Transition due to preferential cluster growth of collective emotions in online communities

Anna Chmiel and Janusz A. Hołyst
Faculty of Physics, Center of Excellence for Complex Systems Research, Warsaw University of Technology, Koszykowa 75, PL-00-662 Warsaw, Poland

(Received 31 July 2012; revised manuscript received 12 December 2012; published 15 February 2013)

We consider a preferential cluster growth in a stochastic model describing the dynamics of a binary Markov chain with an additional long-range memory. The model is driven by data describing emotional patterns observed in online community discussions with binary states corresponding to emotional valences. Numerical simulations and approximate analytical calculations show that the pattern of frequencies depends on a preference exponent related to the memory strength in our model. For low values of this exponent in the majority of simulated discussion threads both emotions are observed with similar frequencies. When the exponent increases an ordered phase emerges in the majority of threads, i.e., only one emotion is represented from a certain moment. Similar changes are observed with increase of a single-step Markov memory value. The transition becomes discontinuous in the thermodynamical limit when discussions are infinitely long and even an infinitely small preference exponent leads to ordered behavior in each discussion thread. Numerical simulations are in a good agreement with the approximated analytical formula. The model resembles a dynamical phase transition observed in other Markov models with a long memory where persistent dynamics follows from a transition to a superdiffusion phase. The ordered patterns predicted by our model have been found in the Blog06 dataset although their number is limited by fluctuations and sentiment classification errors.

DOI: 10.1103/PhysRevE.87.022808 PACS number(s): 89.65.−s, 89.20.Hh, 64.60.De

I. INTRODUCTION

It is well known (see, e.g., Ref. [1]) that a one-dimensional (1D) system with short-range forces cannot undergo a phase transition at a nonzero temperature. The situation changes when the interaction range increases, e.g., the Ising chain displays a second-order phase transition when spin interactions decay with the distance r as $r^{-(1+\sigma)}$ for $\sigma < 1$ and nonstandard critical exponents are observed for $0.5 < \sigma < 1$ [2]. Another example is the 1D long-range q-states Potts model, where depending on the σ exponent and q parameter, a first-order or a second-order phase transition is possible [3].

Some properties of 1D spatial systems with long-range interactions can be mapped to N-step (long memory) Markov chains, where transitional probabilities depend on the system history and the spatial variable corresponds to the time axis. Analytical and numerical solutions for the resulting time-dependent probability distributions were presented in Refs. [4,5] for fixed values of time horizon N. The formalism was extended [6,7] to an infinite-range memory that covers the whole history of a 1D random walker. In such a case, a dynamical phase transition takes place between normal diffusion and superdiffusive behavior. When the parameter describing the influence of memory is small enough, the variance D_L of a walker position scales with the walking time L as $D_L \sim L$. It increases however as $D_L \sim L^{\kappa}$, $\kappa > 1$ when the memory influence parameter crosses a critical value. The results can explain a persistent behavior of coarse-grained DNA sequences, written texts, and financial data [6].

In this work, we consider a stochastic model of preferential cluster growth where a special form of long-memory dynamics results from recent observations of emotional patterns in discussions in online communities [8–13]. In fact, complex phenomena taking place during information search and communication exchange over the internet have been investigated by several authors and diverse methods of statistical physics, see, e.g., Refs. [14–20]. The studies are facilitated by an easy access to massive data sources [21,22]. Diffusion of information and opinion in online communities is frequently compared to epidemiological phenomena [23–29]. However, both processes need separate approaches, as shown, e.g., in recent investigations [30,31] of social contagion in online social networks that emerged during political protests in Spain.

Our model is based on a special collective phenomenon of emotional interactions reported in Ref. [11]. Consecutive comments posted on blogs, the BBC Forum, IRC channels, and the Digg website when represented by binary variables corresponding to posts' emotional valences [32–34] tend to group in clusters of a similar valence and the cluster growth rate can be well described by a sublinear preferential rule [11]. It follows that a negative comment is more likely to be posted after a sequence of five negative messages rather than after four such posts. The persistent dynamics of this system has been confirmed by the Hurst exponent analysis [10]. The aim of this paper is to study the statistical behavior of this system when affective interactions are strong and long clusters are present. In particular we will investigate situations when in a given course of time the process of preferential cluster growth leads to the emergence of a critical cluster followed by posts always displaying the same valence, and secondly, what is a fraction of such an ordered phase in all posts.

This paper is organized as follows. In Sec. II we describe observations of emotional clusters in massive data sets, in Sec. III we define a data-driven model for posts' appearance, and in Sec. IV we present numerical simulations showing a transition between a mostly disordered (hetero-emotional) and a mostly ordered (mono-emotional) phase in a two-state case of such a model. The model extension to a three-state system is studied in Sec. V. In Sec. VI we analyze data sets from selected online communities in order to demonstrate the presence of a mono-emotional phase.

II. PREFERENTIAL GROWTH OF EMOTIONAL CLUSTERS IN REAL DATA

In accordance with the behavior found in several online communities (BBC Forum [35,36], Digg, IRC, Blog06) and presented in Refs. [11,13], the preferential growth mechanism is the main process responsible for forming emotional clusters. The process is manifested by the power-law formula for the conditional probability $p(e|ne)$ that after n comments with the same emotion e [32–34] the next comment will express a similar sentiment. The data (see Fig. 1) reveals the relation $p(e|ne) = p(e|e)n^{\alpha}$ where $p(e|e)$ is the conditional probability that two consecutive messages have the same emotion $e = -1,0,1$ (negative, neutral, positive); $p(e|e)$ is defined by $p(e|e) = p(ee)/p(e)$, where $p(ee)$ is the joint probability of the pair ee that is measured as a number of occurrences of the two consecutive messages with the same valence e divided by the number of all appearing pairs. Here $p(e)$ is the probability of a given emotion e measured as the number of comments with the valence e divided by the total number of comments in the considered data. For the description of automatic sentiment analysis applied to the data retrieval see Refs. [11,37–39]. The characteristic exponent α represents the strength of the preferential process leading to a long-range attraction between posts of the same emotion. The probability of finding the cluster of size n is proportional to the factor $C = p(e)p(e|e)^{n-1}[(n-1)!]^{\alpha}$ responsible for the appearance of the sequence of n consecutive messages. It should be also taken into account that the cluster of size n is defined as exactly n posts with mono-emotional expressions.

III. MODEL DESCRIPTION

Here we try to simulate the process of preferential cluster growth in an artificial environment. To make the problem simpler for further analytical investigations, we shall usually consider a two-state system where only positive $e = 1$ or negative $e = -1$ messages can appear in such an artificial discussion. In Sec. V we will show that the behavior of a three-state model is similar. Each thread has the same length L, unlike in real data, where the thread distribution was close to a power-law function (see Supporting Material in Refs. [11] and [13]).

