Advances in Complex Systems
Vol. 16, Nos. 4 & 5 (2013) 1350026 (25 pages)

DOI: 10.1142/S0219525913500264

World Scientific
www.worldscientific.com

ENTROPY-GROWTH-BASED MODEL OF EMOTIONALLY CHARGED ONLINE DIALOGUES

JULIAN SIENKIEWICZ

Faculty of Physics, Centre of Excellence for Complex Systems Research, Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland
julas@if.pw.edu.pl

MARCIN SKOWRON

Interaction Technologies Group, Austrian Research Institute for Artificial Intelligence, Freyung 6/3/1a, A-1010 Vienna, Austria
marcin.skowron@ofai.at

GEORGIOS PALTOGLOU

School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, United Kingdom
g.paltoglou@wlv.ac.uk

JANUSZ A. HOŁYST

Faculty of Physics, Centre of Excellence for Complex Systems Research, Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland
jholyst@if.pw.edu.pl

Received 28 September 2012
Revised 29 March 2013
Accepted 16 April 2013
Published 28 June 2013

We analyze emotionally annotated massive data from Internet relay chat (IRC) as well as from BBC forum website and model the dialogues between chat participants by assuming that the driving force for the discussion is the entropy growth of emotional probability distribution. This process is claimed to be responsible for a power-law distribution of the discussion lengths observed in the dialogues. We perform numerical simulations based on the noticed phenomenon obtaining a good agreement with the real data. Finally, we propose a method to artificially prolong the duration of the discussion that relies on the entropy of emotional probability distribution.

Keywords: Correlations; collective phenomena; sociophysics.

1. Introduction

The extensive records of data opened new possibilities of examining communication between humans ranging from face-to-face encounters [7, 29, 69, 72], through mobile

telephone calls [46, 78], surface-mail [45] short messages [77] to typical Internet activities such as e-mail correspondence [17], bulletin board system (BBS) dialogues [24], forum postings [33], web browsing [10] or Twitter microblogging [62].

Communication and its evolution is one of the key aspects of a modern life, which in an overwhelming part is governed by the circulation of information. In the most fundamental part, the communication is based on a *dialogue* — an exchange of information and ideas between two people [54]. Assuming an ideal situation, if the highest priority would be given to *acquiring* certain information, from a layman point of view the dialogue should be free from any additional components that could restrain conversation's participants from achieving the common goal.

In reality, a holistic view on the communication should in fact treat it rather as a *discourse*, i.e., it needs to be defined by language use, communication of beliefs and social interactions [73] or even social context [36]. In this sense, the meaning emerges through a mutual relation between communicators and their social contexts. On the other hand, early models of communication focused on the generation of meaning by words themselves, creating a system of signs, governed by rules and used to signify objects [55]. This is in fact, a very reductionism view, treating the language as made up of distinct units that can be studied in separation from their environment. Then again, using another approach, we can also treat the dialogue as an entity governed by conversational rules [25]. In this the concept of turn-taking is placed — apportioning of who is to speak next and when [54]. Recent studies in this area prove that although there are differences across the languages in the average gap between turns, all tested languages exhibit a universal behavior of avoidance to overlap and of minimizing the silence between discussion turns [70].

Clearly the approach that lays in the closest proximity to the area of complex systems is this given by Shannon and Weaver in the late 1940s [58]. In their view, called the information theory, a message is transmitted by a channel from a source to a receiver that interprets it. The channel is characterized by its bandwidth, defining the capacity and resulting level of information. Thus, a channel with high quality transmits the message itself while a poor quality channel may convey a contaminated content. Such an approach is deliberately free from taking into account the content of the message.

There is also another classification connected to dialogues. According to Buber [3] one can distinguish three different types of dialogues: genuine, technical and disguise. The third one is in fact a monologue disguised as dialogue, the first one is bound to establish a living mutual relation between the parts. For the purposes of this study the second one is the most important — it is defined by the need of objective understanding.

As a rule, in Western intellectual tradition, use of emotions cues in language is considered to be of purely rhetorical function [6], enhancing the impact exerted on the conversational partner. There are several studies with respect to use of figurative language in verbal emotional communication [20], the role of emotional

information processing in treatment [4] or more generally the observed frequencies of typical emotive words used in everyday conversations [60]. As it concerns a more quantitative view on the influence of emotions there are certain studies that show the rise in the attention that interlocutors pay to emotional words as compared to neutral ones [18]. However it seems that it can often be the nonverbal component that gives a hint about the stage or proximity dialogue's end, e.g., by the duration of mutual gaze [13].

As compared to the offline communication, the exchange of information in the Internet is claimed to be more biased toward the emotional aspect [66]. It can be explained by an online disinhibition effect [71] — the sense of anonymity that almost all Internet users possess while submitting their opinions on various fora or blogs. Nevertheless, it is the very Internet that gives the opportunity to acquire massive data, thus making it possible to perform a credible statistical analysis of common habits in communication. As the recent research shows, it is already possible to spot and model certain phenomena of the Internet discussion participants while looking just at the emotional content of their posts [8, 9, 11, 12, 15, 16, 21, 22, 42, 44, 52, 56, 74]. One of them is the collective emotional behavior [11, 21], the other is clear correlation between the length of discussion and its emotional content [11, 12, 52].

In this paper we argue that a simple physical approach based on the observation of entropy of emotional probability distribution during the conversation can serve as an indicator of a discussion about to finish. We give arguments supporting the observation of the maximum entropy rule in the emotional dialogues regardless of the type of the medium in question (i.e., negative, neutral), which results in creation of a tool that can be used to distinguish between the initial and final stage of the dialogue. The process of entropy maximization is claimed to be responsible for a power-law distribution of the discussion length and serves as a key idea for the numerical simulations of the dialogues which confirm that such assumed rules lead to good agreement between the observed and simulated discussion lengths.

The paper is organized as follows: Section 2 gives a brief description of the used data as well as of the emotional classification method, Sec. 3 presents our observations regarding the discussion length distribution, equalization of the emotional probabilities and entropy growth, in Sec. 4 we show the description of simulations rules which results are given in Sec. 5. Finally, Sec. 6 describes a potential application of the observed phenomenon. Four Appendices include precise technical details of the dialogue extraction method, assumed definitions, classifier quality and error analysis.

2. Data Description

As a source of data for analyzing online dialogues we chose the Internet relay chat (IRC) [27] logs. Some of the the major IRC channels are being automatically archived by the channel operators, the logs are often accessible to a general public,

and include the records of real-time, chat-like communication between numerous participants. The presented analysis is limited only to one of the channels, namely `#ubuntu` [28] in the period 1st January 2007 — 31st December 2009. In this work we focused on dialogues that included only two participants. The final output, after several levels of data processing (for details see Appendix A) consists of $N = 93329$ dialogues with the length L between $L_{\min} = 11$ and $L_{\max} = 339$ each. Each dialogue can be represented as a chain of messages (see Fig. 1) where all odd posts are submitted by one user and all even by another one. For the sake of comparison we also used the previously examined [11, 12] BBC Forum dataset that consists of several multiuser discussions gathered from such categories as "World News" or "Religion" between June 2005 and June 2009. Fundamental properties of both datasets are shown in Table 1.

The emotional classifier program that was used to analyze the emotional content of the discussions is based on a machine-learning (ML) approach. The algorithm functions in two phases: during the training phase, it is provided with a set of documents classified by humans for emotional content (positive, negative or objective) from which it learns the characteristics of each category. Then, during the application phase, the algorithm applies the acquired sentiment classification knowledge to new, unseen documents. In our analysis, we trained a hierarchical language model [38, 43, 57] on the Blogs06 collection [37, 47] and applied the trained model to the extracted IRC dialogues, during the application phase. The algorithm is based on a two-tier solution, according to which a post is initially classified as objective or subjective and in the latter case, it is further classified in terms of its polarity, i.e., positive or negative. Each level of classification applies a binary language model [43, 51]. Posts are therefore annotated with a single value $e = -1, 0$ or 1 to quantify their emotional content (to be more precise — their valence [19]) as negative, neutral or positive, respectively (for details on the choice of relevant values see Appendix B). The accuracy of the classifier (see Appendix C for details)

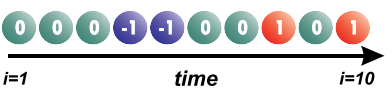


Fig. 1. (Color online) An exemplary dialogue of $L = 10$ comments. Each bullet corresponds to a comment with a negative (marked as -1), neutral (marked as 0) or positive (marked as 1) content.

