Twitter Critical Phases Identification Based on Time Series of Microposts Analysis

Andrey Dmitriev[a], Victor Dmitriev[a] and Stepan Balybin[b]

[a] National Research University Higher School of Economics, Moscow, Russia
[b] Department of Physics, Lomonosov Moscow State University, Moscow, Russia

Abstract

Based on the basic principles of the self-organized criticality theory, we proposed an identifiers of network criticality. The identifiers allow you to determine the subcritical and supercritical phases of Twitter, using only the results of the analysis of the time series of microposts. The most significant result is the existence of two classes of time series of microposts and tweet Ids corresponding to them. The first class of the time series corresponds to the subcritical phase of the network. On the contrary, the second class corresponds to the supercritical phase.

Keywords 1

Self-Organized Criticality, Subcritical Phase, Supercritical Phase, Twitter Time Series, Detrended Fluctuation Analysis

1. Introduction

Some of the objects and phenomena studied by sociophysics are social networks and critical phenomena, such as phase transitions, observed in them (e.g., see the reviews [1,2] and references therein). In the thermodynamics theory of irreversible processes, it is stated that significant structure reconstructions occur when the external parameter reaches a certain critical value and has the character of a kinetic phase transition [3]. The critical point is reached as a result of finetuning of the system external parameters. In a certain sense, such critical phenomena are not robust.

At the end of the 1980s, Bak et al. [4,5] found that there are complex systems with a large number of degrees of freedom that go into a critical mode as a result of the internal evolutionary trends of these systems. A critical state of such systems does not require fine-tuning of external control parameters and may occur spontaneously.

The motivation of our investigation is the following. There is a number of studies (e.g., see the works [6–15]), in which it is established that the observed flows of microposts generated by microblogging social networks (e.g., Twitter) are characterized by avalanche-like behavior. Time series of microposts depicting such streams are the time series with a power-law distribution of probabilities, with $1/f$ noise and long memory. Despite this, there are no studies on the critical phases identification on Twitter based on the time series microposts analysis. The critical phases identification is the purpose of our research.

2. Identifiers of the critical phases on Twitter

To determine the network phases, it is necessary to determine the size of avalanche microposts, which will allow the social network to be assigned to one of the critical phases.

The key features of the complexity of the social networks at the level of the time series generated by them are the power law for the probability distribution function (power-law PDF) of the time series of microposts, the power spectral density (PSD) of the time series, which is characterized by $1/f$ noise,

Modeling and Analysis of Complex Systems and Processes - MACSPro'2020, October 22–24, 2020, Venice, Italy & Moscow, Russia
EMAIL: a.dmitriev@hse.ru (Andrey Dmitriev); vdmitriev@hse.ru (Victor Dmitriev); sn.balybin@physics.msu.ru (Stepan Balybin)
ORCID: 0000-0002-7092-9101 (Andrey Dmitriev); 0000-0003-1544-6462 (Victor Dmitriev); 0000-0002-9845-6684 (Stepan Balybin)

and the power law for the autocorrelation function (power-law ACF), which is characterized by the presence of the long memory in the time series [16–18].

2.1. Power-law distribution

In the general case, the power-law PDFs can be considered as a statistical value of the scale invariance of the time series of microposts:

$$p(\eta) \propto \eta^{-\alpha}, \quad (1)$$

where $\alpha \in (1,3)$, η is number of the microposts. It should be noted that usually power-law PDFs are characterized by $\alpha \in (2,3)$ [19]. We considered the most common case belonging to power-law PDF.

Power-law PDF (1) refers to distributions with heavy tails, for which, unlike compact distributions, the well-known 3σ rule (the possibility of neglecting the values of the number of microposts exceeding 3σ) is not satisfied. If the distribution (1) is fulfilled, then rare large events do not occur infrequently enough for their probability to be neglected. The possibility of gigantic, extraordinary events appearing on Twitter indicates the network's tendency for disasters.

2.2. Flicker noise

Another characteristic of the scale-invariant properties of the time series is $1/f$ noise, which is observed in the power form of PSD at low frequencies f:

$$S(f) \propto f^{-\beta}. \quad (2)$$

The β value in PSD (2) determines the color of the noise. For $1/f$ noise, $\beta \in (0.5,1.5)$. The case of $\beta = 1$ is usually referred to as pink noise. $1/f$ noise is characteristic of all complex systems, regardless of their nature. If in the time series η_t there is $1/f$ noise, then for the social network there are no periodically repeated values of the number of microposts. This is due to the fact that, in the time series of microposts, it is impossible to distinguish one characteristic scale responsible for the appearance of large values of the number of microposts. (e scale-invariant type of PSD demonstrates a strong nonlinearity of social network signals when it is impossible to isolate individual components in the spectrum and offer its physical interpretation. (us, the dynamics of Twitter microposts, in which $1/f$ noise is observed, cannot be decomposed into separate components. Twitter, operating in a self-organized state, generates oscillations of microposts with PSD of the form (2).

