Fluctuation and Burst Response in Social Media

Mizuki Oka
University of Tsukuba
Tennodai 1-1-1 Tsukuba
Ibaraki, 305-8577, Japan
mizuki@cs.tsukuba.ac.jp

Yasuhiro Hashimoto
The University of Tokyo
Kashiwanoha 5-1-5, Kashiwa
Chiba, 277-8561, Japan
hashimoto@vis.k.u-tokyo.ac.jp

Takashi Ikegami
The University of Tokyo
Komaba 3-8-1, Meguro-ku
Tokyo, 153-8902, Japan
ikeg@sacral.c.u-tokyo.ac.jp

ABSTRACT

A salient dynamic property of social media is bursting behavior. In this paper, we study bursting behavior in relation to the structure of fluctuation, known as *fluctuation-response relation*, to reveal the origin of bursts. More specifically, we study the temporal relation between a preceding baseline fluctuation and the successive burst response using a frequency time series of 3,000 keywords on Twitter. We find three types of keyword time series in terms of the fluctuation-response relation. For the first type of keyword, the baseline fluctuation has a positive correlation with the burst size; as the preceding fluctuation increases, the burst size increases. These bursts are caused endogenously as a result of word-of-mouth interactions in a social network; the keyword is sensitive only to the internal context of the system. For the second type, there is a critical threshold in the fluctuation value up to which a positive correlation is observed. Beyond this value, the size of the bursts becomes independent from the fluctuation size. Our analysis shows that this critical threshold emerges because the bursts in the time series are endogenous and exogenous. This type of keyword is sensitive to internal and external stimuli. The third type is mainly bursts caused by exogenous bursts. This type of keyword is mostly sensitive only to external stimuli. These results are useful for characterizing how *excitable* a keyword is on Twitter and could be used, for example, for marketing purposes.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Servies - Web-based services; H.1.2 [**Models and Principles**]: User/Machine Systems; J.4 [**Computer Appication**]: Social and Hehavioral Sciences - Sociology

1. INTRODUCTION

Social media such as Facebook, Twitter, and Google Plus have established their role as information-sharing tools, both personally and commercially [7]. With the introduction of these new forms of social media, one can observe how people respond to specific information on the web. When information receives collective attention, the information appears as a *burst*, an increase in the number of appearances about the information for a certain period of time on social media. For example, if we take the number of tweets that contain the keyword *earthquake* as depicted in Figure 1, the bursts in the keyword time series show a strong correlation with the occurrences of earthquakes. This is because when there is an earthquake, people tend to tweet about it using the keyword *earthquake*. These bursts occur aperiodically in accordance with the timing of the earthquakes.

Another example of bursts is in the keyword time series such as *school* as depicted in Figure 1. We observe daily periodic bursts since people attend school every day on weekdays and people tweet about it. As these examples show, by aggregating the time series of keywords on social media, such as Twitter, we can extract patterns that exhibit underlying natural phenomena to human behaviors. There is also a keyword such as *joy*, which does not show obvious bursts but continuous fluctuations in the number of tweets as in Figure 1.

A *burst* is one of the most salient temporal features on Twitter. Several studies have investigated the properties of these bursts to reveal insight into people's collective behavior [2, 8]. Crane and Sornette analyzed a property of a burst in terms of endogenous and exogenous bursts [2]. Exogenous bursts are caused by external influences such as earthquakes or appearances in the mass media. Endogenous bursts are caused as a result of word-of-mouth interactions in a social network. Crane and Sornette found that whether a burst is exo- or endo- can be found by looking at the peak ratio of the burst; when the peak ratio is *small*, then the burst is endogenous, otherwise exogenous. Lehmann et al. applied their findings to a large-scale record of tweets, specifically hash-tagged tweets, and using endogenous and exogenous bursts, demonstrated that tweets can be clustered into four classes [8].