Table 1. Fundamental properties of the datasets: number of comments C, number of dialogues (IRC)/discussions (BBC) N, shortest dialogue/discussion length $L_{\min}$, longest dialogue/discussion length $L_{\max}$, average valence $\langle e \rangle$, probability of finding negative, neutral or positive emotion (respectively $p(-)$, $p(0)$ and $p(+)$).

dataset	C	N	$L_{\min}$	$L_{\max}$	$\langle e \rangle$	$p(-)$	$p(0)$	$p(+)$
IRC	1889120	93323	11	339	0.17	0.15	0.53	0.32
BBC	2474781	97946	1	6789	−0.44	0.65	0.16	0.19

checked for 950 humanly annotated comments in IRC data is 62.49% for subjectivity detection and 70.25% for polarity detection while in the case of BBC data the numbers are, respectively, 73.73% and 80.92% for 594 annotated documents (see [48] and *Materials and Methods* section in [11]).

3. Common Features

The obtained dialogues have been divided into groups of constant dialogue length L. For such data we follow the evolution of mean emotional value $\langle e\rangle_i^L$ and average emotional probabilities $\langle p(e)\rangle_i^L$ ($\langle e\rangle_i^L$. In both cases the $\langle\cdots\rangle_i^L$ symbol indicates taking all dialogues with a specific length L and averaging over all comments with number i, thus, for example, $\langle p(-)\rangle_i^L$ is the probability that at the position i in all dialogues of length L there is a negative statement. The characteristic feature observed regardless of the dialogue length is that the $\langle e\rangle_i^L$ at the end of the dialogue is higher than at the beginning (upper row in Fig. 2). In fact, there is especially

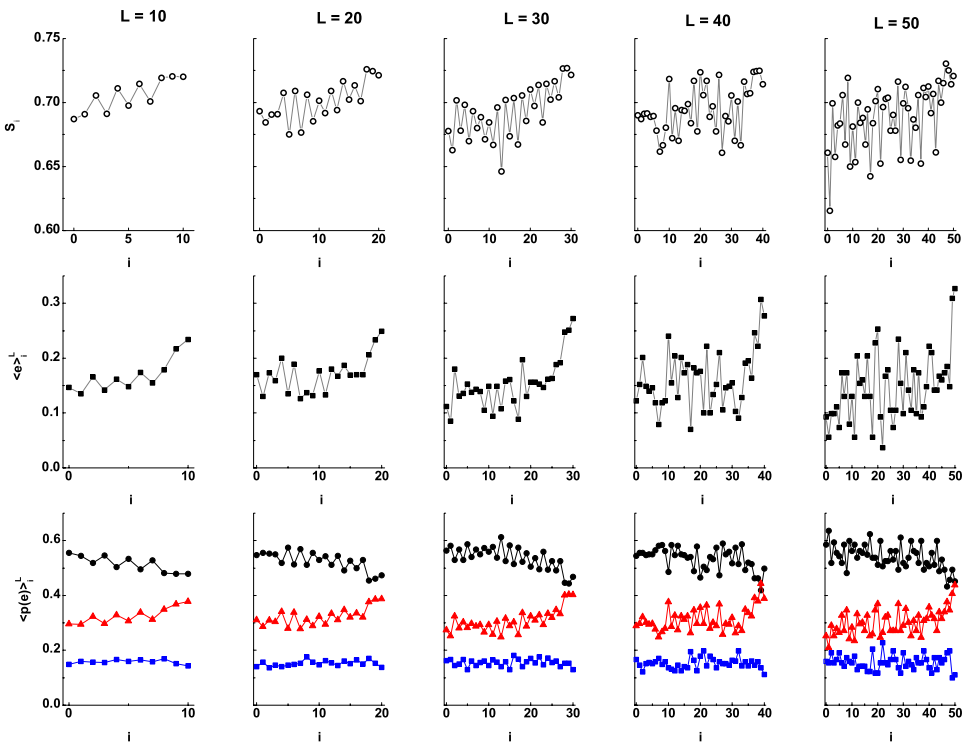


Fig. 2. (Color online) IRC data: entropy S_i of the emotional probabilities distribution $\langle p(e)\rangle_i^L$ (top row), average emotional value $\langle e\rangle_i^L$ (middle row) and average emotional probabilities $\langle p(-)\rangle_i^L$ (squares), $\langle p(0)\rangle_i^L$ (circles), $\langle p(+)\rangle_i^L$ (triangles) in the ith timestep for dialogues of specific $L = 10$ (first column), $L = 20$ (second column), $L = 30$ (third column), $L = 40$ (fourth column) and $L = 50$ (fifth column).

a rapid growth close the very end of the dialogue, which is probably caused by participants who acknowledge others' support issuing comments like "thank you", "you were most helpful", etc.

The direct reason for such behavior is shown in the bottom row of Fig. 2, which presents the evolution of the average emotional probabilities $\langle p(-)\rangle_i^L$, $\langle p(0)\rangle_i^L$ and $\langle p(+)\rangle_i^L$. The observations can be summarized in the following way:

- the negative emotional probability $\langle p(-)\rangle_i^L$ remains almost constant,
- $\langle p(+)\rangle_i^L$ increases and $\langle p(0)\rangle_i^L$ has an opposite tendency,
- $\langle p(+)\rangle_i^L$ and $\langle p(0)\rangle_i^L$ tend to equalize in the vicinity of dialogue end.

The analysis proving that the presented results are of statistical significance of those results is shown in detail in Appendix D.1.

Other manifestation of the system's features can be spotted by examining the level of the entropy S of the emotional probabilities $\langle p(e)\rangle_i^L$. Entropy or other information theoretic quantities as mutual information [14], Kullback–Leiber divergence [34] or Jensen–Shannon divergence [34] have been already used to quantify certain aspects of human mobility [67], semantic resemblance or flow between Wikipedia pages [39, 40] or correlations between consecutive emotional posts [74]. Moreover, basing on entropy, it has also been shown how the coherent structures in the e-mail dialogues arise [17] or how to predict conversation patterns in face-to-face meetings [72]. The concept of entropy is often used in such nonphysical areas as ecology, for example as a tool for tracing the biodiversity [61]. However, as Bailey [1] — the initiator of the social entropy theory (SET) — states, in the case of social sciences the term "entropy" had hardly been used until 1980's, spare the works of Miller [41], Rothstein [53] and Buckley [5] who employed it for the examination of sociological organization structure. In this paper, the entropy is used after Shannon's definition [59], i.e.,

$$S_i^{sh} = -\sum_{e=-1,0,1} \langle p(e)\rangle_i^L \ln\langle p(e)\rangle_i^L. \tag{1}$$

In Fig. 3 we show a schematic plot illustrating the meaning of Eq. (1). If the distribution of some feature is equiprobable (e.g., each of three political parties get exactly 1/3 of the total number of votes, 3(a)), the resulting value of entropy is maximal. In the opposite situation (one party gets the majority of votes, 3(b)), the entropy is very low — in an extreme situation, when all the votes are gathered by one party, the entropy is minimal and equals 0. Thus, entropy can serve as an indicator of the state of the system, showing if it is ordered (low S) or disordered (high S).

Here, taking into account the fact that $\langle p(-)\rangle_i^L$ is constant in the course of dialogue, we paid attention only to $\langle p(+)\rangle_i^L$ and $\langle p(0)\rangle_i^L$, thus the observed entropy had a form of

$$S_i = -[\langle p(0)\rangle_i^L \ln\langle p(0)\rangle_i^L + \langle p(+)\rangle_i^L \ln\langle p(+)\rangle_i^L]. \tag{2}$$

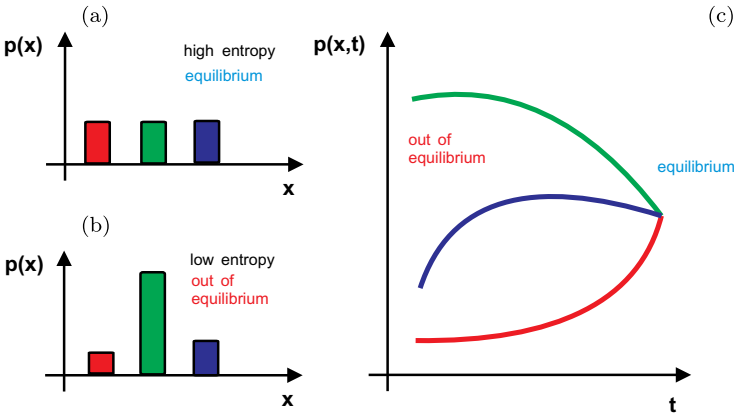


Fig. 3. (Color online) Schematic plot illustrating the meaning of Eq. (1). (a) Equiprobable distribution of some feature, resulting value of entropy is maximal. (b) A dominant feature present in the probability distribution — resulting value of entropy is very low. (c) System is initially out of equilibrium equilibrates in the course of time, acquiring the state of maximal entropy.

