

Investigating and Forecasting User Activities in Newsblogs: A Study of Seasonality, Volatility and Attention Burst

Christian Bauckhage, César Ojeda and Rafet Sifa
Fraunhofer IAIS
St. Augustin, Germany

Christian Backhage and Rafet Sifa
University of Bonn
Bonn, Germany

Abstract—The study of collective attention is a major topic in the area of Web science as we are interested to know how a particular news topic or meme is gaining or losing popularity over time. Recent research focused on developing methods which quantify the success and popularity of topics and studied their dynamics over time. Yet, the aggregate behavior of users across content creation platforms has been largely ignored even though the popularity of news items is also linked to the way users interact with the Web platforms. In this paper, we present a novel framework of research which studies the shift of attentions of population over newsblogs. We concentrate on the commenting behavior of users for news articles which serves as a proxy for attention to Web content. We make use of methods from signal processing and econometrics to uncover patterns in the behavior of users which then allow us to simulate and hence to forecast the behavior of a population once an attention shift occurs. Studying a data set of over 200 blogs with 14 million news posts, we found periodic regularities in the commenting behavior. Namely, cycles of 7 days as well as 24 days of activity which may be related to known scales of meme lifetimes.

I. INTRODUCTION

Much recent research on Web analytics has concentrated on developing theories of collective attention where the main object of study is the evolution of the popularity of topics, ideas, or sets of news [1], [2]. In this context, the concept of a *meme* has arisen as the main *atom* of modern quantitative social science [1], [3] and researchers seek to understand whether a particular meme will remain popular and, if so, for how long. Under this line of research, virality is the main phenomenon to model. Usually, a contagious (i.e. network based) approach is followed, and virality is literally treated as an infection in a given population: as an item becomes popular, i.e. as a piece of news is retweeted or discussed in the media, the population unit, whether blog or twitter account, is considered infected.

It is important to note that, this paradigm of research, ignores the impact of the source of the meme. The website or content generating media plays a key role in generating (or hindering) the evolution of a particular idea or topic. Naturally, if a given website has a large number of users, it is likely, that certain topics become popular and capture the attention of the population. We might thus ask whether the baseline behavior of the users of a website is enough to generate popularity. Is it the content which is popular or is it the website which is popular? If we can use the baseline behavior of a given

website as a reference, we might use such information as a general guideline for forecasting. Information regarding the population behavior over a website will provide the basis for understanding both the evolution of that particular website as well as the possible success of the future content.

The population behavior on a given website is unavoidably stochastic, as we cannot know a priori whether a particular user or a set of users will visit or not (see Fig. 1 for an exemplary time series of activities of blog commenters). Statistical time series analysis has a rich history of success in fields as diverse as electronics, computer science, and economics and we know that, given that relevant information regarding the behavior of the system is properly modeled in the phenomena that is measured, and randomness is realized under proper bounds, estimates can be achieved and predictive analytics becomes possible.

In this work we present a bottom up case study to analyze attention of blog users (particular commenters) to understand and predict their activity patterns which exploits the fact that many websites as well as blogs have years of user history which can be mined in search for relevant patterns.

II. RELATED WORK

The study of information diffusion on the Web intends to model pathways and dynamics through which ideas propagates. One of the goals is to infer the structure of networks in which information propagates [4]. Researchers often try to devise equations which govern the aggregate behavior [2]; these equations typically have parameters which depend on the population size, the rate of the spreading process, and universal features of how attention fades. They model time series which show how much activity a particular topic or meme attract. Forecasting can be then performed once the parameters of a particular population have been determined. In [5] Matsubara et al. learned the parameters of the attention time series related to the first Harry Potter movie and then predicted the behavior for the next movies by measuring only the initial population reaction. More importantly, the *natural* limitations of human attention and human behavior have shown to define the overall behavior of the population as they access the information [6], [7], [8], [9].

Blog dynamics are traditionally studied in the context of conversation trees established between different blogs [10],

number of blogs	203
number of posts	713,122
number of comments	14,883,752
time span of activity	2006-2014

TABLE I: Characteristics of the Wordpress Data Set

[11]. In this context, model of the structural and dynamical properties of networks of hyperlinks are sought for. McGlohn et al. [11] try to find representative temporal and topological features for grouping blogs together and to understand how information propagates among different blogs. The model in [10] defines rules which indicate how the different links are created. Typical properties include the degree distribution of the edges to a blog, which presents a power law behavior, the size of conversation trees (how many posts are created in a given theme) and the popularity in time, as measured with the amount of edges in time.

