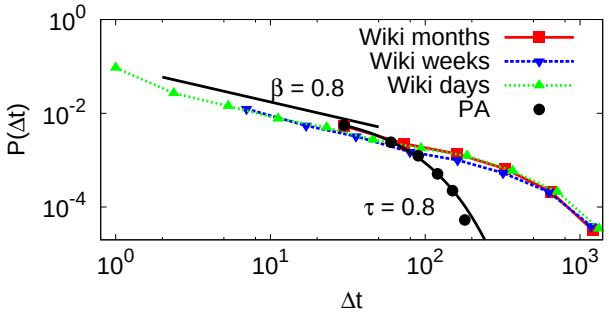
The distribution of the magnitude of $\Delta x/x$ for the two popularity measures at representative time resolutions is illustrated in Figure 3. All curves provide evidence for a wide variability of the burst magnitude that spans 8 orders of magnitude. In all cases and at all granularity it is possible to observe a heavy-tail behavior for the occurrence of large magnitude events. The observed long tails are stable and fairly well approximated by a power law $p(\Delta x/x) \sim (\Delta x/x)^{-\alpha}$ with exponent α between 1.9 and 2.6 estimated by maximum likelihood methods [32]. This indicates that a statistically appreciable fraction of events corresponds to increases in popularity by factors of 10–10^3 or more. Such a disproportionate jump of interest occurs not only for young or lesser known pages, but for a broad range of popularity, which we confirmed by considering the distribution of degree for for pages about to experience a burst, before the burst occurred. In all three datasets these distributions were broad; while they were not as broad as the overall degree distribution, we can conclude by this that even pages with large indegree can still experience dramatic changes. Yet additional evidence is that when pages below a certain age (e.g. 3 months) are ignored, the distributions in Figure 3 are unchanged. In other words, these popularity spikes are statistically possible for all documents almost independently of their popularity.

The heavy-tailed burst magnitude distributions suggest a dynamics characterized by the lack of a typical scale for measuring the bursts. This is typical in a wide range of "critical" physical, economic, and social systems, such as avalanches, earthquakes, and stock market bubbles and crashes [19], [33], [34], but it had never before been observed in information networks.

Another way to characterize the dynamics of bursty systems is to study the distribution of times between successive events. In traditional systems where this behavior is modeled by queueing theory, we expect this distribution to be Poissonian. On the other hand, systems that lack a typical scale in the event size are generally associated with a lack of characteristic time scale and to long-range time correlation among consecutive events. To test the presence of non-Poissonian dynamics we analyzed the time distribution between bursts, shown in Figure 4 for the English Wikipedia. We consider bursts such that $\Delta k/k > 1$ after January 1st, 2003. This necessarily includes pages which undergo smaller bursts (in absolute terms); e.g. pages whose popularity measure goes from 1 to 2. However, we observed that thresholding did not change the statistical properties of burst events — recall that even pages with high popularity can experience large bursts. The intervals between bursts are broadly distributed in a power-law fashion with a finite size cutoff, as in Omori's law of earthquakes and other avalanche phenomena [35]. The analogy between online popularity dynamics and critical avalanche phenomena calls for a stylized model able to explain the observed features in terms of shifts in collective attention.



(a) Empirical distribution of $\Delta k/k$ from the English Wikipedia, together with analogous data as generated by a preferential attachment model. The long tail of burst sizes, highlighted by a power-law guide to the eye taken from a maximum likelihood fit, is missed by the PA model, which generates data only in the gray area.



(b) Empirical distribution of time between bursts Δt (in normalized units) in the English Wikipedia, together with the distribution generated by a preferential attachment model. This is the same data as in Figure 4, and the fact that the lines overlap when using normalized time units (except for finite time cutoff) supports the power-law fit (cf. Figure 4). However, the PA distribution fits an exponential $P(\Delta t) \sim e^{-\Delta t/\tau}$ with parameter $\tau = 0.8$.

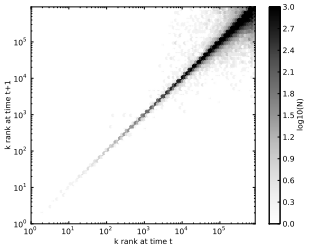
Fig. 5. Comparison of the empirical data with what would be expected from a preferential attachment process. The PA process fails to produce wide distributions of event size and temporal spacing.