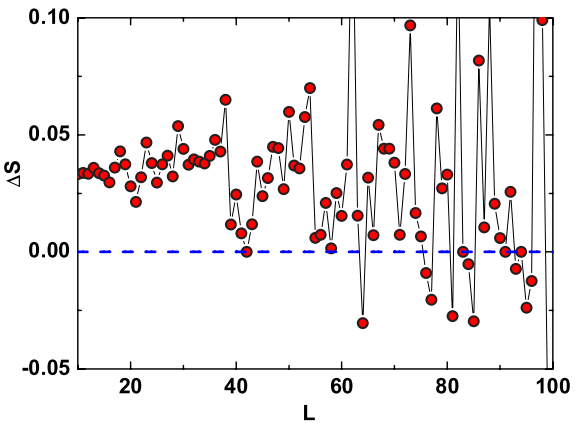


Fig. 4. (Color online) Difference between terminal and initial entropy value ΔS versus the dialogue length L.

Plotting the difference between terminal and initial entropy ΔS versus the length of the dialogue L it is possible to see that for the dialogues up to $L \approx 50$ this difference is always above zero (see Fig. 4). The statistical relevance of this observation is presented in detail in Appendix D.2. It implies a following likely scenario for the dialogue: it evolves in the direction of growing entropy. In the beginning of the dialogue, the probabilities $\langle p(0)\rangle_i^L$ and $\langle p(+)\rangle_i^L$ are separated from each other, contributing to low value of initial entropy S_p. However, then the entropy grows, the probabilities $\langle p(0)\rangle_i^L$ and $\langle p(+)\rangle_i^L$ equalize leading to high value entropy (i.e., higher then the initial one) at the end of the dialogue.

However, it is essential to notice that the observed behavior in the IRC data is only one of the possible scenarios of the more general phenomenon of the principle of maximum entropy [30], governing also certain aspects of biological [76] or social systems [31] (at the level of social networks). The tendency for the isolated system to increase its entropy and to evolve to reach the state characterized by the maximum entropy (MaxEnt) is a well-know physical phenomenon previously observed in many real-world systems [26]. It is a sign of the situation when the system is initially out of equilibrium and in the course of time it equilibrates [Fig. 3(c)], acquiring the state of maximal entropy. Social sciences had incorporated the idea of equilibrium long before entropy [68], although it has then been used rather as a synonym of system integration and stability [49, 50]. In the physical case (and also in this study) it is essential that growing entropy indicates the direction of time. Thus, this behavior should be irrelevant of the type of the system in question. Let us stress that in many settings there are constrains in system's dynamics, e.g., due to interactions with the environment. As a result an equilibrium state is not a state of homogeneous probability distribution since this symmetry can be broken by an external influence. This is observed also in our social dynamics experiment — the fraction of negative comments is constant in time and different from 1/3.

In order to test the assumption on universality of our approach, we performed an analysis analogous to this for the IRC data with respect to emotionally annotated dataset from the BBC Forum (see [11] and [12]) consisting of over 2×10^6 comments and almost 10^5 discussions. In this case each discussion was treated as a natural "dialogue", although it usually consisted of more than 2 users communicating to each other. Following the line of thought presented for IRC data we grouped all discussion of constant length and calculated the quantities $\langle p(-)\rangle_i^L$, $\langle p(0)\rangle_i^L$, $\langle p(+)\rangle_i^L$ and S_i^{sh}. The results, shown in Fig. 5, bear close resemblance to those obtained for IRC data: one can clearly see that while the negative component decreases, the positive and objective (partially) ones increase. In has an instant effect on the value of entropy which grows during the evolution of the discussion (topmost row in Fig. 5). The main difference between IRC and BBC forum results concerns the component whose value decreases during the discussion evolution: for IRC it is the $\langle p(0)\rangle_i^L$ while for BBC forum — $\langle p(-)\rangle_i^L$. It is directly connected to the fact that the above mentioned components play the role of "discussion fuel" [11] propelling thread's evolution. BBC forum data come from such categories as "World News" and "UK News" and as such may lead the discussion participants to place comments of very negative valence. On the other hand `#ubuntu` IRC channel servers rather as a source of professional help which is normally expressed in terms of neutral dialogue. As the discussion lasts, the topic dilutes (BBC forum) or the problem is being solved (IRC) and the dominating component dies out leading to maximization of entropy. Here entropy can serve as a kind of indicator measuring the way emotional states are changing. It can be directly applied as a tool to distinguish between the initial state that is later subject to a sort of thermalization and the final phase where all the emotions get mixed up. Thus, one may regard it

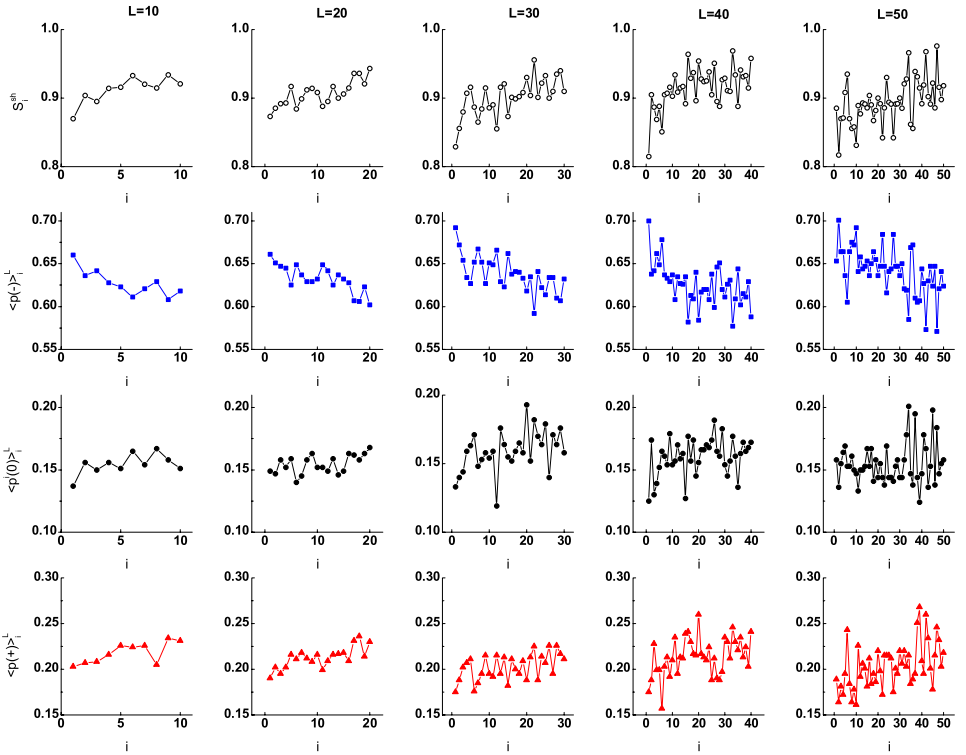


Fig. 5. (Color online) BBC data: entropy S_i^{sh} of the average emotional probabilities distribution $\langle p(e)\rangle_i^L$ (top-most row) and average emotional probabilities $\langle p(-)\rangle_i^L$ (squares) $\langle p(0)\rangle_i^L$ (circles) and $\langle p(+)\rangle_i^L$ (triangles) in the ith timestep discussions of specific length $L = 10$ (first column), $L = 20$ (second column), $L = 30$ (third column), $L = 40$ (fourth column) and $L = 50$ (fifth column).

as an index of the dialogue phase — regardless of the overall emotional character of the medium (i.e., neutral, negative).

There is also another process taking place in the system in question that displays a nontrivial behavior. As shown previously in [11], we can talk about grouping of similarly emotional messages. To quantify the persistence of a specific emotion one can consider the conditional probability $p(e\,|\,ne)$ that after n comments with the same emotional valence the next comment has the same sign. As it is easy to prove, if e would be treated as an identical and independently distributed (i.i.d.) variable the conditional probability $p(e\,|\,ne)$ should be independent of n and equal to $p(e)$, i.e., the probability of a specific emotion in the whole dataset (see Table 1). In the case of the IRC data, the analysis shows (see Fig. 6) that $p(e\,|\,ne)$ is well approximated by

$$p(e\,|\,ne) = p(e\,|\,e)n^{\alpha}, \tag{3}$$

where $p(e\,|\,e)$ is the conditional probability that two consecutive messages have the same emotion (see Appendix D.3 for discussion). The discrepancy between the data

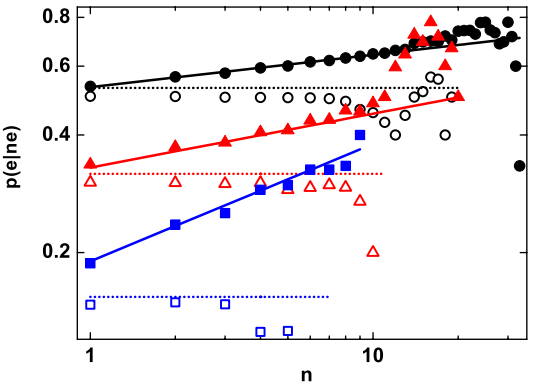


Fig. 6. (Color online) Conditional probability $p(e\,|\,ne)$ of consecutive emotional post of the same sign versus the size n. Full triangles, squares and circles are data points (respectively: negative, neutral and positive messages), empty symbols are shuffled data, solid lines come from Eq. (3) and dotted lines represent relation $p(e\,|\,ne) = p(e)$.