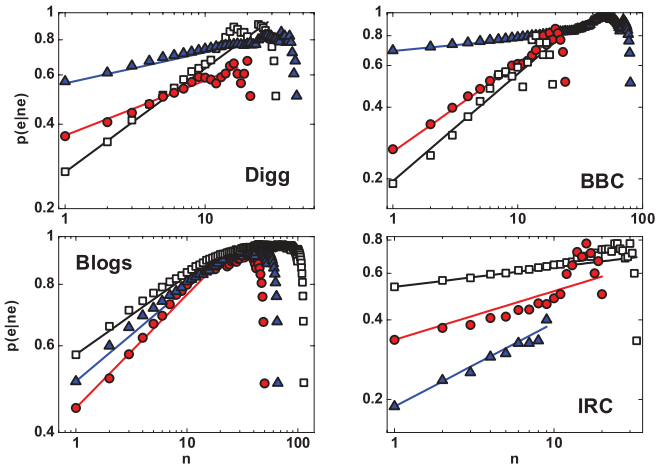

FIG. 1. (Color online) The conditional probability $p(e|ne)$ of the next comment occurring with the same emotion e for Digg, BBC, blogs, and IRC data [11,13]. Symbols are data (blue triangles, red circles, and white squares, for negative, positive, and neutral clusters, respectively), and lines represent the fit to the preferential attraction relation $p(e|ne) = p(e|e)n^{\alpha}$. Values of $p(e|e)$ and α are shown in Table I.

The evolution rules of this two-state system are as follows.

(i) the emotion in the first message is randomly chosen with even probabilities $p(e = 1) = p(e = -1) = 1/2$.

(ii) the probability of emotion e in the next message is dependent on the discussion history. Information about this history is coded in size n of a recently observed emotional cluster. The cluster of size n is defined as a subchain of length n of consecutive states with the same values as the valences e [11].

(iii) The process of the cluster growth is based on the behavior observed in real data. The conditional probability that a cluster containing n consecutive messages with the same valence e increases its length to $n + 1$ is given by the equation

$$p(e|ne) = x_e n^{\alpha_e}, \tag{1}$$

where x_e is a constant that can depend on the cluster valency e. It amplifies the cluster growth rate and equals to a one-step memory parameter $p(e|e)$ that can be calculated from real data. Usually we will disregard the dependence x_e on the valence e and use a valence independent value x for simulations and analytical calculations. The exponent α_e, where $0 < \alpha_e < 1$ describes a strength of preferential interactions for the emotion e as it was described in Sec. II. In the numerical simulations of emotional patterns in each time step we randomly choose a value between [0; 1]. If it is smaller than $p(e|ne)$, then the cluster of the emotion e is continued; otherwise, the cluster is terminated, and the opposite emotion $(-e)$ appears.

(iv) if $p(e|ne) = 1$, then the cluster reaches its critical size n_c,

$$n_c^e = (x_e)^{-1/\alpha_e}, \tag{2}$$

which means that from this moment on the discussion will be permanently ordered and that all following messages in this thread will possess the same emotional valence e.

Since the critical cluster usually does not start in the beginning of a thread, a characteristic time T_c can be thus defined when the cluster reaches its critical length n_c. In numerical simulations we shall use the $\langle T_c \rangle$ as the average over R realizations (threads); in almost all cases we take $R = 10^4$.

IV. TWO-STATE SYSTEM

Unless otherwise stated we consider the simplest case $x = x_1 = x_{-1} = 0.5$ and $\alpha_{-1} = \alpha_1 = \alpha$. The probabilities of the appearance of both emotions when calculated in an unordered phase (before the critical cluster occurrence) are the same $p(-1) = p(1) = 0.5$, and the distribution of the observed cluster lengths is very similar to the one observed in real data.

After transition time T_c, i.e., when the critical cluster appears, the discussion changes into a mono-emotional thread (MET). Starting here, the probabilities $p(-1)$ and $p(1)$ become 0 and 1 (or 1 and 0). This means that half of the threads is nearly

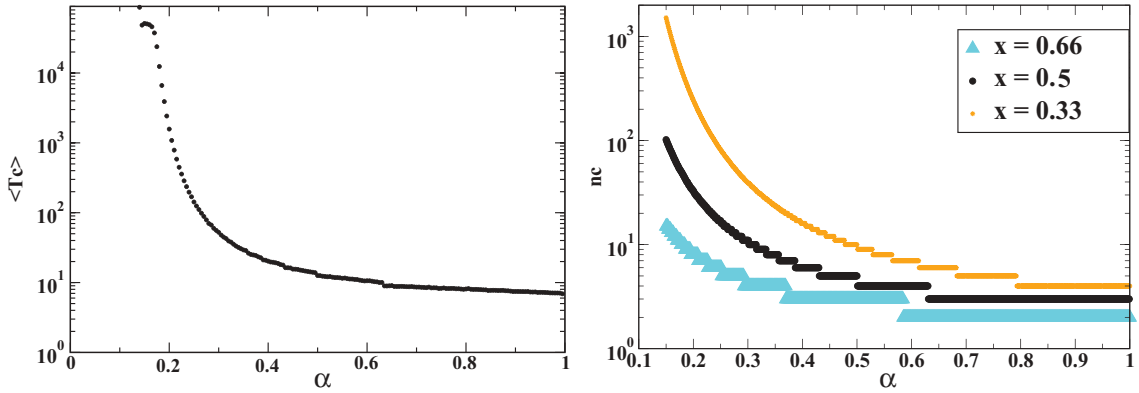

FIG. 2. (Color online) Left: time T_c needed for the emergence of the critical cluster for $x = 0.5$, $L = 10^7$. Right: size of critical cluster as a function of α for $x = 0.66$ (sky blue), $x = 0.5$ (black), and $x = 0.33$ (orange) (from bottom to top).

all positive, and the other half nearly all negative (if the threads are long enough). It is obvious that the average critical time $\langle T_c \rangle$ should depend on the strength of emotional interactions, i.e., on the exponent α. It is also obvious that $\langle T_c \rangle$ has to be larger or equal to the critical size of the cluster $\langle T_c \rangle \geqslant n_c$ (see Fig. 2). Values of $\langle T_c \rangle$ are received from numerical simulations and n_c from Eq. (1).