Here, we intend to make use of methods of statistical time series modeling in order to study and forecast the behavior of a population on a given webpage. In particular, we focus on time series of users commenting patterns. We base our work on well establish methods in economics [12], [13] and signal processing [14]. In a nutshell, we attempt to describe the statistical process which generates the time series, modeling both temporal correlations as well as signal noise. Due to the world financial crises of the years 2008 a great deal of attention in the field is devoted at understanding volatility or strong fluctuations in the market [15], [16]. At their core, the problems of volatility and collective attention shifts are similar and we exploit these similarities in our approach.

III. DATA SET DESCRIPTION

The empirical basis for our work consists in a collection of blogs hosted on *Wordpress*¹. Wordpress blogs are personal or company owned web sites that allow users to create content in the form of posts. Users (or, in blogger terminology, writers or authors) decide whether or not their readers can comment on particular posts. Comments on posts provide a significant proxy when it comes to measuring the attention a topic receives.

Wordpress is a content management system and a frequently used blogging platform accounting for more than 23.3% of the top 10 million websites as of January 2015². As of this writing, there are 43 million posts and more than 60 millions users. Notably, the user-base of Wordpress ranges from individual hobby bloggers to publishers such as Time magazine³. And while traditional mass media limit the number of topics accessible to the public [17], blogs have arisen as a *social medium* allowing for personal opinion and information variety. Nevertheless, inequality is a pervasive characteristic of social phenomena [18], [19] and many of the consumers of

¹<https://www.wordpress.com/>

²Information retrieved on 5.3.2013 from http://w3techs.com/technologies/overview/content_management/all/

³<https://vip.wordpress.com/2014/03/06/time-com-launches-on-wordpress-com-vip/>

Blog URL	News Genre	Viral Post Count
le-grove.co.uk	Sport	620
politicalticker.blogs.cnn.com	Politics	542
order-order.com	Politics	294
sloone.wordpress.com	Personal	248
technologizer.com	Technology	136
smkorean.wordpress.com	Entertainment	116
collegecandy.com	Magazine	108
seokyalways.wordpress.com	Personal	96
religion.blogs.cnn.com	Religion	93
kickdefella.wordpress.com	Personal	89

TABLE II: Top 10 Newsblogs from our dataset.

blog contents concentrate on sites with a large community of frequent users.

Our dataset contains the entire post and comment history of newsblogs (from year 2006 to 2014) that ever made it to the daily updated list of *Blogs of the Day*⁴. Having crawled the content of over 6580 different blogs, we choose to work with the time series of activities through comments and posting of the most active (throughout the available time span) 203 blogs from variety of genres in order to study long term dynamics. The blogs we analyze overall contain 713,122 posts and 14,883,752 comments. Table I summarizes the overall characteristics of our data and Table II shows the top 10 blogs sorted with respect to the number of viral posts. So as to protect the identity of commenters, we randomly hashed the author identities and key features we extracted from each post are

- CommentsTime: Time of each comment
- MainURL: URL of the post
- CommentsInfo: Content of the comment
- Title: Title of the post
- PubDate: Date of last update of the post
- Blog: URL of the blog
- Info: Summary of the Post

IV. EMPIRICAL OBSERVATIONS

Our main objective in this paper is understanding the dynamics of the aggregate behavior of a population of users of a webpage or blog. For each blog we determine the number of comments C_t on all posts at time t . Figure 1 presents a typical example of such a time series of daily commenting behavior spanning a period of 3 years together with a monthly moving average. Averaged over months, the time series shows a slowly varying behavior which reflects the fact that, over time, the attention the blog receives does not vary much. On the other hand, on the time scale of days rapid variations can be observed, as well as strong peaks which indicate the presence of particularly engaging content. We aim at applying forecasting methods to these very fluctuations. The main goal of our work is to propose a forecasting procedure that can characterize these shifts of attention.