Table 2. Conditional probabilities $p(e\,|\,e)$ and scaling exponents for the power-law cluster growth α_e with errors.

Emotion sign	$p(e\,\|\,e)$	α_e
Positive ($e = 1$)	0.34	0.138 ± 0.004
Neutral ($e = 0$)	0.53	0.083 ± 0.001
Negative ($e = -1$)	0.19	0.30 ± 0.01

and the relation obtained by random insertion of emotional comments (see open symbols in Fig. 6) is significant. The exponents α and the conditional probabilities $p(e\,|\,e)$ are gathered in Table 2.

4. Simulation Description

The methodology described above proves to be successful in finding the prominent characteristic of the data in question, however it is rather useless if one would like to perform the simulations of the dialogues. It is crucial to choose other way for calculating the average emotional probabilities "on the fly" and, using the results, decide on the further dialogue evolution. Thus, we decided to work with moving time window, i.e., the probability of the specific valences in the ith timestep are

$$\begin{cases} \bar{p}_i^M(+) = \dfrac{1}{M}\displaystyle\sum_{j=1}^{j=M} \delta_{e(i-j),+1}, \\ \bar{p}_i^M(0) = \dfrac{1}{M}\displaystyle\sum_{j=1}^{j=M} \delta_{e(i-j),0}, \\ \bar{p}_i^M(-) = \dfrac{1}{M}\displaystyle\sum_{j=1}^{j=M} \delta_{e(i-j),-1}, \end{cases} \tag{4}$$

for $i \geq M$, where δ is the Kronecker delta symbol and M is the size of the window. Consequently, entropy S_i is also calculated using the probabilities $\bar{p}_i^M(+)$ and $\bar{p}_i^M(0)$ as

$$\bar{S}_i^M = -[\bar{p}_i^M(0)\ln\bar{p}_i^M(0) + \bar{p}_i^M(+)\ln\bar{p}_i^M(+)] \tag{5}$$

expressing in fact the entropy in the ith time window. The practical way of application is shown in Fig. 7 for a dialogue of $L = 30$ comments. In this case the size of the time window is set to $M = 10$.

The data-driven facts presented in the previous section lie at the basis of the simulation of dialogues in IRC channels data. The key point treated as an input parameter for this model is the observation of the preferential attraction of consecutive emotional messages. This idea "runs" the dialogue, whereas the discussion is terminated once the difference between the entropy in the given moment and its initial value exceeds certain threshold. Those features are implemented in the following algorithm:

(i) start the dialogue by drawing the first emotional comment with probability $p(e)$,
(ii) set the next comment to have emotional valence e of the previous comment with probability $p(e\,|\,ne) = p(e\,|\,e)n^{\alpha_e}$
(iii) if the drawn probability is higher than $p(e\,|\,ne)$, set the next comment one of two other emotional values (i.e., if the original $e = 1$, then the next

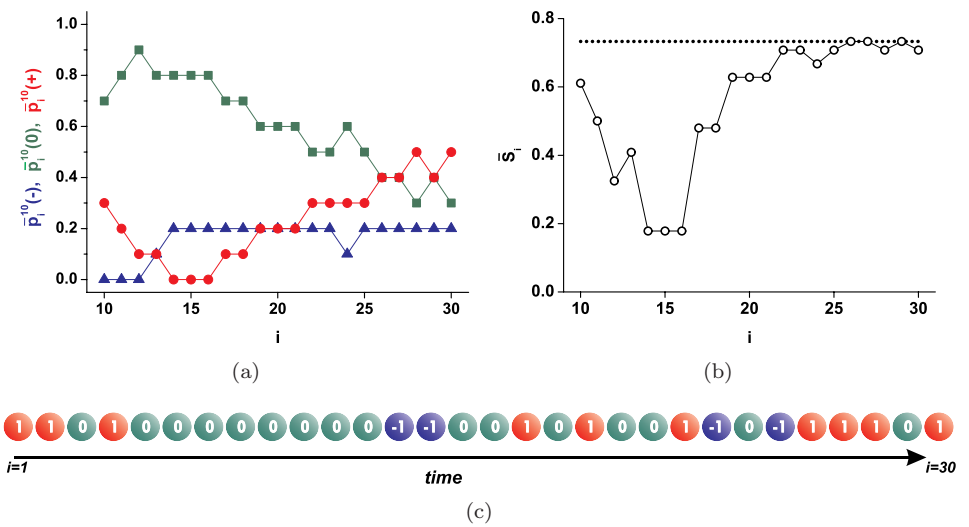


Fig. 7. (Color online) (a) Probabilities of specific valence $\bar{p}_i^M(-)$ (triangles), $\bar{p}_i^M(0)$ (squares) and $\bar{p}_i^M(+)$ (circles) in the ith time window given by Eq. (4) for the exemplary dialogue shown in panel (c). (b) Entropy $\bar{S}_i$ in the ith time window defined by Eq. (5) for the exemplary dialog shown in panel (c). The dotted line marks the maximal value of entropy in Eq. (5) i.e., $\bar{S}_i^{\max} = \frac{2}{5}\ln\frac{5}{2} \approx 0.73$. The dialogue is real-world example from IRC data.

comment valence is 0 with probability $p(0)/[p(0)+p(-)]$ or -1 with probability $p(-)/[p(0)+p(-)]$)

(iv) if the difference between entropy in this time-step and the initial entropy is higher than threshold level ΔS terminate the simulation, otherwise go to point (ii).

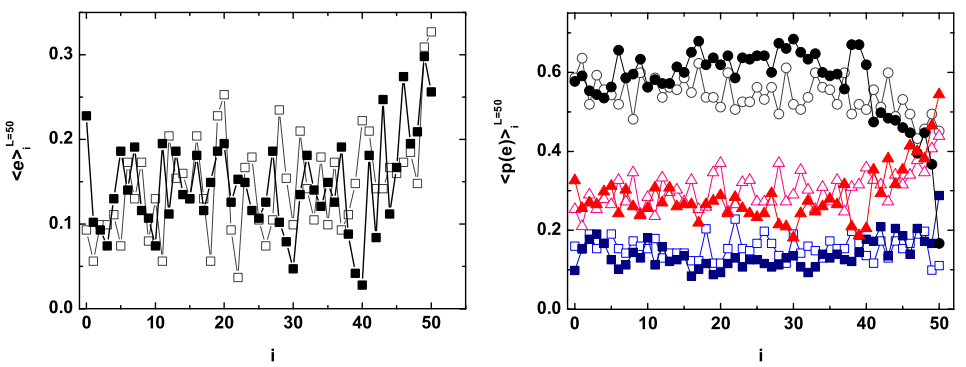
The observed valence probabilities in this simulation are always calculated using quantities in a moving time window given by Eqs. (4) and (5) with $M = 10$.

There is another crucial parameter connected to the simulation process, i.e., the initial entropy threshold S_T. When time-step $i = M$ is reached, the entropy $\bar{S}_i^M$ is calculated for the first time and then decision is taken: if $\bar{S}_M^M < S_T$ the simulation runs further, otherwise it is canceled and repeated. The total number of successfully simulated dialogues is equal to this observed in the real data.

5. Simulation Results

Figure 8 shows a comparison of the average emotional value $\langle e\rangle_i^L$ and average emotional probabilities $\langle p(e)\rangle_i^L$ for the real data and simulations performed according to the algorithm described in the previous section for dialogues of length $L = 50$. As one can see that the plots bear close resemblance apart from only one detail, i.e., the rising value for the $\langle p(-)\rangle_i^L$ close to the end of the dialogue.

Moreover, the simulation strongly depends on the exact value of the initial entropy threshold S_T which can be clearly seen in Fig. 9(a), where the dialogue length distribution is presented. If the S_T is restricted to values between 0.1–0.5 (downward and upward triangles) the distribution of dialogue lengths is exponential and does not follow the one observed in the real data (circles). Higher values of S_T



(a) (b)

Fig. 8. (Color online) Comparison of average emotional value $\langle e\rangle$ (panel a) and probability of specific emotion (panel b, $\langle p(-)\rangle_i^{L=50}$ — squares, $\langle p(0)\rangle_i^{L=50}$ — circles, $\langle p(+)\rangle_i^{L=50}$ — triangles) for simulations performed according to the procedure presented in Sec. 4 (full symbols) and for real data (empty symbols) for dialogue length $L = 50$. The real data shown are identical with those shown in the fifth column of Fig. 2.