V. Modeling Popularity Trends

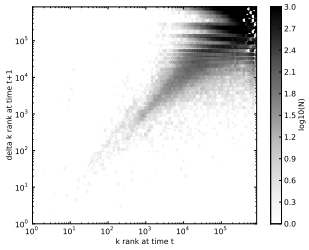
A. Preferential Attachment

Among the many growth models in the general family of preferential attachment, we chose the directed version [36] of the linear preferential attachment model [7] as a baseline, and used it to generate a graph. This rich-get-richer mechanism does produce graphs with the same degree distribution as in our data sets; however, preferential attachment alone fails to reproduce the long tails observed in the distributions of both $\Delta k/k$ and inter-burst time (Figure 5).

Another way to explore the limitations of PA in explaining the observed dynamics is to visualize the relationship between the rank of a node's indegree k at a given time step, and its behavior in the time step that follows. In Figure 6(a) we show a scatter plot comparing a node's rank in k at time t with its rank in k at time $t+1$. Figure 6(b) similarly shows the



(a) Rank of k at a representative time t vs. rank of k in the subsequent timestep.



(b) Rank of k at a representative time t vs. rank of Δk in the subsequent timestep.

Fig. 6. Scatter plots visualizing changes in rank of k and Δk between timesteps. In preferential attachment, the ranking based on degree (and change in degree) should be nearly static between timesteps, producing points only on the main diagonal of the plots. The presence of points well off the main diagonal indicates change not well correlated with present degree.

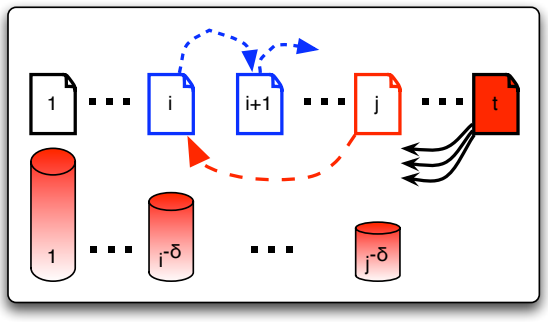


Fig. 7. Illustration of the rank-shift model in an example where popularity is measured by indegree. New nodes are added at each timestep, illustrated by the node t; each node's probability of receiving a new link is proportional to their rank. In the diagram the node j is being reranked, pushing down the ranks of $i, i+1, ...$

```
Given real δ, ρ and ranking function r(),
    desired number of nodes N
for t in 0 .. N do
    # Growth step
   Create new node t
   Assign links from t to existing nodes k,
       with P(k) ~ r(k)^-δ.
    # Reranking step
    for each k (r(k) = j), with probability ρ,
      choose random i < j and set r(k) = i
       for r(ℓ) in i .. j do
          set r(ℓ) = r(ℓ) + 1
       end
    end
end
```

Fig. 8. Pseudocode for the rank-shift model. In the case of traffic, instead of assigning links to existing nodes, we simply increment their traffic counts.

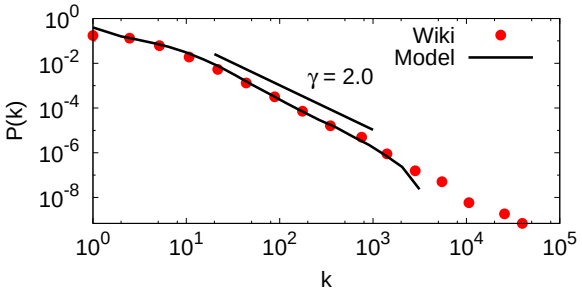
relationship between a node's rank in k at time t and its rank in Δk at time $t+1$. This visualization suggests that the empirical data has an underlying preferential attachment component, but with a strong chance of large changes, especially for nodes of lower degree.