2.3. Long memory

The third universal feature of complexity associated with power laws (1) and (2) is the existence of the long memory in the time series of microposts. In simple systems, the time correlation function (for example, the autocorrelation function), which shows the extent of which the time series "remembers" its history, has the following form:

$$\rho(\tau) \propto \exp(-\tau/\tau_s). \quad (3)$$

Complex systems are characterized by a power-law decrease in ACF as the time lag τ increases:

$$\rho(\tau) \propto \tau^{-\gamma}, \quad (4)$$

where $\gamma \in (0,1)$.

The existence of power-law ACF for the time series of microposts means that the current number of microposts largely depends on the past number of microposts generated by Twitter, as well as the absence of characteristic times at which information about the previous appearance of microposts would be lost.

It is fundamentally important that the existence of long temporal correlations states the fact of the emergence of Twitter. This is fact determines the possibility of the emergence of the avalanche of microposts (extremal events).

2.4. Spectrum of criticality exponents

If for the time series of microposts relevant to a certain topic, power laws (1), (2), and (4) are fulfilled, then the following important consequences are possible.

Firstly, the relevant Twitter segment distributing microposts relevant to a particular topic, is in the SOC state. Secondly, power laws describe large-scale invariance in the structure of time series of microposts generated by the self-organized critical social network.

PDF, PSD, and ACF in the form of power laws make it possible to use the range of interval indicators α, β, γ as the indicator of the self-organized criticality of the Twitter. If the social network is in the SOC state or the supercritical (SupC) phases, then for such states the indicators of power laws (spectrum of the criticality exponents, $\{\alpha, \beta, \gamma\}$) take the values from the intervals $(1,3)$, $(0.5,1.5)$, $(0,1)$. Otherwise, Twitter is in the subcritical (SubC) phase.

3. Identification of critical phases

3.1. Mining Twitter time series data and methods their analysis

The most suitable data source for mining of Twitter time series data that contain tweet ids (unique identifiers of tweets) regarding different events, such as political elections and natural disasters, is Harvard Dataverse. It contains the datasets of tweets ids on 12 different topics, and each dataset consists of more than 2 million unique tweet ids in the form of the 18-digit numbers (for example, 1128408193699340294) combined into one text file (.txt). Harvard Dataverse collected data using Social Feed Manager, which is the open source software that harvests social media data and web resources from Twitter. (e reason why it is necessary to start to work with tweets ids, rather than tweets itself, is the fact that per Twitter's Developer Policy, tweet ids may be publicly shared for academic purposes, but tweets may not.

Nevertheless, in order to get Twitter time series, it is necessary to hydrate the obtained datasets of tweet ids. Hydrating is the process of loading JSON objects from tweets based on available tweet ids. It can be done using the API-interface of Twitter, as well as using third-party applications. We did it with a Hydrator version 0.0.3 software. According to the obtained data, it is possible to build the interaction structure of users and time series of tweets (including retweets and other mentions).

As a result, we got twelve equidistant (a step is 1 second) time series of microposts $\{\eta_i\}, i = 1,2,\dots,N$ of different lengths N, each of which is relevant to some topic (tweet Ids).

The traditional approach to the time series analysis relies on the measurement of PSD and ACF. However, only the implementation of Gaussian processes is exhaustively described by their second moments. Outside of such implementations, a complete statistical description requires an estimate of higher order moments. In addition, higher order moments do not always have such a clear physical meaning as ACF and PSD. Therefore, evaluations of a small number of values that can be given a certain meaning become important. (ese values include the fractal dimensions of the time series.

The fractal dimension is closely related to the scaling index s, which can be the Hurst exponent, estimated by the method of normalized range or fluctuation analysis (FA) [20], or the generalized Hurst exponent, estimated by the method of detrended fluctuation analysis (DFA) [21].

The DFA method is an efficient method for analysis of the time series characterized by the presence of the long memory or $1/f$ noise. The DFA method is a generalization of the FA method for analysis of the scale invariance of nonstationary time series. The DFA method allows both to estimate the scaling indicator of the time series s and to obtain indirect estimates of β_s and γ_s indicators, calculated from the generalized scaling indicator s of the time series.

3.2. Data analysis results and their discussion

Table 1 presents the ordinary least squares estimates of the spectrum criticality and DFA estimates of scaling indicators s, β_s, and γ_s. The corresponding p values are shown in brackets.

The symbol "–" denotes the absence of statistically significant DFA estimates for β_s and γ_s indicators. Statistically significant values of the exponents are denoted in bold.