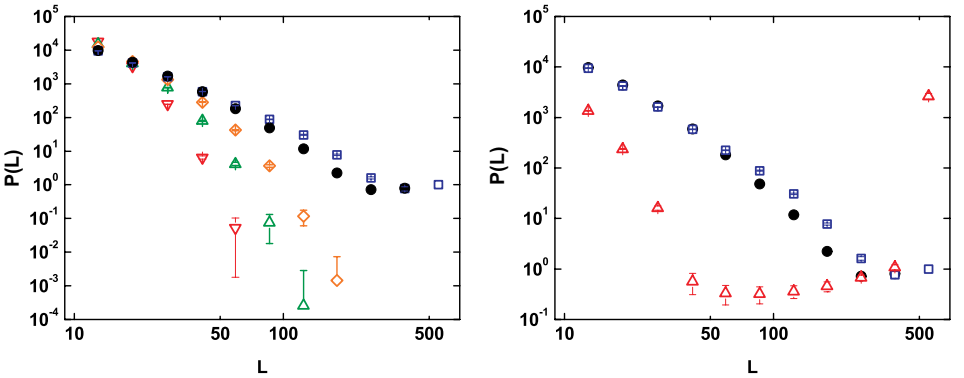


Fig. 9. (Color online) (a) Dialogue length distribution $P(L)$ for real data (circles) and simulations for different values of the initial entropy threshold S_T parameter: $S_T = 0.1$ (downward triangles), $S_T = 0.5$ (upward triangles), $S_T = 0.6$ (diamonds) and $S_T = 0.63$ (squares). (b) Dialogue length distribution $P(L)$ for: real data (circles), simulations with $S_T = 0.63$ (squares) and simulations with $S_T = 0.63$ and insertion of the additional neutral comments (triangles). Each simulation data point is an average over 100 realizations, error bars correspond to standard deviations. Data are logarithmically binned with the power of 1.45.

($S_T = 0.6$, diamonds) shift the curve closer to the data points, nevertheless the character is still exponential. Its only after tuning the S_T parameter to 0.63 that the results obtained from the simulations (squares) are qualitatively comparable with the real data. Full quantitative analysis of the way the parameter S_T was chosen is included in Appendix D.4.

6. Application

It is possible to consider a direct application of the above described model for changing the "trajectory" of the dialogue. For example let us assume that a dialogue system [2, 63, 75] is included as part of the conversation and that its task is to prolong the discussion. In such situation, the system that could rely on the above presented properties would attempt to detect any signs indicating that the dialogue might come to an end and react against it. According to observations presented in Sec. 3 a marker for such event should be the growth of the entropy. In other words the dialogue system should prevent an increase of the entropy in the consecutive time-steps.

In the described case, such action would be an equivalent to an insertion of an objective comment. In this way, an equalization between $\bar{p}_i^M(+)$ and $\bar{p}_i^M(0)$ is prevented and dialogue can last further. An implementation of this rule is presented in Fig. 9(b), where one can compare the real data (again empty circles), a simulation including the entropy-growth rule (again full circles) and a simulation following the insertion of objective comments (empty triangles). While there is a drop-down in the numbers for the small dialogue lengths, the vast majority of the dialogues has

the maximal length (a point in the top-right corner). In this way the insertion of the objective comments is in line with the expected idea of dialogue prolonging.

It is essential to stress that this kind of a theoretical application could be presumably useful and suited only in certain situations and only for particular interactive environments. The key feature observed in the IRC channels data, i.e., the equalization of the emotional probabilities and entropy growth during the time of dialogue does not need to be present in other situations (see e.g., [11, 12]).

On the other hand one could argue that prolonging the discussion on an IRC channel that serves for resolving problems is of little use. We would like to stress that this analysis aims at showing the outline of a more general problem. In fact, this idea could be applied to such media as BBC Forums as well as have a therapeutic usage. To some extent, introduction of deliberately biased emotional comments and scenarios in a human-bot discussion has already taken place [65] resulting in congruent responses issued by participants.

7. Conclusion

Analysis performed on the emotionally annotated dialogues extracted from IRC data demonstrate that following such simple metrics as probability of specific emotion can be useful to predict the future evolution of the discussion. Moreover, all the analyzed dialogues share the same property, i.e., the tendency to evolve in the direction of a growing entropy. Those features, combined together with the observations regarding the preferential growth of clusters, are sufficient to reproduce the real data by a rather straightforward simulation model. In the paper, we also proposed a procedure to directly apply the observed rules in order to modify the way the dialogue evolves. It appears, for example that insertion of objective comments prolongs the discussion by lowering the entropy value. Those observations may be helpful for designing the next generation of interactive software tools [23, 64, 65] intended to support e-communities by measuring various features of their interactions patterns, including their emotional state at the individual, group and collective levels.

Acknowledgments

This work was supported by a European Union grant by the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT. It is part of the CyberEmotions (Collective Emotions in Cyberspace) project (contract 231323). J.S. and J.A.H. acknowledge support from Polish Ministry of Science Grant 1029/7.PR UE/2009/7.

Appendix A. Dialogue Extraction Method

In total, we used 994 daily files with 4600 to 18000 utterances that share a format presented in the first column from the left in Table 3: *post_number* [*timestamp*] ⟨*user_id*⟩ *sentiment_class* with the *sentimentclass* $e = \{-1; 0, 1\}$

Table 3. The process of dialogue extraction in the IRC channel data. Columns from the left show consecutive steps of the algorithm: first and second show the raw data, third is data after application of the searching procedure, fourth is data after averaging multiple posts from the same user and fifth column gives the final output. $[hh\,{:}\,mm]$ defines the timestamp in hours (hh) and min (mm), $\langle user\_id \rangle$ gives the id of the user that addresses the post, $\langle adressing\_user\_id \rangle \rightarrow \langle addresed\_user\_id \rangle$ gives the ids of both addressing and addressed users and value $\{-1, 0, 1\}$ shows the valence of the post.

Original data	User-to-user info	Output 1	Output 2	Final output
1 [00 : 03] ⟨20422⟩ 1	[00 : 03] ⟨20442⟩			Dialogue 1
2 [00 : 04] ⟨55⟩ 1	[00 : 04] ⟨55⟩ → ⟨20442⟩	⟨55⟩ → ⟨20442⟩ 1	⟨55⟩ → ⟨20422⟩ 1	⟨55⟩ ↔ ⟨20422⟩
3 [00 : 05] ⟨20422⟩ 0	[00 : 05] ⟨20442⟩ → ⟨55⟩	⟨20442⟩ → ⟨55⟩ 0	⟨20422⟩ → ⟨55⟩ 0	1
4 [00 : 05] ⟨55⟩ −1	[00 : 05] ⟨55⟩ → ⟨20442⟩	⟨55⟩ → ⟨20442⟩ −1	⟨55⟩ → ⟨20422⟩ −1	0
5 [00 : 08] ⟨20422⟩ 1	[00 : 08] ⟨20422⟩ → ⟨55⟩	⟨20442⟩ → ⟨55⟩ 1	⟨20422⟩ → ⟨55⟩ 1	−1
6 [00 : 08] ⟨55⟩ 0	[00 : 08] ⟨55⟩ → ⟨20442⟩	⟨55⟩ → ⟨20442⟩ 0	⟨55⟩ → ⟨20442⟩ 0	1
7 [00 : 09] ⟨27⟩ 0	[00 : 09] ⟨27⟩ → ⟨20442⟩	⟨27⟩ → ⟨20442⟩ 0	⟨27⟩ → ⟨20442⟩ 0	0
8 [00 : 13] ⟨20422⟩ 0	[00 : 13] ⟨20422⟩	⟨20442⟩ → ⟨27⟩ 0	⟨20422⟩ → ⟨27⟩ 0	Dialogue 2
9 [00 : 13] ⟨2⟩ −1	[00 : 13] ⟨2⟩			⟨20422⟩ ↔ ⟨27⟩
10 [00 : 14] ⟨20422⟩ −1	[00 : 14] ⟨20422⟩ → ⟨20442⟩	⟨20442⟩ → ⟨27⟩ −1		0
11 [00 : 14] ⟨20422⟩ 0	[00 : 14] ⟨20422⟩	⟨20442⟩ → ⟨27⟩ 0		0
12 [00 : 59] ⟨171⟩ −1	[00 : 59] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ −1	⟨171⟩ → ⟨13692⟩ 0	Dialogue 3

Table 3. (*Continued*)