In these studies, Twitter is a system that directly reflects people's responses. However, a system, even an artificial one, should exhibit an autonomous internal structure [9, 11, 1]. This autonomous internal structure can be found, for example, in a chemically made self-running *oil-droplet* [5, 4]. An oil droplet is a very simple artificial system made of olive oil and alkaline water. Although the system is simple, through interactions with the environment, the droplet starts to move around autonomously as a result of the emergence of convection flow within the droplet. Another au-

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579208.

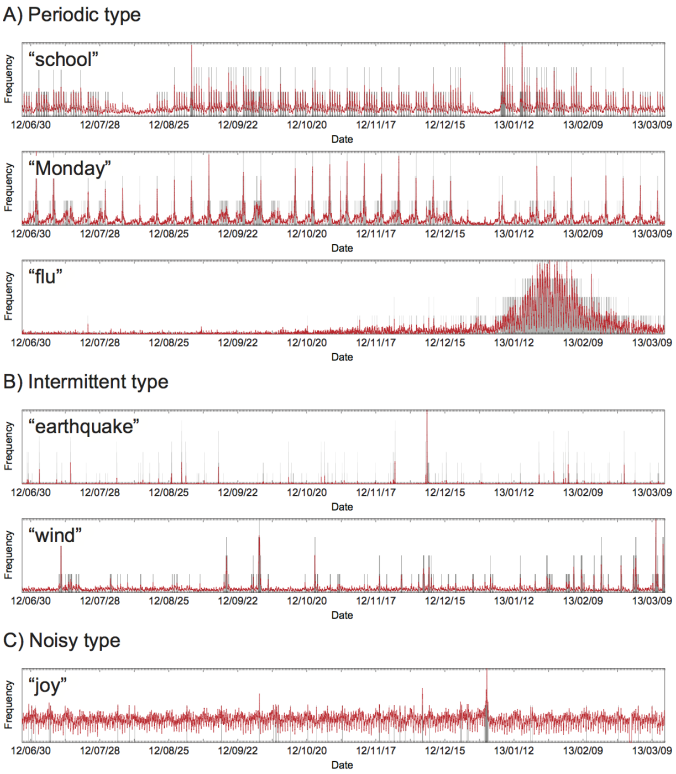

Figure 1: Examples of time series (red lines) and detected bursts (gray bars; the height indicates the burst level). A) The periodic type with a small time scale (`school`, `Monday`), the periodic type with a large time scale (`flu`), B) the intermittent type (`earthquake`, `wind`), and C) the noisy type (`joy`).

tonomous system, among many, is a *neuron* in the brain. The brain shows active patterns consuming more metabolic energy when the brain is resting; no input is fed into the brain. This mode in the brain, known as the default mode, has been extensively studied and revealed insights into several functionalities of the brain [12, 13].

Our interest here is Twitter, which exhibits bursting behaviors similar to the firing of a neuron in the brain and has an internal structure that defines the response size, i.e., burst size, within the system. To investigate this notion quantitatively, we use a theory called *fluctuation-response relation* in statistical physics [14]. This theory says that if a system exhibits fluctuation, the response size to an external stimulus has a linear relation with the size of the fluctuation. That is, the larger the fluctuation, the larger the response size. For example, imagine a ball in a plate, and the ball is fluctuating within the plate. If the fluctuation is small, the ball would not get out of the plate when there is a stimulus (e.g., tilt the plate with a human finger). However, if the fluctuation of the ball is large, the ball would get out of the plate with a stimulus from outside. In general, the fluctuation-response relation holds in the thermally equilibrium system but is phenomenologically extended to many non-equilibrium open systems from physics [14] to biology [10, 16] and economics [15].

Table 1: General statistics for the dataset

total number of tweets	297,792,366
total number of users	12,677,098
total number of keywords	1,550,770

We are interested in whether this fluctuation and response relationship also holds in the Twitter keyword time series as an autonomous property of Twitter. More specifically, we regard the size of the burst as the strength of the response on Twitter and the standard deviation in the number of occurrences of keywords as the fluctuation and studied the temporal relationships. We found that the fluctuation-response relation holds on Twitter as well. That is, when the fluctuation is small, the response to external stimuli is smaller, and when the fluctuation is larger, the response increases.