Table 1

The spectrum criticality and DFA estimates of scaling indicators

Tweet Ids	α	β	γ	β_s	γ_s	s
2016 United States Presidential Election	**1.23**(0.0121)	**1.29**(0.0182)	**0.12**(0.0201)	**0.92**(0.0036)	**0.08**(0.0036)	**1.04**(0.0036)
Women's March	**2.11**(0.0234)	**1.23**(0.0198)	**0.42**(0.0211)	**0.90**(0.0101)	**0.10**(0.0101)	**1.05**(0.0101)
End of Term 2016 US government	3.24(0.6743)	0.24(0.7235)	5.24(0.6990)	–	–	0.45(0.7699)
Hurricanes Harvey	**2.12**(0.0312)	**1.13**(0.0289)	**0.34**(0.0320)	**0.89**(0.0015)	**0.11**(0.0015)	**1.06**(0.0015)
Hurricanes Irma	**2.23**(0.0234)	**0.98**(0.0194)	**0.18**(0.0209)	**0.96**(0.0098)	**0.04**(0.0098)	**1.02**(0.0098)
Immigration and Travel Ban	**2.18**(0.0401)	**1.09**(0.0320)	**0.21**(0.0128)	**0.97**(0.0094)	**0.03**(0.0094)	**1.02**(0.0094)
Charlottesville	**2.18**(0.0313)	**1.21**(0.0287)	**0.43**(0.0121)	**0.90**(0.0101)	**0.10**(0.0101)	**1.05**(0.0101)
Winter Olympics 2018	3.59(0.7239)	0.22(0.6348)	5.64(0.5341)	–	–	0.52(0.8172)
US Government	3.28(0.6361)	0.19(0.7298)	6.01(0.6399)	–	–	0.48(0.7456)
News Outlet	3.36(0.4275)	0.23(0.3895)	5.50(0.4458)	–	–	0.54(0.6451)
2018 US Congressional Election	**1.47**(0.0281)	**1.05**(0.0398)	**0.22**(0.0435)	**0.95**(0.0099)	**0.05**(0.0099)	**1.03**(0.0099)
115th US Congress	3.99(0.3189)	0.26(0.4197)	5.24(0.5618)	–	–	0.46(0.9999)
Ireland 8th	**2.18**(0.0311)	**1.18**(0.0270)	**0.35**(0.0311)	**0.97**(0.0129)	**0.03**(0.0129)	**1.02**(0.0129)

The most significant result in the context of our study is the existence of two classes of time series of microposts and tweet Ids corresponding to them.

The first class consists of time series for which $\alpha \in [1.23,2.23]$, $\beta \in [1.05,1.29]$, and $\gamma \in [0.12,0.43]$. Indicators of the power laws of such time series belong to the spectrum of indicators of criticality $(1,3)$, $(0.5,1.5)$, $(0,1)$ and, consequently, Twitter, which generates such time series of microposts, is in the SOC state or the SupC phase. The social network is capable of generating extreme events, which are avalanches of microposts of all sizes corresponding to the following tweet ids: "2016 United States Presidential Election," "Women's March," "Hurricanes Harvey," "Hurricanes Irma," "Immigration and Travel Ban," "Charlottesville," "2018 US Congressional Election," and "Ireland 8th." In addition, the current number of microposts largely depends on the past number of microposts generated by Twitter. Indeed, for all the time series of this class indicator ACF, $\gamma \in (0,1)$. It is noteworthy that all tweet ids relate either to protest movements or to political elections or to the population activities during natural disasters. PDF of such time series have infinite η^2 and infinite η for events related to political elections and finite η in all other cases. DFA estimates of βs and cs give close values to the corresponding indicators β and γ, and the presence of statistically significant values of the scaling exponent s determines the scale invariance of time series, which is one of the key features of the selforganized criticality of the social network. In addition, for all time series of the first class $s \cong 1$ and $\beta_s \cong 1$, which corresponds to the presence of pink noise and, accordingly, being Twitter in the SOC state or the SupC phase. The existence of a dependency (4) for the time series of microposts means that the current numbers of microposts largely depend on the past number of microposts generated by Twitter, as well as the absence of characteristic times at which information about previous occurrences of microposts would be lost.

The second class consists of time series for which $\alpha \in [3.24,3.9]$, $\beta \in [0.19,0.26]$, and $\gamma \in [5.24,6.01]$; moreover, estimates of all indicators are not statistically significant: statistical hypothesis is accepted with previously considered p values shown in Table 1. Consequently, for these time series of microposts, at least the power laws (1) and (4) are not satisfied. This result is consistent with the results of the detrended fluctuation analysis, according to which there is no statistically significant estimate of the scaling exponent s; therefore, these time series of microposts are not scale-invariant. Thus, Twitter, which generates these time series, is neither in the SOC state or in the SupC phase.