Original data	User-to-user info	Output 1	Output 2	Final output
13 [00 : 59] ⟨171⟩ 1	[00 : 59] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ 1		⟨171⟩ ↔ ⟨13692⟩
14 [00 : 59] ⟨171⟩ 0	[00 : 59] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ 0		0
15 [01 : 00] ⟨171⟩ 1	[01 : 00] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ 1		0
16 [01 : 00] ⟨13692⟩ 0	[01 : 00] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ 0	⟨13692⟩ → ⟨171⟩ 0	1
17 [01 : 01] ⟨171⟩ 1	[01 : 01] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ 1	⟨171⟩ → ⟨13692⟩ 1	1
18 [01 : 01] ⟨171⟩ 1	[01 : 01] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ 1		1
19 [01 : 01] ⟨13692⟩ 1	[01 : 01] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ 1	⟨13692⟩ → ⟨171⟩ 1	1
20 [01 : 01] ⟨171⟩ 1	[01 : 01] ⟨171⟩	⟨171⟩ → ⟨13692⟩ 1	⟨171⟩ → ⟨13692⟩ 1	−1
21 [01 : 02] ⟨171⟩ 1	[01 : 02] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ 1		1
22 [01 : 02] ⟨171⟩ 1	[01 : 02] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ 1		−1
23 [01 : 02] ⟨13692⟩ 1	[01 : 02] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ 1	⟨13692⟩ → ⟨171⟩ 1	1
24 [01 : 02] ⟨13692⟩ 0	[01 : 02] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ 0		
25 [01 : 02] ⟨171⟩ −1	[01 : 02] ⟨171⟩ → ⟨13692⟩	⟨171⟩ → ⟨13692⟩ −1	⟨171⟩ → ⟨13692⟩ −1	
26 [01 : 03] ⟨13692⟩ 1	[01 : 03] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ 1	⟨13692⟩ → ⟨171⟩ 1	
27 [01 : 03] ⟨13692⟩ −1	[01 : 03] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ −1		
28 [01 : 03] ⟨13692⟩ 1	[01 : 03] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ 1		
29 [01 : 03] ⟨171⟩ −1	[01 : 03] ⟨171⟩	⟨171⟩ → ⟨13692⟩ −1	⟨171⟩ → ⟨13692⟩ −1	
30 [01 : 03] ⟨13692⟩ 1	[01 : 03] ⟨13692⟩	⟨13692⟩ → ⟨171⟩ 1	⟨13692⟩ → ⟨171⟩ 1	

used as marker for the emotional valence through this study. Moreover, we could also use information that specifies which user communicates, i.e., directly addresses, another user (see second column in Table 3, shown as $\langle addressing\_user\_id \rangle \rightarrow \langle addressed\_user\_id \rangle$). The discovery of the direct communication links between two users in the IRC channel was based on the discovery of another userID at the beginning of an utterance, followed by a comma or semicolon signs; a scheme commonly used in various multiple users communication channels. However, one has to bear in mind that this kind of information can be sometimes incomplete, i.e., in many cases users do not explicitly specify the receiver of his/her post. Another issue that arises is that the data consist of several overlapping dialogues held simultaneously on one channel. It is also sometimes difficult to indicate the receiver of the message as only part of them are annotated with a user ID they are dedicated to. We created an algorithm that addresses this issue. It consists of two different approaches:

(a) if user A addresses user B in some moment in time and later A writes consecutive messages without addressing anybody specific we assume that he/she is still having a conversation with B
(b) if user A addresses user B and then B writes a message without addressing anybody specific we assume that he/she is answering to A.

The main parameter of such algorithm is the time t in which the searching is being done; in our study we use $t = 5$ min as the threshold value. An exemplary output from the algorithm is shown in the third column in Table 3. In this way we are able to extract a set of dialogues from each of the daily files. After processing the file according to above described rules another issue emerges: it often happens that a user gives a set of consecutive messages directed to one receiver (e.g., the 8th, 10th and 11th line in the third column in Table 3). To create a standardize version of the dialogue (A to B, B to A, A to B and so on), we decided to accumulate the consecutive emotional messages of the same user, calculate the average value $\bar{e}$ in such series and then transform it back into a three-state value according to the formula

$$\begin{cases} e^i = -1 & \bar{e} \in \left[-1; -\dfrac{1}{3}\right] \\ e^i = 0 & \bar{e} \in \left(-\dfrac{1}{3}; \dfrac{1}{3}\right) \\ e^i = 1 & \bar{e} \in \left[\dfrac{1}{3}; 1\right] \end{cases} \tag{A.1}$$

The choice of the transformation form is selected in such a way that a continuous range $[-1; 1]$ is separated into an equal-range division in order to recover the original set of values $\{-1, 0, 1\}$. In effect we obtain the set shown in the fourth column in Table 3. One could also use other ways to transform consecutive emotional messages into one value — we have also tried taking only the last valence, however it did not have any impact on the further analysis and results. The final step of the

data preparation is to divide it into separate dialogues as shown in the 5th column in Table 3. In total, the algorithm produces $N = 93329$ dialogues with the length between $L = 11$ and $L = 339$ (all the dialogues with $L \leq 10$ were omitted).

Appendix B. Numerical Values of Valence

The set of the values $e = \{-1, 0, 1\}$ attached to the concepts of negative, neutral and positive valence may seem to be chosen arbitrary, especially as it leads to a following definition of the average emotional value

$$\langle e \rangle = -1 \times p(-) + 0 \times p(0) + 1 \times p(+) = p(+) - p(-) \tag{B.1}$$

which does not include the value of $p(0)$. However, let us note that taking any set of values $e = \{\lambda - \delta, \lambda, \lambda + \delta\}$ (λ and δ are real numbers) and applying the condition $p(-) + p(0) + p(+) = 1$ gives in effect

$$\langle e \rangle = \delta[p(+) - p(-)]. \tag{B.2}$$

Thus, any linear combination of λ and δ results in definition of $\langle e \rangle$ proportional to the one shown in Eq. (B.1).

Appendix C. Emotional Classifier Quality

In order to check the ability of the classifier to recognize the correct emotion one uses a set of classified messages and annotate them manually. A typical way to quantify classification quality is to use the accuracy measure defined as

$$a = \frac{TP + TN}{TP + FP + FN + TN}, \tag{C.1}$$

where TP is the number of comments that were correctly classified as being in the class c (true positives), FP stands for the number of comments falsely classified to class c (false positives), FN denotes the number of comments that were incorrectly classified as not belonging to c (false negatives) and finally TN is the number of comments correctly classified as not being in the class c (true negatives) [32].

In case of the sentiment analysis it is common [48] to use following class distinctions: objective (i.e., neutral) versus subjective (i.e., positive or negative) and positive versus negative resulting in two values of a: $a_{\mathrm{obj/sub}}$ (accuracy for subjectivity detection) and $a_{\mathrm{pos/neg}}$ (accuracy for polarity detection).

Appendix D. Error Analysis

D.1. *Emotional probabilities*

To support the concepts of the equalization of probabilities a proper error analysis should be performed with respect to the data shown in Figs. 2 and 4. However, as it is impossible to obtain directly the error values of the mentioned quantities we decided to use the definitions shown in Eq. (4) that, due its structure (i.e.,

calculation in a moving window), enables us to obtain the standard deviations $\sigma_L^M(-)$, $\sigma_L^M(0)$, $\sigma_L^M(+)$ of each of the quantities $\bar{p}_L^M(-)$, $\bar{p}_L^M(0)$, $\bar{p}_L^M(+)$. For further simplicity in notation we omit superscript, assuming that in all cases the moving window has the size $M = 10$. In effect one can determine the differences $\Delta\bar{p}_L(e) = \bar{p}_L(e) - \bar{p}_{L-10}(e)$ ($e = -1, 0$ or 1) that show the change of probability value between the end and the start of the dialogue. Its errors that are given on the other hand by

$$\sigma_{\Delta\bar{p}_L(e)} = \sqrt{\sigma_L^2(e) + \sigma_{L-10}^2(e)}. \tag{D.1}$$

The plot of $\Delta\bar{p}_L(e)$ with error bars versus the length of the discussion L shown in Fig. 10 implicate that up to $L \approx 50$ following relations hold true $\Delta\bar{p}_L(+) > 0$, $\Delta\bar{p}_L(0) < 0$ and $\Delta\bar{p}_L(-) = 0$. After crossing $L = 50$ the values start to fluctuate heavily, nonetheless their average value calculated in the moving window of last 10 values confirm trend tendency, thus supporting paper's key arguments.