B. A Rank-Based Model

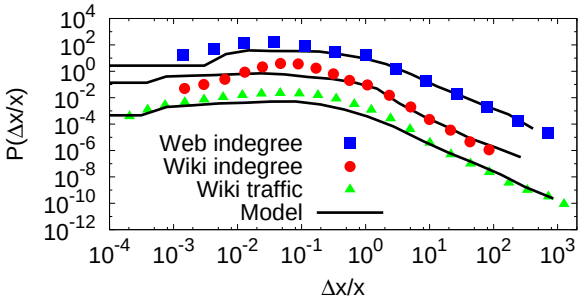
Seeking a very simple model able to capture the critical dynamics observed empirically, we note that the accumulation of attention is not obviously related to the exact degree of a document, information that is seldom available. Popularity is instead likely related to the relative ranking that is always established by users according to some criterion: age, degree, relevance to a user query (if the nodes are Web pages), or some arbitrarily distributed prestige function. We consider a generalization of the *ranking model* [12] where items are sorted according to some popularity criterion and accumulate units of popularity such that the probability that an existing item i receives a unit is $p(i) \sim r_i^{-\delta}$, where r_i is the rank of i according to some arbitrary ranking function, and $\delta > 0$ is a free parameter. This simple model leads in the asymptotic limit to scale-free popularity distributions $p(x) \sim x^{-\gamma}$, where $\gamma = 1 + 1/\delta$. The behavior is very robust with respect to choices of the ranking criterion and of the exponent δ. We stress that, since popularity is distributed based on the ranks of the nodes, and not on their popularity values, the ranking model does not belong to the class of fitness-based models [37], [38].

In our study of Web and Wikipedia pages we focus on the statistics of extreme events, represented by popularity "bursts." We define a burst as a variation of popularity Δx (within a given time window) larger than the original popularity value x of the page, i.e., an event with logarithmic derivative $\Delta x/x > 1$. The distribution of the time elapsed between consecutive bursts of the same node has a Poissonian decay for the ranking model, at variance with our empirical observations. Therefore, a modification of the model must be devised. Pending further study, we want our model to be agnostic to the actual cause of the bursts; in real data, they could be caused by external events (such as interest sparked by news stories) or due to other dynamics of the system (recall Figure 1). For now, observe that the net effect of such a burst for the node in question is to change its popularity rank with respect to the other nodes in the system. Therefore, let us introduce *rank shifting* in the model: at each iteration, with a small probability ρ each node is assigned a new rank, chosen uniformly between 1 and its current rank, simulating a sudden increase in the attention paid to the node. (Figure 7). Thus our *rank-shift* popularity model has two parameters: δ regulating the probability of accumulating popularity as a function of rank, and ρ defining the frequency of rank perturbations for each page. These sudden improvements of the rank lead to abrupt variations of popularity, as observed in the empirical data.

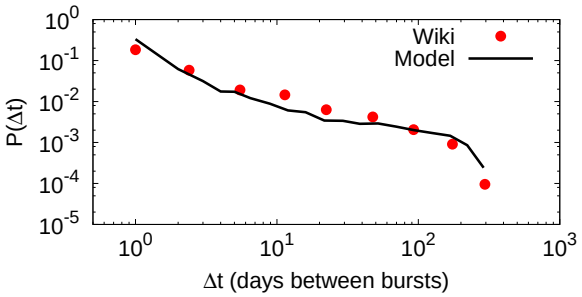
The model works as follows. Each node is assigned an arbitrary position in an initial ranking. Then two steps are performed iteratively. First, a new node t is added, and linked to existing nodes according to their rank; a node with rank r receives a link with probability $p(r) \sim r^{-\delta}$. Second, with probability ρ, each node is *reranked,* i.e. moved to a new position toward the front of the list. The new position i is



(a) Agreement between the empirical indegree distribution from the December 2003 Wikipedia and the popularity distribution produced by the model. Both curves are consistent with a power law $P(k) \sim k^{-\gamma}$ with $\gamma \approx 2.0 \pm 0.2$. Results for traffic are similar.



(b) Agreement between the empirical popularity burst distributions and those produced by the model (data from 2003 for Chilean Web indegree, December 2003 for Wikipedia indegree, and a week in February 2008 for Wikipedia traffic). The curves are shifted for illustration purposes.



(c) Agreement between the Wikipedia inter-burst time distribution and that produced by the model (data from the entire year 2003).

Fig. 9. Comparison of the empirical data with the predictions of the rank-shift model. The model is able to replicate the dynamics observed.

chosen randomly with uniform distribution between 1 (the top position) and the node's current rank j, thus focusing on positive bursts (see Figure 8 for pseudocode). The node previously occupying position i is moved back to $i+1$, and so on. Simulations of this model were performed using the empirical number of nodes N (cf. Table I), and various values of the parameters ρ and δ. The effect of varying these parameters is discussed next.

C. Evaluation of Model

For $\rho = 0$ we recover the original ranking model, and the distribution of $\Delta x/x$ (for instance, the traffic or indegree of nodes) matches that of preferential attachment (cf. Figure 5(a)). This describes the behavior of the many topics that do not undergo sudden, large bursts of attention. These dynamics are reflected in the lognormal portion of the burst magnitude distributions (cf. Figure 3).