Since in some threads the critical cluster is not observed at all, $\langle T_c \rangle$ is not an appropriate observable, and a more convenient measure is a mean inverse of the critical time

$$\langle \lambda(x,\alpha) \rangle = \frac{1}{\widetilde{R}} \sum_{i=1}^{i=\widetilde{R}} \frac{1}{T_c^i}, \tag{3}$$

where $\widetilde{R}$ is the number of threads that were ordered during the simulation, which means that their critical times were smaller than the thread's length. In Fig. 3 we present a relation between $\langle \lambda \rangle$ and α. The left plot is presented in the linear scale and clearly displays the staircase shape of this dependence that follows from the integer values of T_c (compare Fig. 2). The right plot presents in the log-linear scale a rapid decrease in $\langle \lambda \rangle$ for $\alpha \approx 0.15$. The multisteps shape for $\alpha > 0.3$ and a rapid decrease observed for $0.13 < \alpha < 0.2$ are only weakly dependent on the system size L. We tested this behavior for different values of L; for clarity, we show only representative simulations for $L = 10^6$, $L = 2 \times 10^7$, and $L = 5 \times 10^7$. Of course, the length of the thread L influences the value α when the order is observed for the first time in our ensemble of $R = 10000$ samples. It is $\alpha = 0.13$ for a system of the size $L = 5 \times 10^7$ and $\alpha = 0.15$ when $L = 10^3$.

Since parameter $\langle \lambda \rangle$ is the average $\langle 1/T_c \rangle$ it thus equals to a probability that a comment is at the end of the critical cluster. On the other hand, the probability of finding the critical cluster can be obtained from a distribution of cluster sizes

$$P(n) = A(x,\alpha) x^n [(n-1)!]^{\alpha}, \tag{4}$$

that is similar to a relation presented in [11]. It follows

$$\langle \lambda(x,\alpha) \rangle = P(n_c), \tag{5}$$

where n_c is given by Eq. (2).

The inverse of normalization constant in Eq. (4)

$$A(x,\alpha) = \left[\sum_{n=1}^{n=n_c} x^n [(n-1)!]^{\alpha} \right]^{-1} \tag{6}$$

was calculated numerically and is presented in Fig. 4. A compact analytical approximation for this constant can be received when $\alpha = 0$ and $x < 1$. Then we get

$$A(x,\alpha) \approx [x/(1-x)]^{-1}. \tag{7}$$

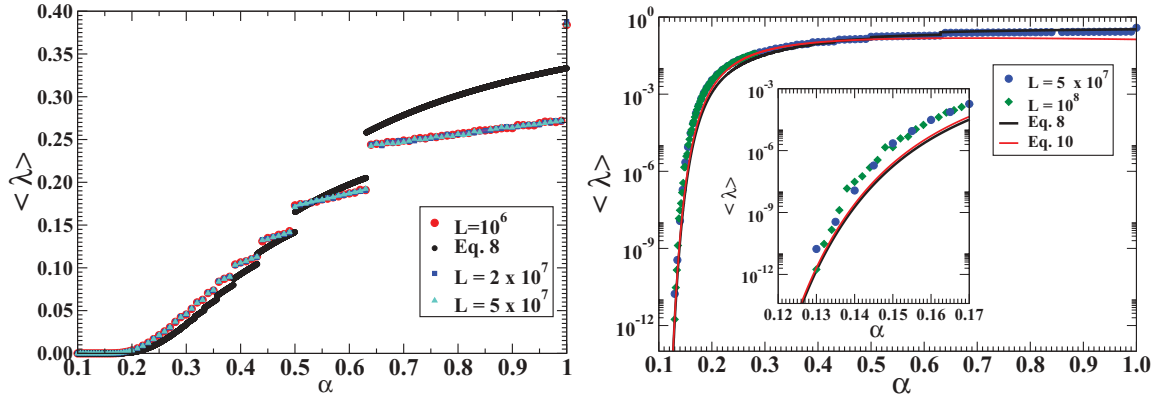

FIG. 3. (Color online) Relation between the inverse of the critical time $\langle \lambda \rangle$ and the exponent of affective interactions α for $x = 0.5$ for different values of discussions lengths L. Red circles: $L = 10^7$, blue squares: $L = 2 \times 10^7$, sky blue triangles: $L = 5 \times 10^7$. Black circles follow from Eq. (8) and are very close to the red line from Eq. (10).

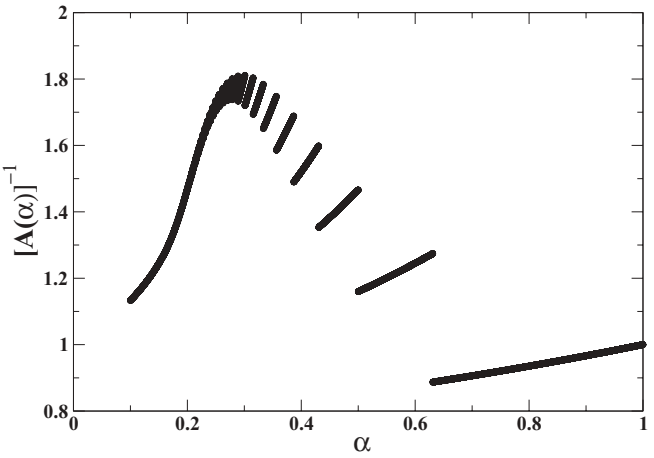

FIG. 4. Values of $1/A(\alpha)$ estimated for $x = 0.5$ using MATHEMATICA.

In fact for $x = 1/2$ the above approximation gives $A = 1$, which is not very far away from exact numerical values presented at Fig. 4 especially when $\alpha \ll 1$.

After using Eq. (2) in Eq. (4) we get from Eq. (5)

$$\langle\lambda(x,\alpha)\rangle = A(x,\alpha)x^{x^{-1/\alpha}}[(x^{-1/\alpha}-1)!]^{\alpha} \tag{8}$$

that well fits the behavior of $\langle\lambda(\alpha)\rangle$ received from numerical simulations (see the right panel in Fig. 3). The value of $\langle\lambda(\alpha=1)\rangle$ can not be obtained from Eq. (8) but may be easily calculated from a simple branching process as

$$\langle\lambda(x=0.5,\alpha=1)\rangle = 2\sum_{n=2}^{n=n_c}\left(\frac{1}{2}\right)^n\frac{1}{n} = 2\ln 2 - 1 \approx 0.386. \tag{9}$$

In the limit $\alpha \ll 1$ Eq. (8) reduces to

$$\langle\lambda(x,\alpha)\rangle \approx (1-x)\exp(-\alpha x^{-1/\alpha}) \tag{10}$$

and we get $\langle\lambda(x,0)\rangle = 0$

Let us consider discussions in an ensemble of threads of length L with affective interactions described by the characteristic exponent α and the parameter x and let us define a fraction of discussions that are mono-emotionally ordered from certain moments as $r(\alpha,x,L) = \frac{\tilde{R}}{R}$. This value is also the probability of the MET occurrence before time $t = L$. It follows the value of r can be written as

$$r(\alpha,x,L) = 1 - [1-\lambda(\alpha,x)]^L, \tag{11}$$

where an explicit form can be received by inserting into Eq. (11) results (7) and (8)

$$r(\alpha,x,L) \approx 1 - [1-(1-x)\exp(-\alpha x^{(-1/\alpha)})]^L. \tag{12}$$

Results of numerical simulations and Eq. (12) are presented in Fig. 5. As one could expect, a fraction r of the MET phase in all threads increases with the increase of α exponent and thread length L. Moreover for longer threads the agreement between Eq. (12) and numerical simulations is better and the transition between the states $r \approx 0$ and $r \approx 1$ is steeper. In the thermodynamical limit $L \to \infty$ this transition becomes