The rest of the paper is organized as follows. In section 2, we describe the data set used for the study as well as the method for detecting bursts that divide each keyword time series into the fluctuation periods (non-bursting periods) and bursting periods. In section 3, we show the detailed analysis of fluctuation and response relation and report the results. We also further classify these bursts as endogenous or exogenous by looking at the peak ratio of each burst and discuss the ratio in relation to the fluctuation-response relation. In section 4, we conclude and discuss future work.

2. MATERIALS AND METHODS

We describe the data used for this study as well as the methods used for detecting bursts in the time series.

2.1 Data

We collected tweets (in Japanese) over a two-year period beginning in July 2011, using Streaming API with the sampling method available at the Twitter developers site[1]. Then, we applied morphological analysis using MeCab software, state-of-the-art software for Japanese morphological analysis[2]. We then extracted the 3,000 most frequently used Japanese nouns as keywords in the tweets.

In the collected data were many automated tweets posted by programs called bots. Some of the data had peculiar statistics due to these bots. To mitigate the bot effect, we used the number of unique users to count the frequency of the keywords, rather than the number of tweets. The basic statistics of the data are shown in Table 1. We chose the 3,000 most popular keywords from 1,550,770 distinct keywords and created a time series for each keyword by counting the number of unique users in 10-minute time intervals. We then smoothed each time series using a Gaussian kernel with a standard deviation of 30 minutes.

2.2 Detection of bursts and fluctuations

Bursts and fluctuations form a continuous spectrum in a time series; thus, we may not have the necessary criterion for distinguishing between the two. However, we believe that they are intuitively and qualitatively different. One supporting argument is that the form of frequency distribution suggests that it consists of a log-normal distribution with a

[1]Streaming API collects at most 1% of all tweets produced on Twitter at a given time according to the documentation available at https://dev.twitter.com.
[2]Available at https://code.google.com/p/mecab/.

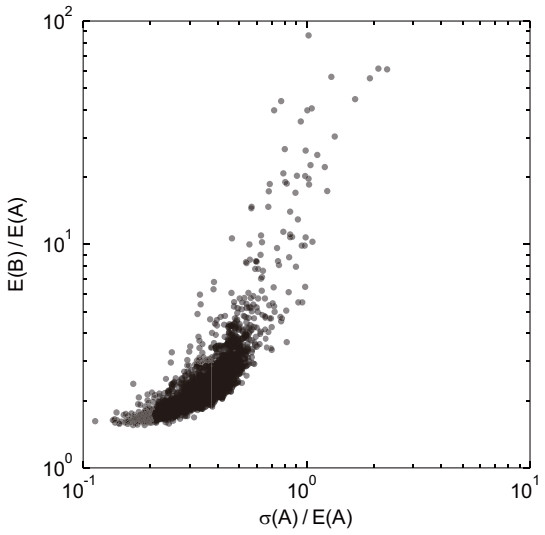

Figure 2: Overall fluctuation-response relation for the 3,000 keywords. $\sigma(A)$ and $E(A)$ are the standard deviation and the mean frequency of the baseline periods, respectively. $E(B)$ is the mean frequency of the burst periods in a time series.

long-tail part. The log-normal distribution corresponds to the fluctuation and the long-tail part to bursts. We thus label each period of the time series as either a burst or baseline fluctuation.

We symbolize the time series by baseline fluctuation (A_i) and burst (B_i) periods. That is, each time series is translated into a symbol sequence of $A_1, B_1, A_2, B_2, \ldots, A_n, B_n$. Each baseline period A_i is always followed by a burst B_i. As we will see in detail, a burst period, B_i, is roughly defined to exist when the frequency is two times the overall average according to the Kleinberg algorithm [6] and otherwise as a baseline period, A_i.