D.2. *Entropy*

One can make use of the method presented in the previous section to test the validity of the assumption of entropy growth proposed in Sec. 3. In fact, using the definition (5) one can express entropy difference as $\Delta\bar{S}_L = \bar{S}_L - \bar{S}_{L-10}$ and its errors as

$$\sigma_{\Delta\bar{S}_L} = \sqrt{\sum_{e=0,1} [s_L^2(e) + s_{L-10}^2(e)]}, \tag{D.2}$$

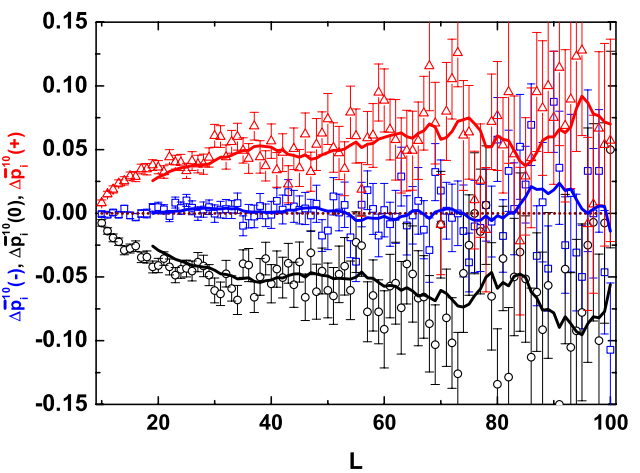


Fig. 10. (Color online) Differences of emotional probabilities $\Delta\bar{p}_L(-)$ (squares), $\Delta\bar{p}_L(0)$ (circles), $\Delta\bar{p}_L(+)$ (triangles) with corresponding error bars given by Eq. (D.1) versus the length of the dialogue L. Solid lines indicate moving average taken over the last 10 values. Dashed line marks $\Delta p = 0$.

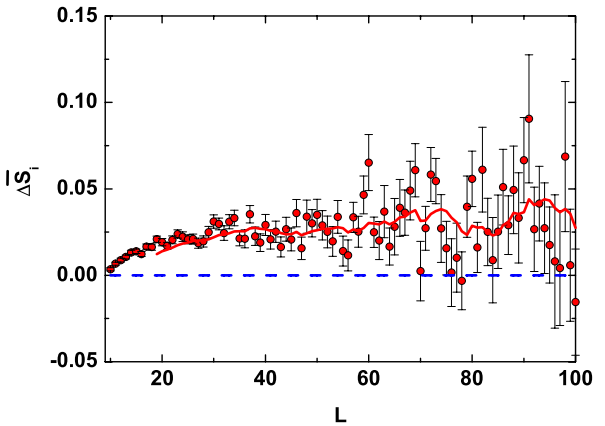


Fig. 11. (Color online) Entropy difference $\Delta\bar{S}_L = \bar{S}_L - \bar{S}_{L-10}$ with corresponding error bars given by Eq. (D.2) versus dialogue length L. Solid line indicate moving average taken over the lat 10 values. Dashed line marks $\Delta\bar{S}_L = 0$.

where

$$s_L(e) = [1 + \ln \bar{p}_L(e)]\sigma_L(e), \tag{D.3}$$

$$s_{L-10}(e) = [1 + \ln \bar{p}_{L-10}(e)]\sigma_{L-10}(e). \tag{D.4}$$

The results are shown in Fig. 11. The concept of entropy growth (i.e., $\Delta\bar{S}_L > 0$) is fulfilled strictly up to $L \approx 70$. After crossing that point small statistics of data leads to large fluctuations, nonetheless moving average calculated for the last 10 points (solid lines) remains above zero.

D.3. *Conditional probability*

The concept of conditional probabilities $p(e\,|\,ne)$ following a power-law relation comes from the paper by Chmiel *et al.* [11], where it has been shown that this process could be responsible for a specific shape of the probability distribution of emotional cluster lengths observed in that data from blogs, *Digg.com* portal and BBC Forum. The data in the mentioned study are of the similar structure as in the IRC case (i.e., chains with values $e = \{-1, 0, 1\}$ representing valence of comments). However in this study the range of the data on both axis (see Fig. 6) is very narrow thus it is essential to check other possibilities of fitting functions. Due to the large fluctuations of data caused by underrepresentation of large clusters (e.g., there are only few positive clusters with $n > 10$) we limited our analysis to the range $n \in [1; 10]$. We decided to check the following linear (p_{LIN}), exponential (p_{EXP}) and power-law (p_{POW}) test functions:

$$p_{\mathrm{LIN}}(e\,|\,ne) = \alpha(n-1) + p(e\,|\,e), \tag{D.5}$$

$$p_{\mathrm{EXP}}(e\,|\,ne) = p(e\,|\,e)e^{\alpha(n-1)}, \tag{D.6}$$

$$p_{\mathrm{POW}}(e\,|\,ne) = p(e\,|\,e)n^{\alpha}. \tag{D.7}$$

Table 4. Values of R^2 coefficient for specific fitting functions.

	Negative	Neutral	Positive
Power-law	0.94	0.99	0.96
Linear	0.91	0.84	0.94
Exponential	0.85	0.80	0.90

The form of these functions is chosen in such a way that for $n = 1$ they recover the value of $p(e \mid e)$, thus they have only one free parameter i.e., α. The obtained values of R^2 coefficient which show the correspondence between fittings and real data are shown in Table 4.

As an outcome of this fitting procedure we use the power-law function p_{POW} characterized by exponents gathered in Table 2.

D.4. *Dialogue length probability*

Here we give the rationale for choosing the specific value of S_T used in the simulations. The key idea behind introducing S_T is that it prevents drawing specific values of $p(+)$ and $p(0)$ (i.e., values close to equilibrium) that would lead to a premature or even instant convergence of the simulation. Moreover, it enables tuning the shape of dialogue length histogram $P(L)$. We check the validity of the chosen parameter by performing the Kolmogorov–Smirnov goodness-of-fit test [35]. We proceed as follows: first we bin the data and form the assumed cumulative distribution function (CDF) F_0. Then, for each considered value S_T we perform 100 separate simulations, in each case we bin it and construct an empirical CDF F. Finally, for each separate

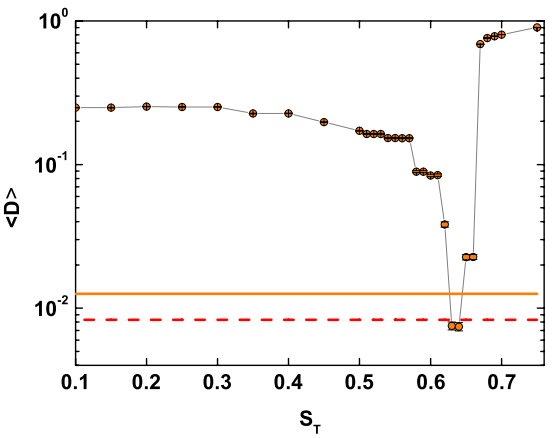


Fig. 12. (Color online) Average Kolmogorov–Smirnov statistic $\langle D \rangle$ (circles) with its standard deviation σ_D (error bars) versus initial entropy threshold parameter S_T. Lines mark acceptance limits for significance levels $\alpha = 0.05$ (solid line) and $\alpha = 0.20$ (dashed line). Solid line linking the data points is for eye guidance only.

simulation we calculate the Kolmogorov–Smirnov statistic defined as

$$D = \sup_L |F(L) - F_0(L)| , \tag{D.8}$$

where L runs over all bins. The obtained values D are then used to calculate the average value $\langle D \rangle$ and its standard deviation σ_D. Those values are plotted against S_T in Fig. 12 together with two lines marking acceptance limits for significance levels $\alpha = 0.05$ (solid line) and $\alpha = 0.20$ (dashed line) evaluated using equations $1.36/\sqrt{\tilde{N}}$ and $1.07/\sqrt{\tilde{N}}$, respectively [35], where $\tilde{N} = 16628$ stands for number of elements in the binned data. This analysis reveals that only in case of $S_T = 0.63$ and $S_T = 0.64$ the hypothesis that the dialogue length distribution obtained via simulation comes from the original distribution F_0 can be accepted.

References

[1] Bailey, K. D., Sociological entropy theory: Toward a statistical and verbal congurence, *Qual. Quantity* **18** (1983) 113–133.

[2] Bohus, D. and Rudnicky, A., Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda, in *Proc. Eurospech 2003* (2003), pp. 597–600.

[3] Buber, M., *Between Man and Man* (Routledge & Kegan Paul, London, 1947).

[4] Bucci, W., Pathways of emotional communication, *Psychoanalytic Inquiry* **21** (2001) 40–70.

[5] Buckley, W., *Sociology and Modern Systems Theory* (Prentice-Hall, Engelwood Cliffs, 1967).

[6] Caffi, C. and Janney, R. W., Toward a pragmatics of emotive communication, *J. Pragmat.* **22** (1994) 325–373.

[7] Cattuto, C., van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F. and Vesipignani, A., Dynamics of person-to-person interactions from distributed RFID sensor networks, *PLoS ONE* **5** (2010) e11596.