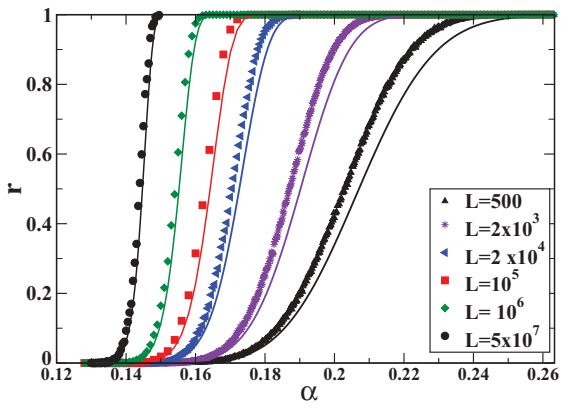

FIG. 5. (Color online) Fraction of ordered threads as a function of the exponent α for various thread lengths L. Lines correspond to Eq. (12).

discontinuous since

$$\lim_{L\to\infty}\lim_{\alpha\to 0^+} r(\alpha,x,L) = 0 \tag{13}$$

and

$$\lim_{L\to\infty} r(\alpha>0,x>0,L) = 1. \tag{14}$$

Let us define the characteristic value α^* as a strength of affective interactions for which we observed a half of ordered realizations $\alpha^* = \alpha(r = 0.5)$. After a short algebra we get from (12)

$$1-(1-x)\exp[-\alpha^* x^{(-1/\alpha^*)}] \approx 2^{-(1/L)}. \tag{15}$$

For the symmetrical case $x = 1/2$ and $L \gg 1$ (unless otherwise stated we use the same assumptions further in analytical calculations) we get a simpler relation

$$\alpha^* 2^{(1/\alpha^*)} \approx \ln(L) - \ln[2\ln(2)] \tag{16}$$

that can be disentangled as

$$\alpha^* \approx \frac{-\ln(2)}{W_{-1}\{-\ln(2)/\ln[L/\ln(4)]\}}, \tag{17}$$

where $W_{-1}(.)$ is the lower branch of Lambert W function [40]. A quantitative measure of the system behavior near α^* is the slope

$$\tan\phi = \left(\frac{\partial r(\alpha,x,L)}{\partial\alpha}\right)_{\alpha^*} \tag{18}$$

that can be expressed as

$$\tan\phi \approx -\frac{\ln(2)}{2}x^{-1/\alpha^*}\left[1+\frac{\ln(x)}{\alpha^*}\right]. \tag{19}$$

For $x = 1/2$ Eq. (19) can be written as an explicit function of the length L using the result (17). Relations (17) and (19) are presented at Fig. 6 where we see a good fit to corresponding numerical simulations. In the limit $L \to \infty$ the value $\alpha^*(L)$ calculated from (17) tends to zero while the slope $\phi(L)$ diverges to infinity, which is a sign of a discontinuous transition in the thermodynamical limit. It should be stressed that for $\alpha = 0$ the MET phase does not exist, which is shown in Eq. (13).

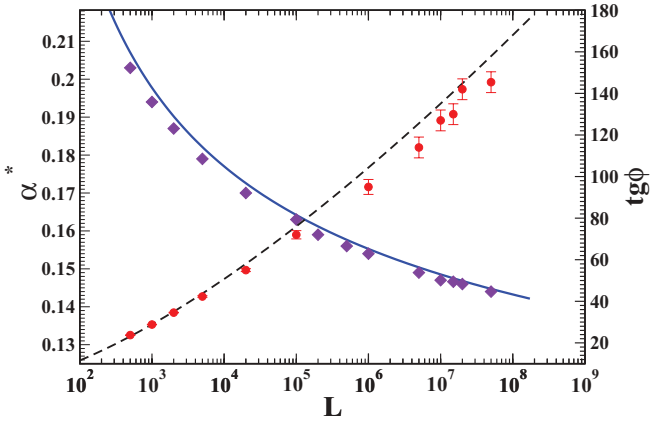

FIG. 6. (Color online) Dependence of the critical value α^* (violet diamonds) and the slope of $\tan\phi$ (red circles) on the system size L. Solid line corresponds to Eq. (17) and a dashed one to Eq. (19) where the value of α^* was taken from Eq. (17).

Until now we were focused on the relation of the fraction r of ordered realizations and the exponent α for a fixed value of the parameter x corresponding to the system short memory $p(e|e)$ [see Eq. (1)]. Figure 7 shows the influence of x parameter on the fraction r where we used the solution Eq. (12). As it could be expected the increase of the short memory makes the MET occurrence more probable. In the thermodynamical limit $L \to \infty$ we have a discontinuous transition similar to that observed at Fig. 5. In fact we get from Eq. (12)

$$\lim_{L\to\infty}\lim_{x\to 0^+} r(\alpha,x,L) = 0, \tag{20}$$

which should be compared to Eq. (13) and Eq. (14).

To show the combined influence of parameters α and x on the system dynamics we have presented the results of simulations for pairs (x, α) fulfilling the condition $r = 0.5$ and solutions of Eq. (12) for different values of L (see Fig. 8). There is a good agreement between numerical simulations and the approximate analytical solution (12) especially for $\alpha \ll 1$. As we can see in Fig. 8 a large fraction of MET phase can

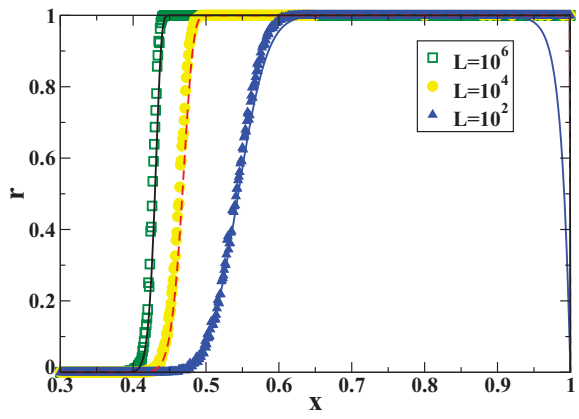

FIG. 7. (Color online) Fraction of ordered threads as a function of the parameter x with fixed value $\alpha = 0.2$ for various thread lengths L. Lines correspond to Eq. (12). A decay of the function $r(\alpha,x)$ for $x \to 1^-$ is an artifact following from Eq. (7).