We assume that a prototypical webpage has a steady user base. This user base can be understood as a *set of followers* and as different followers visit the website, comments will appear as natural stochastic fluctuations. However, if a particular news item is highly relevant or *hot* for a segment of the population,

⁴<https://botd.wordpress.com/>

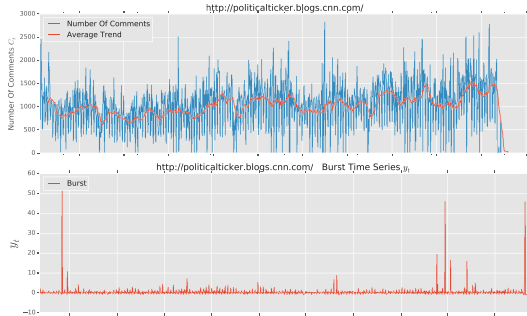


Fig. 1: Time series of the number of comments C_t and bursts y_t for a newsblog. The upper figure shows the daily fluctuations of activities (blue) and the monthly moving average trend (red). The lower figure shows a burst representation that allows for locating noticeable changes in the overall behavior.

then users will show more activity, increase their commenting rate, and also share the content with others users *which are not part of the baseline set of followers of the blog*. Consequently, some of the comments will appear as a result of a spreading process due to the web site followers. *The popularity of web page translate into popularity of the content.*

Assuming a somewhat stable set of users has the added advantage of exploiting seasonality effects, as people are known to follow seasonal behavior [20]. Such periodicities can help to better understand the repeating behavior (for instance weekly visits of average users) as well as to forecast activities.

V. TIME SERIES ANALYSIS

We model the user commenting behavior as a discrete stochastic process, i.e. as a collection of random variables $X_1, X_2, X_3, \dots, X_k$ indexed by time. Since we aim at attention modeling, our first variable of study is C_t , the number of comments an entire blog has at time t . Considering daily samples accords with the pace of publication in most blogs.

A. Seasonality

One approach to time series prediction using stochastic processes requires detecting periodicities. If are recovered by detecting their frequencies and associated amplitudes, we can predict the behavior at each cycle. A common method considers power spectra of the stochastic process at hand via the discrete Fourier transform

$$f_x(\omega) = \sqrt{\frac{1}{T}} \sum_{t=1}^T x_t \exp -i\omega t. \quad (1)$$

We then square the transform

$$I(\omega) = \frac{1}{2\pi} f_x(\omega) \hat{f}_x(\omega) \quad (2)$$

to obtain the so called periodogram of a time series.

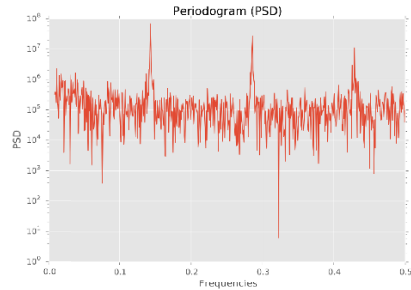


Fig. 2: Seasonality study through a semilog plot of the periodogram for estimation of the power spectral density from an example newsblog. Notice the presence of 3 frequencies among the noisy behavior that account for commenting activity in period of three-and-a-half and seven days.

If seasonal behavior is present, peaks will appear in the periodogram. For example, Fig.2 shows the periodogram for a time series of comment counts from our dataset. Notice the peaks at frequencies of 0.13, 0.29 and 0.31 which account for periods in the comment patterns of *weekly* and *half weekly* seasonal behavior indicating that the users of that particular newsblog leave comments in three-and-a-half and seven days frequencies.

B. Burst and Volatility

In order to characterize the popularity of a particular news source, we require a quantitative measurement of attention shifts. These can be described via an iter-burst measure [21], namely

$$y_t = \frac{C_t - C_{t-1}}{C_t}. \quad (3)$$

Eq. (3) is also known as the logarithmic derivative of the process C_t . Due to its rescaled nature, y_t allows for comparison between different newsblogs and times, but, for our data set, it provides limited value for forecasting. We thus also consider the detrended comment count $\tilde{C}_t$ which is obtained after using the *Baxter-King* filter (see the Appendix).