Kleinberg's algorithm [6] is used to distinguish burst activity (B_i) from baseline activity (A_i). The algorithm assumes the Poisson process for tweets; that is, successive tweets occur independently following Poisson distribution $f(x) = \lambda e^{-\lambda x}$, where λ is the mean frequency and x is the interval of the successive tweets. We define the burst level at each time t ($i(t)$) to distinguish burst activity from baseline activity. This level is not fixed over time but evolves as the mean frequency becomes s times larger than the overall mean frequency. We here set $s = 2$, so that the burst level increases by one when the frequency is twice as large as before. Formally, the mean frequency λ in the formulae is substituted with $\lambda_{i(t)}$ and is defined as $\lambda_{i(t)} = \bar{\lambda} s^{i(t)}$. The entire mean value is denoted by $\bar{\lambda}$.

Another parameter, γ, is also used to control the cost of changing the burst level between successive time points. The burst detection process identifies the time series of the burst level by minimizing the cost function defined by frequency matching and burst level stabilizing under constant parameters s and γ. We used $\gamma = 1$ and extracted the time periods in the time series that are $i(t) > 0$ as burst periods, and otherwise as baseline periods. During the burst period, we also spotted a time point when the frequency was the highest; we called this point the peak of the burst.

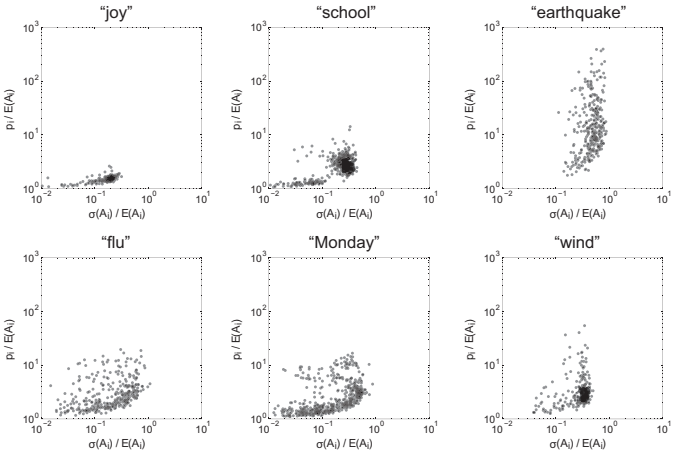

Figure 3: Three types of relationships identified between the baseline fluctuation, $\sigma(A_i)/E(A_i)$, and the peak burst size, $p_i/E(A_i)$. Type I: The response size has a positive correlation to the amplitude of its immediately prior baseline fluctuation (e.g., `joy` and `flu`). Type II: A positive correlation between the response size and the amplitude of the immediate prior baseline fluctuation to a certain threshold and has relatively large responses beyond the threshold (e.g., `school` and `Monday`). Type III: Abrupt responses ranging from small to large at a specific threshold; most importantly, all responses are concentrated around the threshold (e.g., `earthquake` and `wind`).

Using this setting, we labeled each period in a time series as either a baseline fluctuation period or a burst period for the 3,000 keyword time series. We then roughly classified them into three patterns, periodic, intermittent, and noisy, based on the temporal bursting patterns. The original time series are depicted with red lines, and the detected bursts are depicted with gray bars with the height indicating the burst level. The bursts detected in these three typical patterns are depicted in Figure 1. The three patterns are also reported as representative patterns of other online media such as blogs [3].

3. ANALYSIS

We study the temporal relation between the baseline fluctuations and the bursts of the 3,000 keyword time series.

3.1 Fluctuation-response relation

The fluctuation is represented by the standard deviation of the baseline frequency, denoted as $\sigma = \sqrt{\sum x_t^2 - (\sum x_t)^2}$. To study the overall relationship between a baseline fluctuation and a burst for all 3,000 keywords, the average fluctuations and the average burst sizes for each time series are plotted in Figure 2. In the figure, we observe a positive correlation between the fluctuation sizes and the burst sizes.