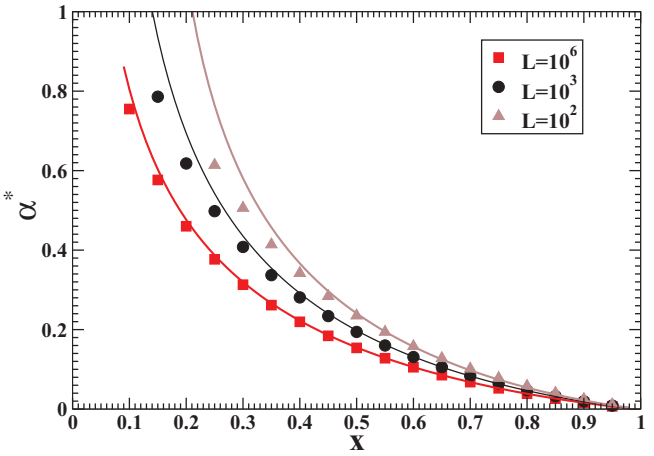

FIG. 8. (Color online) For each value of x we find a corresponding value of α when half of realization are ordered, $r = 0.5$. Points represent numerical simulations for different values L and lines correspond to Eq. (12).

emerge only if both parameters α and x are high enough. For $x < 0.3$ the parameter α needed to observe a half of ordered realizations should be higher than 0.3 even for very long threads $L = 10^6$.

Let us note that derivatives of fraction r in respect to α and x can be considered as corresponding system susceptibilities. A special case is presented by Eq. (19) where we see that the susceptibility against the exponent α calculated at the characteristic point α^* tends to infinity in the thermodynamical limit $L \to \infty$. A similar divergence can be derived by differentiating Eq. (12) against x and assuming the same limit.

V. THREE-STATE SYSTEM

A natural extension of the two-state system is to add one more state, i.e., $e \in \{-1,0,1\}$. To compare properties of such systems with our previous results, we compare a symmetrical three-state model where $x_{-1} = x_0 = x_1 = 0.5$ and $\alpha_{-1} = \alpha_0 = \alpha_1$ with a symmetrical two-state model where $x_{-1} = x_1 = 0.5$ and $\alpha_{-1} = \alpha_1$. Values of the inverse of critical time $\langle\lambda\rangle$ as a function of the exponent α are presented in Fig. 9. Since results for both systems follow the same line, we can state that the number of possible emotional states does not influence a critical time needed for the emergence of MET. This observation can be explained as follows: the occurrence of MET requires growth of a critical cluster of any emotion e. The growth process is dependent only on the conditional probability of cluster growth [Eq. (1)] that is insensitive to the number of possible emotional states. If initial probabilities $p(e)$ of a spontaneous occurrence of every emotional state e are equal and clusters of posts representing each emotion possess identical growth parameters α_e and x_e then an average time needed for the emergence of *any* critical cluster should be independent from the number of possible emotional states.

Figure 10 shows the results for an asymmetrical three-state system with $x_{-1} = x_0 = x_1 = 0.33$. We considered models when one or two emotional states are random ($\alpha_{-1} = 0$ or/and $\alpha_0 = 0$) and the preferential process appears only for the

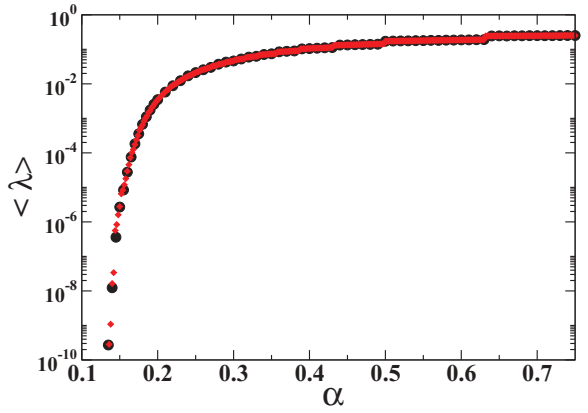

FIG. 9. (Color online) Relation between the observable $\langle\lambda\rangle$ and the exponent α; large black circles: two-state system with parameters $x_{-1} = x_1 = 0.5, \alpha_{-1} = \alpha_1$; small red points: three-state system with $x_{-1} = x_1 = x_0 = 0.5$ and $\alpha_{-1} = \alpha_1 = \alpha_0$.

remaining emotional state. We observe that for a small value of $\alpha < 0.25$ all three considered curves collapsed.

VI. REAL-WORLD DATA

Here we compare theoretical predictions of the MET phase emergence with data corresponding to various online communities: BBC Forum [12,35,36], Digg [41], and Blog [11]. BBC Forum data included discussions posted on the Religion and Ethics and World/UK News message boards starting from the launch of the website (July 2005 and June 2005 respectively) until June 2009. The Blog06 dataset is a subset of the collection of blog posts from December 6, 2005 to February 21, 2006. Only posts attracting more than 100 comments were extracted, as these seemed to initiate nontrivial discussions. The Digg dataset comprises a full crawl of digg.com, one of the most popular social news websites. The data spans February to April 2009 and consists of all the stories, comments, and users that contributed to the site during this period. The data was emotionally annotated by a sentiment classifier, for details of its properties see Refs. [11,37–39,41].

One of differences between our model and real data is the thread length distribution. In the simulation we analyzed

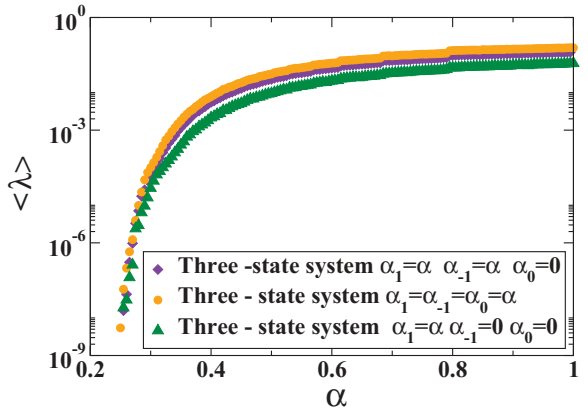

FIG. 10. (Color online) Relation between the observable $\langle\lambda\rangle$ and the exponent α for asymmetrical three-state system for $x = 0.33$ and different values of α; $L = 2 \times 10^6$.

TABLE I. Values of model parameters fitted to real data sets Blog06, BBC Forum, and Digg. α: exponent of preferential emotional interactions. $p(e|e)$: conditional probability of the appearance of two consecutive messages with the same emotional values. n_c: size of the critical cluster (needed for MET emergence) predicted by the model. $n_{\max}$: the maximum cluster size found in data sets. m^d_{th}: the number of threads containing the critical cluster in data sets. The last column m^s_{th} is the average number of threads with the critical cluster received from numerical simulations on the real structure of the data, i.e., the lengths of threads were acquired from the data with parameters α and $p(e|e)$. Simulations were repeated $R = 1000$ times for every thread.