C. Modeling Volatility

Having introduced $\tilde{C}_t$, we next need an approach that can account for fluctuations. Since after the filter is applied we obtain a zero mean behavior, we can assume that the comments behavior will vary as

$$\tilde{C}_t = \sigma_t \epsilon_t \quad (4)$$

where σ_t represents the standard deviation of the fluctuation at time t and $\epsilon_t \sim \mathcal{N}(0, 1)$ a noise value at time t , sampled from a normal distribution of mean 0 and variance 1. In finance, when modeling the fluctuations in returns (as opposed to $\tilde{C}_t$) σ_t is known as the volatility. The simplest model for volatility occurs in econometrics under the names of *ARCH* (autoregressive conditional heteroscedasticity) and generalized *ARCH* or

GARCH [22]. Here, heteroscedasticity refers to variations in the fluctuation of the variable of interest, in our case C_t . Formally, this implies that the covariance $Cov(C_t, C_{t-k})$ depends on time t . The dependence is modelled through past values of the noise e_t as well as past values of σ_t

$$\sigma_t^2 = \omega + \sum_{k=1}^q \alpha_k \epsilon_{t-k}^2 + \sum_{l=1}^q \beta_l \sigma_{t-l}^2. \quad (5)$$

The values of p and q determine how correlated σ_t is with past fluctuations so that the model is called *GARCH*(p, q). To learn these parameters from data, we have to take into account that for *GARCH*(1,0) the square of the fluctuations are $\tilde{C}_t^2 = \alpha_0 + \alpha_1 \tilde{C}_{t-1}^2 + \text{error}$, i.e. that they behave as an *AR*(1) (an autoregressive model of order 1) [12]. Hence, the values of the partial autocorrelations of the square variable give us a hint of the dependence of the model.

VI. MAIN RESULTS

A. Attention Proxy

Traditional attention proxies for collective attention comprise direct measures such as the number of hyperlinks to a web site, retweets, or the number of likes via the facebook platform. Sometimes an internal measure of the website is preferred, for instance, the web site *digg.com* uses a user based ranking mechanism which ultimately decides which content is displayed on the main page. In order to validate our use of comments as an attention proxy, we make use of the fact that Wordpress regularly features the most popular posts⁵. In the following, we refer to normal posts as posts not presented in this list, whereas viral or important posts as the posts presented in the list and we note that, indeed, posts features on the most popular list attract more comments than others.

Since we work with a dataset that covers several years of activity, it is important to note that the amount of comments expected in a given year is different from the amount of comments in another if the blog shows growth or decay in its user base. Yet, this issue is accounted for by our use of *de-trended behavior* $\tilde{C}_t$.

For our whole data set of relevant blogs, we observe average fluctuation value for the important posts of 71.12 whereas for the normal blogs an average fluctuation value of -96.14. This shows that *most of the comment behavior arises from popular or viral posts*.

Figure 3 shows the distribution of comments for normal and important posts where we observe positive skew distributions that can be modeled using pareto and lognorm distributions (see the Appendix). The pareto behavior indicates that there are strong fluctuations in the normal data set, meaning that there are in fact important posts not considered by the Wordpress listing procedure.

B. Seasonality

In this section we perform the periodogram analysis to identify the cycles of commentors leaving comments in the

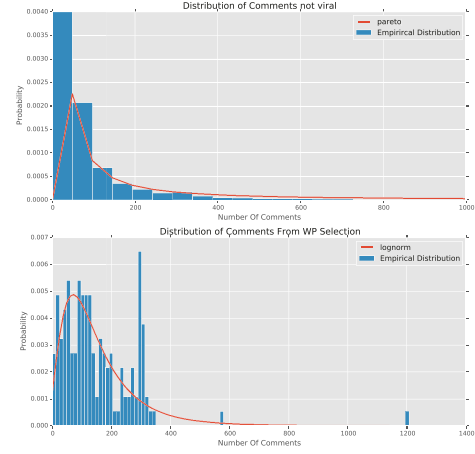


Fig. 3: The distribution of comments. The upper figure shows the distribution of comments for posts pertaining to the non-relevant data set. Pareto distribution is fitted with 43 average value and p value = 0.43. The lower figure on the other hand shows the distribution of comments for relevant posts as determined from the Wordpress data set average 254 comments per post and p value = 0.73.

newsblogs. It is important to note that such an analysis should be realized for each blog independently, as it will reveal the cycles of the followers or induced cycles due to the posting patterns of the particular website we are analyzing. Nevertheless, it is commonly known that both human behavior and virality patterns often show universal trends. For instance, attention to memes grows and fades over time in highly regular patterns [1]. Also, people show repetitive behavior because of work schedules and life habits [20]. If there is in fact such universal behavior for our case, a periodogram analysis should reveal common characteristics across different blogs.