We then closely studied the temporal relation between a baseline fluctuation and a burst for each keyword by plotting each transition from A_i to B_i for n number of pairs. We found three typical classes of the relation exhibited. The three typical plots, type I to type III, are shown in Figure 3. The first type, type I, shows a relationship between

the baseline fluctuation and the burst such that the response size (i.e., the maximum size, or the peak p_i of the burst period B_i) is correlated with the amplitude of the immediately preceding baseline fluctuation σ (e.g., keywords such as `joy` and `flu` in Figure 3, type I).

The second type, type II, has a point up to which the fluctuation gradually amplifies where the fluctuation-burst relation changes qualitatively and causes large bursts (e.g., the keywords `school` and `Monday` in Figure 3, type II). We call this the critical threshold. Below this critical threshold, the burst response has a positive correlation with the preceding baseline fluctuation. Above the critical threshold, the size of the response becomes independent from the fluctuation size. Interestingly, these keywords have occasional bursts due to events that break periodicities. Taking the example of `school`, some periodicities originate in the circadian rhythm. Sometimes this periodicity breaks, and the fluctuation increases, causing the bursts that follow to also be larger (see Figure 4). These periodicity-broken phases correspond to major school breaks, such as spring, summer, and winter holidays. A disruption in repetitive everyday life triggers a large burst.

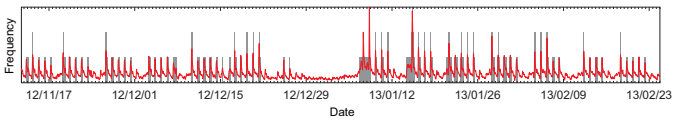

Figure 4: Breaks in the periodicities for the keyword school originating in the circadian rhythm causing the following bursts to increase.

The third type, type III, is found in keywords such as `earthquake` and `wind`, which are depicted in Figure 3 (type III). For this type, the fluctuation-independent bursts range from small to large merged at or above the critical threshold. The threshold value varies from one keyword to another. In the next section, we analyze the causes of the emergence of this critical threshold in terms of endogenous and exogenous bursts.

3.2 Endogenous and exogenous bursts

We identify each keyword's bursts as endogenous or exogenous by extending Crane and Sornette 's work in [2]. Namely, we consider not only the peak ratio but also compare it with the respective burst sizes to classify endogenous and exogenous bursts. More concretely, we measured each burst's peak-size ratio p_i/S_i against its scaled burst size S_i/E (A_i), where, p_i is the burst peak height, and S_i is the burst size. Exogenous bursts are caused by external influences, and the peak-size ratio becomes larger than a certain value; otherwise, the burst is defined as endogenous. Crane and Sornette analyzed a property of a single burst; however, we statistically analyzed a series of bursts at one time point and classified them as one of two distinct types of bursts.

Figure 5 shows the plots for types I, II, and III, represented by the corresponding keywords, respectively. Each point is colored using the fluctuation value depicted in Figure 3. The value is scaled from 0 to 1 for each keyword, with blue the smallest fluctuation value and red the largest fluctuation value. The peak-size ratio is either almost inversely proportional to the burst size on a logarithmic scale and the deviations from the proportional line. When the peak-size ratio is inversely proportional to the burst size, the size of the burst is bounded and rarely causes larger bursts. These bursts are endogenous. The colors in the figure suggest that these bursts have smaller fluctuation values. When the peak-size ratio deviates from the inversely proportional line and remains high, the average burst size is not bounded, resulting in larger bursts. These bursts are exogenous. The colors in the figure suggest that these bursts have larger fluctuation values.

Based on these criteria, we can see that type I bursts, represented by the keywords `joy` and `flu` in Figure 5, almost all the bursts, are endogenous. Type II bursts as represented by the keywords `school` and `Monday` are a mixture of endogenous and exogenous bursts. The bursts with smaller fluctuation values are endogenous, and the ones with larger fluctuation values are exogenous. This mixture of the two types of bursts is the critical threshold for the fluctuation values discussed in the previous section. For type III bursts, represented by the keywords `earthquake` and `wind`, most of the bursts are exogenous with a high peak ratio regardless of the burst sizes or the fluctuation values.