	α	$p(e\|e)$	n_c	$n_{\max}$	m^d_{th}	m^s_{th}
Blog06_+	0.23 ± 0.01	0.45	33	51	4	26
Blog06_-	0.19 ± 0.01	0.51	35	67	4	57
Blog06_0	0.16 ± 0.01	0.58	31	114	35	217
BBC_+	0.38 ± 0.02	0.27	32	25	0	2.08
BCC_-	0.051 ± 0.005	0.69	1672	81	0	0
BBC_0	0.45 ± 0.04	0.2	36	20	0	0.04
Digg_+	0.20 ± 0.01	0.37	115	22	0	0
Digg_-	0.11 ± 0.01	0.56	195	46	0	0
Digg_0	0.37 ± 0.04	0.27	35	33	0	0.001

systems with a fixed length L and various cases were tested for L between $L = 100$ and $L = 10^8$. In real data sets the thread length occurrence is described by specific distributions as presented in Fig. 11. The maximum observed value of L is around 10^3 for Digg and Blog06 and approximately 10^4 for BBC Forum, but those values appear only occasionally. Generally the character of these distribution is close to a power-law decay, thus, the majority of data comprises short discussions.

In Table I, for three data sets we present the values of exponents α (see Ref. [11]), one-step memory parameters $p(e|e)$, values of critical clusters n_c following from Eq. (2) and $n_{\max}$, i.e., maximum sizes of clusters observed in real data. The first step of our data analysis was to look for threads containing a cluster larger than a critical one. Assuming that only clusters appearing at the end of discussions can be treated as examples of a MET phase in real data we have found threads with $n_{\max}$ larger than n_c only in Blog06 data (see Table I).

The appearance of such clusters in Blog06 is in agreement with the analysis presented in Fig. 12, where results of simulations make up the colored background for real pairs (x, α) displayed by various symbols. In case of data corresponding to negative and neutral clusters in the BBC Forum, and negative and positive clusters in Digg the length of the system needed to statistically observe half of realizations with MET is larger than $L = 10^6$. In real Digg data we do not find discussions of such length thus the absence of MET phase is not puzzling for this community. A similar situation takes place in case of the neutral clusters in Digg and positive clusters in BBC. Here one needs threads less than $L = 10^6$ but more than 10^3. On the other hand for positive and negative clusters in Blog06 the data points lie very close to $L = 10^3$ line, while for neutral Blog06 clusters the system size necessary to find half of the ordered threads is only slightly larger than 10^2. These conditions are in a qualitative agreement with observations of

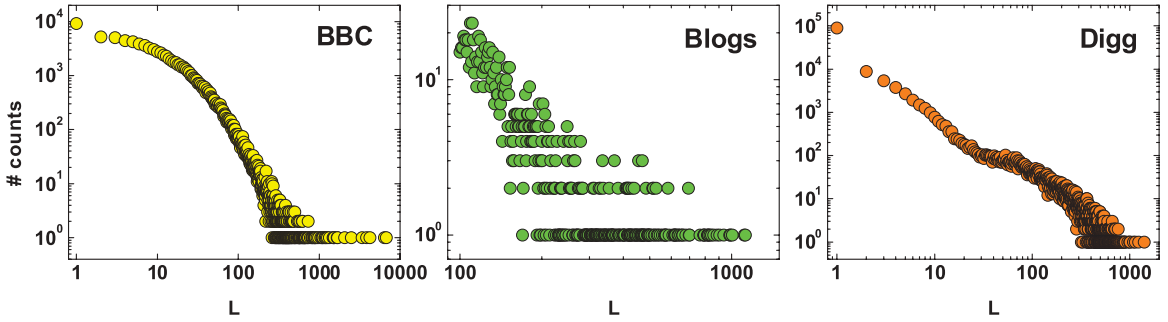

FIG. 11. (Color online) Histograms of thread lengths L for BBC Forum, Blog06, and Digg (see Supporting Material in Ref. [11]).

the MET phase in this community. For Blog06 data we found 35 threads with neutral clusters larger than n_c, negative clusters larger than n_c were found in four threads and positive clusters appeared also in four threads.

In Blog06 data set we find threads containing two or three clusters larger than n_c in the same thread. It means that even after the critical cluster emerged, hetero-emotional comments may afterwards be posted. Such a behavior is not predicted by our model, where after the mono-emotional cluster crosses its critical size no other emotional state can be observed. To explain this discrepancy let us point out that in case of real data there exist fluctuations that might result from mistakes in the classification algorithm [41,42], and also from a random emotional behavior of participants that is not provided for by model.

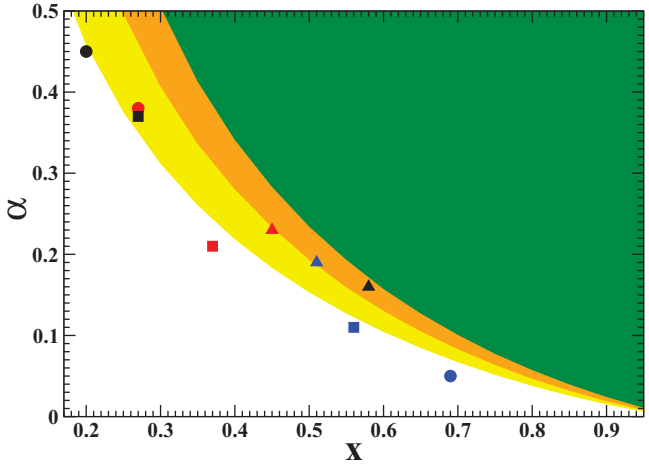

FIG. 12. (Color online) Comparison of conditions for the MET phase emergence in the model and in real data. Results of numerical simulations show pairs of parameters (x, α) for which a half of simulated threads were ordered ($r = 0.5$) and they are displayed by various colors for different threads lengths L. The white color represents $L > 10^6$, the yellow $10^6 < L < 10^3$, the orange $10^3 < L < 10^2$, and the green $L < 10^2$ (similar results can be found at Fig. 8). Symbols present values (x, α) calculated from real data (see Fig. 1 and Table I). BBC Forum is marked with the circles, Digg with the squares, Blog06 with the triangles, the red color presents positive emotions, the blue negative, and black is used for the neutral. In the case of real data the parameter x was estimated as $(p(e|e)$. The plot suggests that the best conditions for the MET phase occurrence are in the Blog06 data set.