For $\rho > 0$ numerical simulations show that the tail of the popularity burst magnitude distribution shifts from a lognormal to a power law while the popularity distribution remains a power law; its exponent remains $\gamma = 1 + 1/\delta$, with an exponential cutoff now depending on ρ. This modification allows the model to capture the dynamics of topics undergoing large bursts of attention. This behavior is manifest empirically in the broad tails of the burst magnitude distributions, which cannot be explained by preferential attachment alone (cf. Figure 5(a)).

Given the relationship between δ and the exponent γ of the indegree distribution (discussed above), we chose $\delta = 1/(\gamma - 1)$ using the empirical γ, finding $1 \leq \delta \leq 1.2$ for our data. We then numerically estimated a value of ρ in order to fit the distribution of $\Delta x/x$, and found $10^{-5} \leq \rho \leq 10^{-3}$. With these parameters, our simple model is able to reproduce many of the critical features observed in the empirical data. Not only does it predict the distributions of both popularity measures for both data sets (Figure 9(a)), but also the long tail of the distributions of indegree and traffic burst size (Figure 9(b)). Further, the model also captures the long-range distribution of inter-burst intervals (Figure 9(c)). The 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. This is sufficient to reproduce the broad distributions in the magnitude of bursts and in their temporal dynamics.

VI. Conclusions

This work analyzes popularity growth across systems of varying scales by using the logarithmic derivative of popularity proxies. We applied this approach to three large data sets, revealing non-trivial and consistent popularity dynamics that they share; namely, that a huge number of attention bursts of large magnitude, akin to booms in financial markets, occur daily in the online world. Such wild dynamics of popularity are not entirely captured by existing rich-get-richer models. They are also different from those observed in news-driven events [13], where attention fades rapidly and overall popularity is lognormal-distributed. We further propose a rank-shift model which outperforms preferential attachment models for capturing these dynamics.

Further analysis is needed to better characterize the long-term temporal correlations among bursts of individual pages and ascertain whether any topic or only certain types of topics are capable of experiencing large surges of attention even after being dormant for a long time. Our model focuses only on the occurrence of large events; with time, it could be combined with models that seek to replicate other features of the resulting graph, such as the *forest fire* model [11]. Though impossible with the English Wikipedia, as the time ranges for relevant data do not overlap, correlation of indegree and strength over time might yield interesting results for

those Wikipedia languages for which dumps are still available. Predictability is emerging as a key question [16]. Future efforts will also extend the present analysis to other measures of popularity available in the Wikipedia — number of revisions and number of unique editors — as well as to other data sets. The present findings are neutral about what drives the popularity bursts: it could be search engines, external events, news, word of mouth, social media, marketing campaigns, or any combination of them. In a companion paper, we show that external events are certainly one important factor [18]. Further study correlating traffic from multiple sources, such as Twitter and news feeds, could shed further empirical light on this question.

VII. Acknowledgments

We thank Ricardo Baeza-Yates for facilitating our access to the TodoCL data set, Virgil Griffith for his assistance in our (unsuccessful) attempts to obtain data from the Internet Archive, Bharat Dravid for his early help obtaining the Wikipedia data set, and Mark Meiss for collecting traffic data. We are also grateful to Vittorio Loreto, Ciro Cattuto, and Massimo Marchiori for useful discussions. 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., the ISI Foundation, and the Indiana University School of Informatics and Computing. S.F. gratefully acknowledges ICTeCollective, grant number 238597 of the European Commission.

References

[1] M. Meiss, F. Menczer, and A. Vespignani, "On the lack of typical behavior in the global Web traffic network," in *Proc. 14th International World Wide Web Conference*, 2005, pp. 510–518.

[2] L. Becchetti, C. Castillo, D. Donato, R. Baeza-Yates, and S. Leonardi, "Link analysis for web spam detection," *ACM Trans. Web*, vol. 2, no. 1, pp. 1–42, 2008.

[3] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li, "Browserank: letting web users vote for page importance," in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2008, pp. 451–458.

[4] A. Ntoulas, J. Cho, and C. Olston, "What's new on the Web? The evolution of the Web from a search engine perspective," in *Proc. 13th Intl. Conf. on World Wide Web*. ACM Press, 2004, pp. 1–12. [Online]. Available: http://doi.acm.org/10.1145/988672.988674

[5] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2002.