Twitter users, that is, in such a state, are not coordinated. This leads to the generation of the time series, for which the spectrum is not performed. It may be the SubC phase, but such a conclusion requires the determination of the explicit form of PDF and ACF dependencies, which is beyond the scope of our study. The only argument in favor of the assumption of Twitter being in the SubC phase is the fact that indicator $\beta \in [0.19,0.26]$ is close to the value corresponding to white noise ($\beta \cong 0$).

4. Conclusion

The presence of a spectrum of criticality indicators $\{(1,3); (0.5,1.5); (0,1)\}$ for the observed time series of microposts is a sufficient feature that Twitter is in a supercritical phase. It is important that the identification of a self-organized critical state of the network does not require a detailed analysis of the interactions between its users at the micro level. Only an analysis of the time series of microposts for being in the spectrum is sufficient, which does not require significant resource costs.

An approach to monitor the social network state based on the spectrum analysis, for example, can be effective for identifying the origin of protest movements for which Twitter is one of the tools. In addition, the approach can be used to study the activity of users on the network related to political elections. For example, if the social network is in the SubC phase and for the corresponding time series of microposts it is possible to find the interval $\Delta t \in \Delta t_{\mathrm{SubC}}$, in which a slow increase in the size of avalanches is observed, then the existence of such Δt is a possible precursor of the appearance of the SOC state and further transition to the SupC phase. Another situation is possible. If it is possible to find a relatively small interval $(t_c, t) \in \Delta t_{\mathrm{SubC}}$, then the existence of such an interval is a possible precursor of the unpredictability of the behavior on the social network. Starting from time t, avalanches of microposts of all sizes will appear.

Note that all this is nothing more than a discussion of possible applications. To conduct such studies, it is necessary to develop and test algorithms for detecting such integrals, but this is beyond the scope of our study.

5. Acknowledgments

This work was partially funded by the Russian Foundation for Basic Research (grant 16-07-01027).

6. References

[1] Dorogovtsev, S., Goltsev, A., Mendes, J.: Reviews of Modern Physics 80(4), 1275–1335 (2008).
[2] Fronczak, P., Fronczak, A., Hołyst, J.: European Physical Journal B 59(1), 133–139, (2007).
[3] Eu, B.: Development and Application 93, 12–54 (1998).
[4] Bak, P., Tang, C., Wiesenfeld, K.: Physical Review Letters 59(4), 381–384 (1987).
[5] Bak, P., Tang, C., Wiesenfeld, K.: Physical Review A 38(1), 364–374 (1988).
[6] Meng, Q.: Chinese Physics Letters (28), Article ID 118901, (2011).
[7] Noel, P., Brummitt, C., D'Souza, R.: Physical Review E 89, Article ID 012807, (2014).
[8] Mollgaard, A., Mathiesen, J.: PLoS One 10, Article ID e0123876, (2015).
[9] Aguilera, M., Morer, I., Barandiaran, X.: Quantifying political self-organization in social media. Fractal patterns in the Spanish 15M movement on twitter. In Proceedings of the 12th European Conference on Artificial Life, 395–402 (2013).
[10] Kirichenko, L., Bulakh, V., Radivilova, T.: Fractal time series analysis of social network activities. In Proceedings of the IEEE 4th International Scientific-Practical Conference Problems of Infocommunications, 456–459, (2017).
[11] De Bie, T., Lijffijt, J., Mesnage, C.: Detecting trends in twitter time series. In Proceedings of the IEEE 26th International Workshop on Machine Learning for Signal Processing, 1–6, (2016).
[12] Bild, D., Liu, Y., Dick, R.: ACM Transactions on Internet Technology 15 (4), (2015).
[13] Remy, C., Pervin, N., Toriumi, F.: Information diffusion on twitter: everyone has its chance, but all chances are not equal. In Proceedings of the IEEE International Conference on Signal-Image Technology and Internet-Based Systems, (2013).

[14] Gleeson, J., Durrett, R.: Nature Communications 8, 1227 (2017).
[15] Liu, C., Zhan, X., Zhang, Z., Sun, G., Hui, P.: New Journal of Physics 17(11), Article ID 113045, (2015).
[16] Kantelhardt, J.: Fractal and multifractal time series. In Mathematics of Complexity and Dynamical Systems, (2012).
[17] Dmitriev, A., Kornilov, V., Maltseva, S.: Complexity 2018, Article ID 4732491, (2018).
[18] Dmitriev, A., Dmitriev, V., Balybin, S.: Complexity 2019, Article ID 8750643, (2019).
[19] Clauset, A., Shalizi, C., Newman, M.: SIAM Review 51(4), 661–703 (2009).
[20] Hurst, H.: Transactions of the American Society of Civil Engineers 116, 770–799 (1951).
[21] Peng, C., Buldyrev, S., Havlin, S., Simons, M., Stanley, H., Goldberger, A.: Physical Review E 49(2), 1685–1689 (1994).