The relationship between fluctuation and response corresponds to endogenous bursts. In other words, when the burst is endogenous, the amplitude of the baseline fluctuation directly influences the burst size. However, the bursts on and above the critical threshold are exogenous, and the size of the burst becomes independent from the size of the fluctuation. By measuring whether bursts in a keyword are caused endogenously or exogenously, we can classify the keyword as one of the three fluctuation relations (type I, II, and III in Figure 3) and use it, for example, to characterize the *excitability* of the keyword to an external stimulus.

4. CONCLUSION AND DISCUSSION

In this paper, a temporal relationship between a baseline fluctuation and the subsequent response as a burst was studied. Twitter has a sensor that responds not only to external stimuli but also to internal dynamics. This is observed in the relationship between the baseline fluctuation and the burst sizes in the keyword time series. In some keywords, a response or a burst increases along with a baseline fluctuation. These bursts are caused endogenously. Some keywords have a threshold at the fluctuation value. Taking the threshold as a dynamic phase transition, we can interpret the fluctuation increase to the transition point as a critical fluctuation. At or above the threshold, the response becomes independent from the baseline fluctuation size, showing a wide range of burst sizes as a result of external influences. This critical threshold emerges as the result of different types of bursts, endogenous and exogenous bursts. The threshold has different values for different keywords and is self-organized due to an external disruption or coupled with different oscillatory behaviors. Other keywords show only exogenous bursts at a specific fluctuation value showing a wide range of burst sizes.

Based on these findings, we can identify, for example, how responsive a keyword is to internal or external stimuli and use it for marketing purposes.

5. ACKNOWLEDGMENTS

This work was supported by the Japan Society for the Promotion of Science Grant-in-Aid for Young Scientists (B)