[6] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," in *Proc. 12th International World Wide Web Conference*, 2003, pp. 568–576.

[7] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[8] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The Web as a graph: Measurements, models and methods," *Lect. Notes Comp. Sci.*, vol. 1627, pp. 1–17, 1999.

[9] F. Menczer, "The evolution of document networks," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 5261–5265, 2004.

[10] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2008, pp. 462–470.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 2, 2007.

[12] S. Fortunato, A. Flammini, and F. Menczer, "Scale-free network growth by ranking," *Phys. Rev. Lett.*, vol. 96, no. 21, p. 218701, 2006.

[13] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 45, pp. 17 599–17 601, 2007. [Online]. Available: http://www.pnas.org/content/104/45/17599.abstract

[14] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Topical interests and the mitigation of search engine bias," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 34, pp. 12 684–12 689, 2006.

[15] M. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, "Ranking web sites with real user traffic," in *WSDM '08: Proceedings of the international conference on Web search and Web data mining*. New York, NY, USA: ACM, 2008, pp. 65–76.

[16] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," arXiv:0811.0405v1 [cs.CY], Tech. Rep., 2008.

[17] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 41, pp. 15 649–15 653, 2008. [Online]. Available: http://www.pnas.org/content/105/41/15649.abstract

[18] J. Ratkiewicz, A. Flammini, and F. Menczer, "Traffic in social media I: paths through information networks," under review.

[19] A.-L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, pp. 207—211, 2005.

[20] Z. Dezso, E. Almaas, A. Lukacs, B. Racz, I. Szakadat, and A. Barabasi, "Dynamics of information access on the Web," *Phys. Rev. E*, vol. 73, p. 066132, 2006.

[21] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, vol. 311, no. 5762, pp. 854–856, 2006.

[22] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the World Wide Web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[23] A. Broder, S. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the Web," *Computer Networks*, vol. 33, no. 1–6, pp. 309–320, 2000.

[24] M. A. Serrano, A. Maguitman, M. Boguna, S. Fortunato, and A. Vespignani, "Decoding the structure of the WWW: A comparative analysis of Web crawls," *ACM Trans. Web*, vol. 1, no. 2, p. 10, 2007.

[25] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, no. 3/4, pp. 425—440, 1955.

[26] R. Baeza-Yates and B. Poblete, "Dynamics of the Chilean web structure," *Comput. Netw.*, vol. 50, no. 10, pp. 1464–1473, 2006.

[27] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, "Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia," *Phys. Rev. E*, vol. 74, no. 3, p. 036116, 2006.

[28] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet, "Wikipedias: Collaborative web-based encyclopedias as complex networks," *Phys. Rev. E*, vol. 74, no. 1, p. 016115, 2006.

[29] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi, "Temporal analysis of the Wikigraph," in *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 45–51.

[30] R. B. Almeida, B. Mozafari, and J. Cho, "On the evolution of Wikipedia," in *Proc. Int. Conf. on Weblogs and Social Media*, March 2007.

[31] P. P. Calada, "The WBR-99 collection," Departamento de Computação, Universidade Federal de Minas Gerias, Tech. Rep. [Online]. Available: http://www.linguateca.pt/Repositorio/WBR-99/wbr99.pdf

[32] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," arXiv:0706.1062v1 [physics.data-an], Tech. Rep., 2007.

[33] B. B. Mandelbrot, *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, ser. Selecta. Springer, 1997, vol. E.

[34] B. Gutenberg and C. Richter, "Frequency of earthquakes in California," *Bull. Seismol. Soc. Am.*, vol. 34, pp. 185–188, 1944.

[35] F. Omori, "On the after-shocks of earthquakes," *J. Coll. Sci. Imp. Univ. Japan*, vol. 7, pp. 111–200, 1894.

[36] S. Dorogovtsev, J. Mendes, and A. Samukhin, "Structure of growing networks with preferential linking," *Phys. Rev. Lett.*, vol. 85, no. 21, pp. 4633–4636, 2000.

[37] G. Bianconi and A.-L. Barabási, "Bose-Einstein condensation in complex networks," *Phys. Rev. Lett.*, vol. 86, no. 24, pp. 5632–5635, Jun 2001.

[38] M. Boguna and R. P. Satorras, "Class of correlated random networks with hidden variables," *Phys. Rev. E*, vol. 68, p. 036112, 2003.