Indeed, based on periodogram analysis, we obtained the most relevant frequencies for each periodogram as the frequencies above 4 standard deviations over the statistics of each individually calculated periodogram. Figure 4 shows the histogram of the relevant frequencies where we observe relevant peaks at frequencies of 17 and 7 days.

The peak with the largest frequency value of 7 days can be attributed to a weekly behavior of the users. Business days constrain the publishing efforts of websites as well as the visiting patterns of users. The 17-day period on the other hand, seems to correspond to the natural attention scale of publishing behavior. A news content provider for example, publishes several posts on a given topic and this 17 day frequency might correspond to the saturation of attention resources people can

⁵<https://botd.wordpress.com/top-posts/>

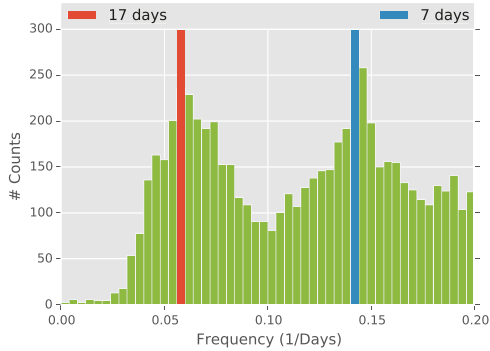


Fig. 4: Seasonality study through analyzing the histogram for the relevant frequencies over the whole population of blogs. For each blog we selected the frequencies which exceed the 4 standard deviation. We observe peaks for 7 and 17 days indicating users' periodic commenting frequencies.

allocate to different subjects [9]. Unless we perform an actual content analysis, however, we will not be able to pinpoint the true latent cause of this periodicity. Nevertheless, our results show a consistent appearance of this frequency throughout our Wordpress data set.

C. Forecasting

Having analyzed the nature of user behavior and having uncovered periodic patterns in the behavior of commentors, we next approach the ultimate goal of analytics which, in our case, is forecasting user fluctuations.

1) *A Forecasting Protocol*: To this end, We propose the following protocol in order to study the fluctuations in attention to blogs:

- Create the time series of comments per day C_t (different time scales such as hours, days, or weeks may be considered, provided that enough samples are available).
- Apply a detrending procedure such as the Baxter King filter in order to obtain a stationary version of the given time series.
- Obtain the periodogram of the time series so as to identify relevant frequencies.
- Compute the auto correlation (Eq. (9)) and the partial autocorrelation of $\tilde{C}_t$.
- If **relevant** autocorrelations can be observed, the blog is amenable for forecasting of fluctuations.
- Fit the *ARCH* or *GARCH* model to the time series.
- Once the values of β_p and α_q are known, we can obtain the conditional volatility, i.e. the values of σ_t given past values of q and p .

In order to fit the parameters of the *ARCH* and *GARCH* processes, many commercial and free software packages⁶ are

⁶An example open source python package for fitting ARCH models can be found under <https://pypi.python.org/pypi/arch>.

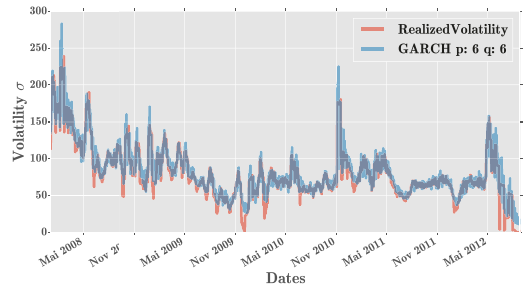


Fig. 5: Realized volatility vs Conditional Volatility as obtained from the best GARCH model for the <http://marisacat.wordpress.com/> website.