The last column of Table I follows from numerical simulations performed on the thread length distribution taken from real data. The average number $m^s_{\rm th}$ of threads with a critical cluster obtained in this way is always larger than the value observed in real data $m^d_{\rm th}$. This discrepancy also follows from fluctuations that are absent in our model. In real data such fluctuations mean splitting of an emotionally homogeneous state into two or more parts by one or a few random messages with other emotional valence. If all separate parts were shorter than n_c our search algorithm did not detect MET in a given thread and thus the number of detected MET cases is lower. The effect of fluctuations is demonstrated in Fig. 13. The left panel is an example of a *clean* MET, where at the end of the discussion the users post only messages with a mono-emotional expression. The right panel is an example of a *noisy* MET where during the presumably mono-emotional phase messages with different emotions randomly appear, or at least are detected by the sentiment classifier. To quantify this behavior, for every thread with a critical cluster, we define a coefficient γ

$$\gamma = \frac{l^e}{l}. \tag{21}$$

Here l^e is a number of messages with a dominant emotional state e (corresponding to the critical cluster) observed in the part of the thread starting from the critical cluster, whereas l is a number of all messages in this part of a thread. In other words $l = l^e + l^d$ where l^d is a number of messages with emotional states different than e that were observed after the critical cluster. Let us note that $1 - \gamma$ is a measure of a noise level. Figure 14 presents values l and γ for all threads containing the MET phase. Mean values of this ratio for clusters of various emotions are $\langle\gamma\rangle_+ = 0.874$, $\langle\gamma\rangle_- = 0.932$, and $\langle\gamma\rangle_0 = 0.969$. It means that for a vast majority of threads the critical cluster was located close to the discussion end and not many posts expressing other emotions following critical clusters were observed.

Let us assume that the calculated ratio $\langle\gamma\rangle_e$ is the probability of a single post with emotion e being not affected by an additional independent and identically distributed random process. Such a process has been neglected in our model although it could lead to additional fluctuations disrupting the MET phase. One can then estimate a suppressing factor Γ that

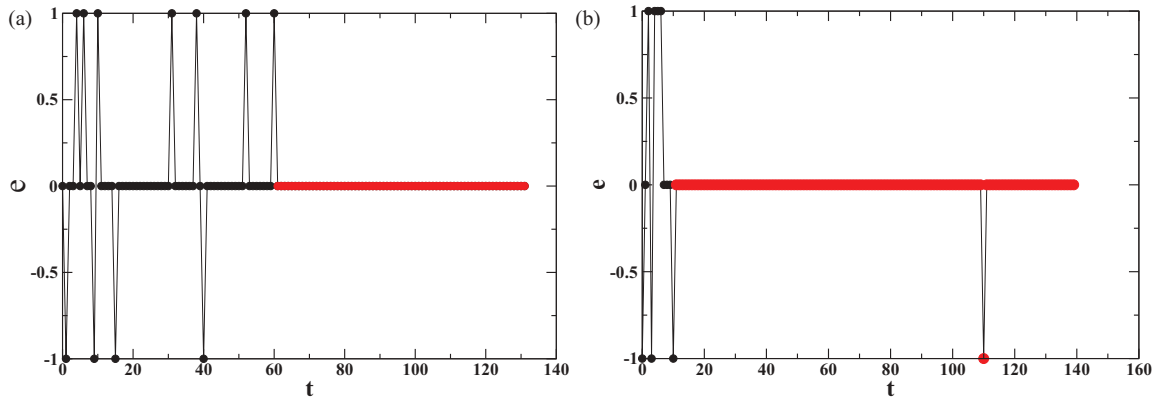

FIG. 13. (Color online) Examples of the threads from Blog06 where l are marked in red. (a) a thread with a cluster of the size $n_{\max} = 72$ that closed the discussion ($\gamma = 1$). (b) a cluster with the size $n_{\max} = 99$ that started a discussion with $l = 129$ comments, where $l^e = 128$ messages expressed the same emotional values and one message was different. The γ coefficient for this thread was 0.992.

limits the observed number of MET events

$$\Gamma(\langle\gamma\rangle_e) = \langle\gamma\rangle_e^{n_c^e}. \tag{22}$$

Using the data from Table I and the $\langle\gamma\rangle_0$ ratio estimated above we get the value $\Gamma_0 = 0.39$ for neutral MET in Blog06. This suppressing factor can be compared to the ratio of a number of observed MET events to a number of predicted MET events following Table I $\Delta_0 = m_{\text{th}}^d/m_{\text{th}}^s = 0.16$. Although our approach to estimate the effects of fluctuations is elementary we receive a fairly good agreement between parameter values Γ_0 and Δ_0. The small statistics of positive and negative METs events ($m_{\text{th}}^d = 4$) was an obstacle for the noise level estimation in those cases.

In summary, the MET phase can be detected in real data sets if affective interactions described by parameters $p(e|e)$ and α are strong enough and the threads are long. In our case the necessary condition was met only for Blog06 data and here indeed we have observed a number of MET cases, however the effects of fluctuations suppress their emergence.

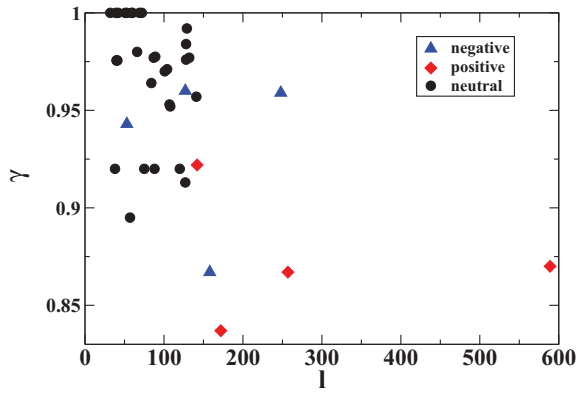

FIG. 14. (Color online) Relation between γ coefficients and MET cluster sizes l for Blog06. Red diamonds correspond to positive MET events, blue triangles to negative events, and black circles to neutral events.

VII. CONCLUSION

We studied a specific long-memory stochastic process that represents a data-driven binary model of emotional online discussion threads. The dynamics is described by non-negative parameters x and α corresponding respectively to the single-step memory and to the characteristic exponent of the preferential process of mono-emotional cluster building. Analytical and numerical calculations show that in such a model persistent mono-emotional threads (MET) can emerge when a cluster reaches a critical size n_c. This phenomenon takes place in time T_c independent from the system size L. It follows that the longer the thread the more likely it is that it will be ordered. For finite threads fraction r of MET-containing threads increases continuously with parameters x and α. However, in the thermodynamical limit $L \to \infty$ there is a discontinuous transition between a phase without mono-emotional threads $r = 0$ and a completely ordered phase $r = 1$. The transition takes place when either $x \to 0_+$ or $\alpha \to 0_+$. If we consider fraction r as an order parameter then corresponding susceptibilities diverge for $L \to \infty$ in such transition points. The extension of the model to a three-state dynamics does not change its main properties, e.g., the critical time T_c depends in a similar way on the emotional interaction exponent α.