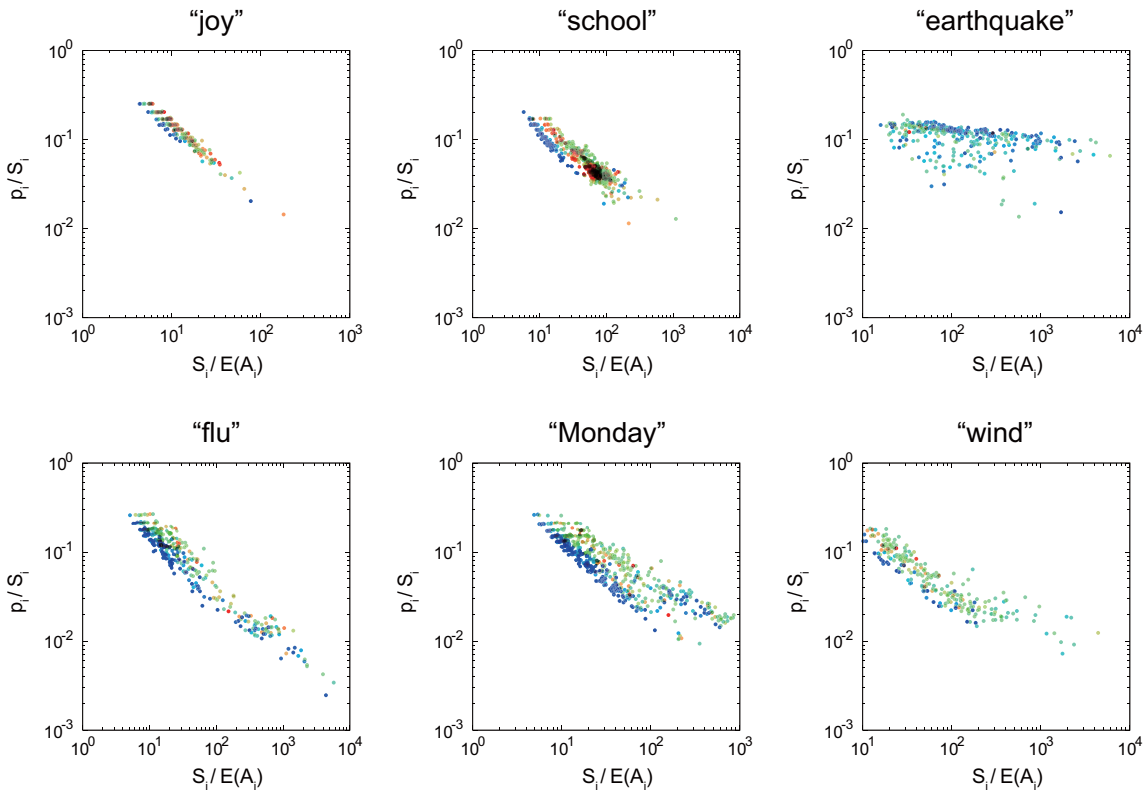

Figure 5: Burst size versus peak-size ratio. The peak-size ratio is either inversely proportional to the burst size (a linear curve with the exponent $= -1$ or deviates from it. p_i is the burst peak height, S_i is the burst size, and $E(A_i)$ is the mean frequency in the baseline period. The size of the plots is proportional to p_i. (Left-column) The burst size versus the peak-size ratio for type I time series (i.e., `joy` and `flu`) shows endogenous bursts, (Middle-column) type II time series (i.e., `school` and `Monday`) shows a mixture of endogenous and exogenous bursts, and (Right-column) type III time series (i.e., `earthquake` and `wind`) shows many exogenous bursts and a small number of endogenous bursts.

(#25730184 “Burst Analysis of Twitter time series based on RT diffusion and its application to Web services”) and partially by Grant-in-Aid for Scientific Research on Innovative Areas (#24120704 “The study on the neural dynamics for understanding communication in terms of complex hetero systems”). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

6. REFERENCES

[1] R. A. Brooks. Intelligence without representation. *Artificial Intelligence Journal*, 47:139–159, 1991.

[2] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc Natl Acad Sci USA*, 105(45):15649–15653, 2008.

[3] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501, 2004.

[4] M. M. Hanczyc and T. Ikegami. Chemical basis for minimal cognition. *Artificial Life*, 16(3):233–243, 2010.

[5] M. M. Hanczyc, T. Toyota, T. Ikegami, N. Packard, and T. Sugawara. Chemistry at the oil-water interface: Self-propelled oil droplets. *J. Am. Chem. Soc.*, 129(30):9386–9391, 2007.

[6] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, 2002.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of the 19th International World Wide Web*, pages 591–600, 2010.

[8] J. Lehmann, B. Gonçalves, and C. C. José J. Ramasco. Dynamical classes of collective attention in twitter. In *Proc. 21st Intl. Conf. on World Wide Web*, pages 251–260, 2012.

[9] S. Nolfi. Evolving non-trivial behaviors on real robots: A garbage collecting robot. *Robotics and Autonomous Systems*, 22:187–198, 1997.

[10] F. Oosawa. Effect of field fluctuation on a macromolecular system. *J. Theor. Biol.*, 52:175–186, 1975.

[11] R. Pfeifer, M. Lungarella, and F. Iida. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318:1088–1093, 2007.

[12] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. Inaugural article: A default mode of brain function. *PNAS*, 98:676–82, 2001.

[13] M. E. Raichle and A. Z. Snyder. A default mode of brain function: A brief history of an evolving idea. *NeuroImage*, 37(4):1083–1090, 2007.

[14] L. E. Reichl. *A Modern Course in Statistical Physics*. University of Texas, 1980.

[15] D. Ruelle. Converstations on nonequilibrium physics with an extraterrestrial. *Physics Today*, 4:48–53, 2004.

[16] K. Sato, Y. I. T. Yomo, and K. Kaneko. On the relation between fluctuation and response in biological systems. *Proc Natl Acad Sci USA*, 100(24):14086–14090, 2003.