Blogs	Best Model	MSE	MSE p=1 q=1	# Comments
le-grove.co.uk/	(7, 7)	2.00	1.87	1625633
politicalticker.blogs.cnn.com/	(3, 2)	4.48	4.16	1461266
order-order.com/	(5, 5)	4.87	3.80	1289484
shoone.wordpress.com/	(8, 7)	1.39	1.16	93493
technologizer.com/	(9, 1)	594.34	649.79	165856
snsdorean.wordpress.com/	(2, 5)	0.40	0.28	83081
collegecandy.com/	(9, 1)	0.71	0.61	51495
seokyualways.wordpress.com/	(1, 6)	3.76	2.41	130460
religion.blogs.cnn.com/	(9, 9)	21.66	18.74	2426437
kickdefella.wordpress.com/	(6, 8)	0.22	0.17	39312

TABLE III: Fitting results for the top 10 newsblogs.

available [22]. We performed a statistical test of heteroscedasticity for all of the blogs in our data set [16] and selected the top 100 web sites. We present the results for the top 10 blogs in Table III. We fitted *GARCH* models with a grid search on the parameters $p \in \{1, \dots, 10\}$ and $q \in \{1, \dots, 10\}$ and selected the best model according to Akaike Information Criterion (AIC) [23]. For comparison, we obtain the realized volatility [22] i.e. the standard deviation for the past 14 days window (σ for $\{\tilde{C}_{i-14}, \dots, \tilde{C}_{i-1}, \tilde{C}_i\}$) and compared to the conditional volatility σ_t as predicted by *GARCH* given the past values of the models one day in advance (according to the selected q and p). Figure 5 shows an exemplary *GARCH* fit of the fluctuations of the commenting behavior of a newsblog. For the top 100 blogs, we obtained average MSE values of 0.739100 for the models with best AIC values and 0.663600 for the models with $p=1, q=1$.

VII. CONCLUSIONS AND FUTURE WORK

We were concerned with the problem of understanding the dynamics of the collective attention of populations of readers of newsblogs and presented a novel time series analysis approach. Our approach relies on two main concepts: seasonality and volatility. The seasonality is uncovered by the analysis of the periodogram function of the time series of comments. Volatility, on the other hand, is uncovered by analyzing the deviation of the variable fluctuations in time. We found universal seasonal behavior of newsblogs users in our dataset which hint at universal patterns of human behavior. Additionally, we showed how *GARCH* modeling of the time series presents a way to forecast possible future scenarios for

users of newsblogs as we can generate fluctuations of user attention based on the previous history. Such patterns can be exploited to perform informed decision making, to develop new time based marketing and advertisement strategies, and to present content to new as well as loyal users.

In future work, we will focus on developing better attention and credibility metrics [24] since cyclic behavior might generate fictitious attention shifts. It is important to note that the periodogram uncovers the importance and period of the cycles present, but in order to fully recover the information at hand, a detailed analysis of the population is yet to be performed. There is also a need to uncover the meaning of the stochastic process parameters from first principles, i.e. from the individual users' behavior. As opposed to finance, where only time series data is available, news websites offer the possibility of tracking individual user behavior over time. Hence, since we might have more information about the individuals, a more detailed study of volatility could provide even better attention models.