The behavior of our model resembles some features of a dynamical phase transition induced by a long memory in Markov chains studied in Refs. [4–7]. The observed increase of fraction r of ordered threads with parameters x and α corresponds to a transition from a diffusive to a superdiffusive behavior reported in Refs. [4–7]. The transition takes place when the memory correlation strength crosses a critical value. Since superdiffusion means that the variance of correlated sequence increases faster than linearly within the observation horizon, it follows that the system is more persistent in this phase. In our case the persistence corresponds to a more frequent occurrence of the MET phase (coherent segments) in various threads. There is however a crucial difference between our model and systems studied in Refs. [4–7]. In our case there is a nonzero fraction r of ordered threads for any $\alpha > 0$ and $x > 0$ and thus critical values of these parameters can not be uniquely defined. Another difference to the phenomena observed in Refs. [4–7] is an intrinsic nonstationarity of our

model. The MET phase emerges in the course of time thus a thread can consist of two different phases. This is not the case of systems studied in Refs. [4–7].

It is interesting that the emergence of MET phase was observed for 43 threads in the Blog06 data set and the presence of this phase could be explained by our model if corresponding parameter values were used. The number of MET events is much lower compared to theoretical estimations, which can by explained be effects of fluctuations and sentiment classification errors. The absence of the MET events in BBC and Digg data sets is consistent with analytical and numerical calculations of MET density in our model.

ACKNOWLEDGMENTS

The work was supported by EU FP7 ICT Project *Collective Emotions in Cyberspace - CYBEREMOTIONS*, European COST Action MP0801 *Physics of Competition and Conflicts*, Polish Ministry of Science Grants No. 1029/7.PR UE/2009/7 and No. 578/N-COST/2009/0 and a special grant of Faculty of Physics of Warsaw University of Technology. A.C.H. would like to thank Julian Sienkiewicz for his useful advice on how make the simulation faster. We are much obliged to the anonymous referees for their critical comments, which allowed us to improve the above paper.

[1] L. D. Landau and E. M. Lifshitz, *Statistical Physics*, 3rd ed. (1980) *Part 1* (Elsevier Butterworth Heinemann, Oxford, 2010).
[2] D. J. Thouless, Phys. Rev. **187**, 732 (1969).
[3] E. Bayong, H. T. Diep, and V. Dotsenko, Phys. Rev. Lett. **83**, 14 (1999).
[4] O. V. Usatenko and V. A. Yampolskii, Phys. Rev. Lett. **90**, 110601 (2003).
[5] O. V. Usatenko, V. A. Yampolskii, K. E. Kechedzhy, and S. S. Melnyk, Phys. Rev. E **68**, 061107 (2003).
[6] U. Keshet and S. Hod, Phys. Rev. E **72**, 046144 (2005).
[7] S. Hod and U. Keshet, Phys. Rev. E **70**, 015104(R) (2004).
[8] P. Sobkowicz and A. Sobkowicz, Eur. Phys. J. B **73**, 633 (2010).
[9] M. Mitrović, G. Paltoglou, and B. Tadić, Eur. Phys. J. B **77**, 597 (2010).
[10] A. Garas, D. Garcia, M. Skowron, and F. Schweitzer, Sci. Rep. **2**, 402 (2012).
[11] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, and J. A. Hołyst, PLoS ONE **6**(7), e22207 (2011).
[12] A. Chmiel, P. Sobkowicz, J. Sienkiewicz, G. Paltoglou, K. Buckley, M. Thelwall, and J. A. Hołyst, Physica A **390**, 2936 (2011).
[13] J. Sienkiewicz, M. Skowron, G. Paltoglou, J. A. Hołyst, J. Phys. Conf. Ser. **410**, 012001 (2013).
[14] A. Chmiel, K. Kowalska, and J. A. Hołyst, Phys. Rev. E **80**, 066122 (2009).
[15] A.-L. Barabási, Nature **435**, 207 (2005).
[16] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási, Phys. Rev. E **73**, 066132 (2006).
[17] B. Gonçalves and J. J. Ramasco, Phys. Rev. E **78**, 026123 (2008).
[18] A. Vázquez, J. G. Oliveira, Z. Dezsö, K. I. Goh, I. Kondor, and A.-L. Barabási, Phys. Rev. E **73**, 036127 (2006).
[19] F. Radicchi, Phys. Rev. E **80**, 026118 (2009).
[20] J-P. Onnela and N. A. Christakis, Phys. Rev. E **85**, 036106 (2012).
[21] D. Lazer *et al.*, Science **323**, 721 (2009).
[22] A. Vespignani, Science **325**, 425 (2009).
[23] B. Voelk and R. Noe, Behav. Ecol. Sociobiol. **64**, 1449 (2010).
[24] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, Physica A **374**, 457 (2007).
[25] Y. Moreno, M. Nekovee, and A. F. Pacheco, Phys. Rev. E **69**, 066130 (2004).
[26] L. Huang, K. Park, and Y. Ch. Lai, Phys. Rev. E **73**, 035103(R) (2006).
[27] B. Karrer and M. E. J. Newman, Phys. Rev. E **82**, 016101 (2010).
[28] J. Gu, W. Li, and X. Cai, Eur. Phys. J. B **62**, 247 (2008).
[29] P. S. Dodds and D. J. Watts, Phys. Rev. Lett. **92**, 218701 (2004).
[30] J. Borge-Holthoefer and Y. Moreno, Phys. Rev. E **85**, 026116 (2012).
[31] J. Borge-Holthoefer, A. Rivero, and Y. Moreno, Phys. Rev. E **85**, 066123 (2012).
[32] Emotional valence is probably the most important of emotion components. It stands usually at the center of emotion experience since the relevant aspect of any object that elicits emotions is whether we like it or not, or whether it is good for us or not. For a broader description of emotional valence see Refs. [33,34].
[33] L. A. Feldman, J. Personality Soc. Psychol. **69**, 153 (1995).
[34] R. B. Zajonc, Am. Psychol. **35**, 151 (1980).
[35] http://www.bbc.co.uk/dna/mbreligion.
[36] http://www.bbc.co.uk/dna/mbfivelive/F2148565; http://www.bbc.co.uk/dna/mbfivelive/F2148564.
[37] M. Thelwall, D. Wilkinson, and S. Uppal, J. Am. Soc. Inf. Sci. Tech. **61**, 190 (2010).
[38] M. Thelwall, K. Buckley, and G. Paltoglou, J. Am. Soc. Inf. Sci. Tech. **63**, 163 (2012).
[39] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, J. Am. Soc. Inf. Sci. Tech. **61**, 2544 (2010).
[40] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, Adv. Comput. Math. **5**, 329 (1996).
[41] G. Paltoglou, S. Gobron, M. Skowron, M. Thelwall, and D. Thalmann, in Proc. Engage, Springer LNCS State-of-the-Art Survey, 2010, pp 13–25.
[42] The accuracy of detection of the subjectivity amounts: 72 percent. The accuracy of detection of polarity amounts: 67.2 percent.