REFERENCES

- [1] C. Bauckhage, "Insights into internet memes," in *Proc. of ICWSM*, 2011.
- [2] A. Vespignani, "Modelling dynamical processes in complex socio-technical systems," *Nature Physics*, vol. 8, no. 1, 2012.
- [3] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proc. of ACM SIGKDD*, 2009.
- [4] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proc. of ICDM*, 2010.
- [5] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, "Rise and fall patterns of information diffusion: model and implications," in *Proc. of ACM SIGKDD*, 2012.
- [6] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, May 2005.
- [7] R. Sifa, C. Bauckhage, and A. Drachen, "The Playtime Principle: Large-scale Cross-games Interest Modeling," in *Proc. of IEEE CIG*, 2014, pp. 365–373.
- [8] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, 2007.
- [9] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," *Scientific reports*, vol. 2, 2012.
- [10] M. Goetz, J. Leskovec, M. McGlohon, and C. Faloutsos, "Modeling blog dynamics," in *Proc. of ICWSM*, 2009.
- [11] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance, "Finding patterns in blog shapes and blog evolution," in *Proc. of ICWSM*, 2007.
- [12] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, 1979.
- [13] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica: Journal of the Econometric Society*, 1982.
- [14] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*. Artech House Norwood, 2005, vol. 46.
- [15] C. T. Brownlee, R. F. Engle, and B. T. Kelly, "A practical guide to volatility forecasting through calm and storm," *Available at SSRN 1502915*, 2011.
- [16] A. J. Patton and K. Sheppard, "Evaluating volatility and correlation forecasts," in *Handbook of financial time series*. Springer, 2009, pp. 801–838.
- [17] E. S. Herman and N. Chomsky, *Manufacturing consent: The political economy of the mass media*. Random House, 2008.
- [18] M. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary Physics*, vol. 46, no. 5, 2005.
- [19] G. Szabo and B. Huberman, "Predicting the Popularity of Online Content," *Comm. of the ACM*, vol. 53, no. 8, 2010.
- [20] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. Amaral, "A poissonian explanation for heavy tails in e-mail communication," 2008.
- [21] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Characterizing and modeling the dynamics of online popularity," *Phys. Rev. Lett.*, vol. 105, Oct 2010.
- [22] J. D. Hamilton, *Time series analysis*. Princeton university press Princeton, 1994, vol. 2.
- [23] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [24] B. Ulicny and K. Baclawski, "New metrics for newsblog credibility," in *Proc. of ICWSM*, 2007.
- [25] R. G. King and S. T. Rebelo, "Low frequency filtering and real business cycles," *Journal of Economic dynamics and Control*, vol. 17, no. 1, 1993.

VIII. APPENDIX

A. Distributions

1) *Pareto Distribution*:: Given the shape parameter α the probability density function (PDF) of the pareto distribution is assigned non-zero values starting from the scale parameter x_{min} . The PDF of pareto distribution is :

$$f_X(x; \alpha, x_{min}) = (\alpha - 1)x_{min}^{\alpha-1}x^{-\alpha}. \quad (6)$$

2) *Log-normal distribution*:: A random variable whose logarithm is normally distributed follows lognormal distribution. Given respectively scale and location parameters σ and μ the PDF of the lognormal distribution is defined as:

$$f_X(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0. \quad (7)$$

B. Filters

The original role of the Baxter King filter [25] is to isolate the business cycle over a particular econometric time series. It is accomplished via a transformation of the stochastic process as $B(l)C_t = \hat{C}_t$. The desired effect is accomplished by looking at how the transformation is carried in the spectral density. In the frequency space the filter is expressed via: $B^*(e^{i\omega}) = \sum_{-\infty}^{\infty} a_j e^{ij\omega}$. This is impossible to perform since we only have a finite time series. We approximate via:

$$a_j = \begin{cases} a_j * +\theta & \text{if } |j| < J \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

We impose $B(1) = 0$ eliminating zero frequencies and in turn rendering the process stationary.

1) *Autocorrelations*: The autocorrelation function measures the correlation of the values of the stochastic process at two different points in time. How similar two values are. In the case of signal S_t of period τ for example, the correlation will be maximum at the period. It is a function of the time lag between two points in the time series. It is defined as:

$$R(s, t) = \frac{E[(X_t - \mu)(X_s - \mu)]}{\sigma_t \sigma_s} \quad (9)$$

For a stochastic process x_t we define the partial correlation (the conditional correlation) is defined as

$$P(k) = \frac{Cov(x_t, x_{t-k})}{\sqrt{V_t V_k}} \quad (10)$$

where V_t and V_k are defined as

$$V_t = Var(x_t | x_{t-1}, \dots, x_{t-k}) \quad (11)$$

$$V_k = Var(x_{t-k} | x_{t-1}, \dots, x_{t-k-1}). \quad (12)$$

Data Science – Analytics and Applications
Proceedings of the 1st International Data Science
Conference – iDSC2017

Haber, P.; Lampoltshammer, Th.J.; Mayr, M. (Hrsg.)
2017, XXIII, 105 S. 65 Abb. Book + eBook., Hardcover
ISBN: 978-3-658-19286-0