[8] Chmiel, A. and Hołyst, J. A., Flow of emotional messages in artificial social networks, *Int. J. Mod. Phys. C* **21** (2010) 593–602.

[9] Chmiel, A. and Hołyst, J. A., Transition due to preferential cluster growth of collective emotions in online communities, *Phys. Rev. E* **87** (2013) 022808.

[10] Chmiel, A., Kowalska, K. and Hołyst, J. A., Scaling of human behavior during portal browsing, *Phys. Rev. E* **80** (2009) 066122.

[11] Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A. and Hłyst, J. A., Collective emotions online and their influence on community life, *PLoS ONE* **6** (2011) e22207.

[12] Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M. and Hołyst, J. A., Negative emotions boost user activity at BBC forum, *Physica A* **390** (2011) 2936–2944.

[13] Cordell, D. M. and McGahan, J. R., Mutual gaze duration as a function of length of conversation in male-female dyads, *Psychol. Rep.* **94** (2004) 109–114.

[14] Cover, T. M. and Thomas, J. A., *Elements of Information Theory* (Wiley, New York, 1991), pp. 18–26.

[15] Czaplicka, A. and Hołyst, J. A., Modeling of internet influence on group emotion, *Int. J. Mod. Phys. C* **23** (2012) 1250020.

[16] Czaplicka, A., Chmiel, A. and Hołyst, J. A., Emotional agents at the square lattice, *Acta Phys. Pol. A* **117** (2010) 688–694.
[17] Eckmann, J.-P., Moses, E. and Sergi, D., Entropy of dialogues creates coherent structures in e-mail traffic, *Proc. Natl. Acad. Sci. USA* **101** (2004) 14333–14337.
[18] Eviatar, Z. and Zaidel, E., The effects of word-length and emotionality on hemispheric contribution to lexical decision, *Neuropsychologia* **29** (1991) 415–428.
[19] Feldman, L. A., Valence focus and arousal focus: Individual differences in the structure of affective experience, *J. Personality Social Psychol.* **69** (1995) 153–166.
[20] Fussell, S. R. and Moss, M. M., Figurative language in emotional communication, in *Social and Cognitive Approaches to Interpersonal Communication*, eds. Fussell, S. R. and Kreuz, R. J. (Lawrence Erlbaum Associates, Mahwah, 1998).
[21] Garas, A., Garcia, D., Skowron, M. and Schweitzer, F., Emotional persistence in online chatting communities, *Sci. Rep.* **2** (2012) 402.
[22] Garcia, D., Garas, A. and Schweitzer, F., Positive words carry less information than negative words, *EPJ Data Sci.* **1** (2012) 3.
[23] Gobron, S., Ahn, J., Paltoglou, G., Thelwall, M. and Thalmann, D., From sentence to emotion: A real-time three-dimensional graphics metaphor of emotions extracted from text, *Vis. Comput.* **26** (2010) 505–519.
[24] Goh, K.-I., Eom, Y.-H., Jeong, H., Kahng, B. and Kim, D., Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions, *Phys. Rev. E* **73** (2006) 066123.
[25] Grice, H.P., *Logic and Conversation in Syntax and Semantics*, Vol. 3, eds. Cole, P. and Morgan, J. (Academic Press, New York, 1975).
[26] Harte, J., *Maximum Entropy and Ecology* (Oxford University Press, Oxford, 2011).
[27] http://en.wikipedia.org/wiki/Internet_Relay_Chat.
[28] https://help.ubuntu.com/community/InternetRelayChat.
[29] Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., van den Broeck, W., Gesualdo, F., Pandolfi, E., Rava, L., Rizzo, C. and Tozzi, A. E., Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors, *PLoS ONE* **6** (2011) e17144.
[30] Jaynes, E. T., Information theory and statistical mechanics, *Phys. Rev.* **106** (1957) 620–630.
[31] Johnson, S., Torres, J. J., Marro, J. and Muñoz, M. A., Entropic origin of disassortativity in complex networks, *Phys. Rev. Lett.* **104** (2010) 108702.
[32] Kohavi, B. and Provost, F., Glossary of terms, *Mach. Lear.* **30** (1998) 271–274.
[33] Kujawski, B., Hołyst, J. A. and Rodgers, G. J., Growing trees in internet news groups and forums, *Phys. Rev. E* **76** (2007) 036103.
[34] Lin, J., Divergence measures based on the Shannon entropy, *IEEE Trans. Inf. Theory* **37** (1991) 145–151.
[35] Lindgren, B. W., *Statistical Theory*, 3rd edn. (Macmillan, New York, 1976).
[36] Littlejohn, S. W. and Foss, K. A., *Theories of Human Communication* (Waveland Press, Long Grove, 2010).
[37] Macdonald, C., Ounis, I. and Soboroff, I., Overview of the TREC-2007 Blog Track, in *The Sixteenth Text REtrieval Conf. (TREC 2007) Proc.* (Gaithersburg, 2007).
[38] Manning, Ch. D. and Schütze, H., *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, 1999).
[39] Masucci, A. P., Kalampokis, A., Eguíluz, V. M. and Hernández-García, E., Extracting directed information flow networks: An application to genetics and semantics, *Phys. Rev. E* **83** (2011) 026103.

[40] Masucci, A. P., Kalampokis, A., Eguíluz, V. M. and Hernández-García, E., Wikipedia information flow analysis reveals the scale-free architecture of the semantic space, *PLoS ONE* **6** (2011) e17333.
[41] Miller, G. A., What is information measurement? *Am. Psychol.* **8** (1953) 3–11.
[42] Mitrović, M. and Tadić, B., Bloggers behavior and emergent communities in blog space, *Eur. Phys. J. B* **73** (2010) 293–301.
[43] Mitrović, M., Paltoglou, G. and Tadić, B., Networks and emotion-driven user communities at popular blogs, *Eur. Phys. J. B* **77** (2010) 597–609.
[44] Mitrović, M., Paltoglou, G. and Tadić, B., Quantitative analysis of bloggers' collective behavior powered by emotions, *J. Stat. Mech., Theory Exp.* (2011), P02005.
[45] Oliveira, J. G. and Barabási, A.-L., Human dynamics: Darwin and Einstein correspondence patterns, *Nature (London)* **437** (2005) 1251.
[46] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. and Barabási, A.-L., Structure and tie strengths in mobile communication networks, *Proc. Natl. Acad. Sci. USA* **104** (2007) 7332–7336.
[47] Ounis, I. and Macdonald, C., The TREC blogs06 collection: Creating and analysing a blog test collection, Technical Report (Dept of Computing Science, University of Glasgow, 2008).
[48] Paltoglou, G., Gobron, S., Skowron, M., Thelwall, M. and Thalmann, D., Sentiment analysis of informal textual communication in cyberspace, in *Proc. ENGAGE 2010, Lecture Notes in Computer Science, State-of-the-Art Survey* (Springer, Heidelberg, 2010), pp. 13–25.
[49] Pareto, V., *The Mind and Society* (Harcourt, Brace, New York, 1935).
[50] Parsons, T., *The Social System* (Free Press, Glencoe, 1951).
[51] Peng, F., Schuurmans, D. and Wang, S., Language and task independent text categorization with simple language models, in *NAACL '03*, eds. Hovy, E., Hearst, M. and Ostendorf, M. (Association for Computational Linguistics, Edmonton, 2003), pp. 110–117.
[52] Pohorecki, P., Sienkiewicz, J., Mitrović, M., Paltoglou, G. and Hołyst, J. A., Statistical analysis of emotions and opinions at digg website, *Acta Phys. Pol. A* **123** (2013) 604–614.
[53] Rothstein, J., *Communication, Organization, and Science* (Wing Press, Indian Hills, 1958).
[54] Sacks, H., Schegloff, E. A. and Jefferson, G., A simplest systematics for the organisation of turn-taking for conversation, *Language* **50** (1974) 696–735.
[55] Saussure, F., *Course in General Linguistics* (Fontana/Collins, Glasgow, 1977).
[56] Schweitzer, F. and Garcia, D., An agent-based model of collective emotions in online communities, *Eur. Phys. J. B* **77** (2010) 533–545.
[57] Sebastiani, F., Machine learning in automated text categorization, *ACM Comput. Surveys* **34** (1-472002).
[58] Shannon, C. E. and Weaver, W., *The Mathematical Theory of Communication* (The University of Illinois Press, Urbana, 1949).
[59] Shannon, C. E., A Mathematical theory of communication, *Bell Syst. Tech. J.* **27** (1948) 379–423.
[60] Shimanoff, S. B., Commonly named emotions in everyday conversations, *Perceptual Motor Skills* **58** (1984) 514.
[61] Shipley, B., Vile, D. and Garner, É., From plant traits to plant communities: A statistical mechanistic approach to biodiversity, *Science* **314** (2006) 812–814.
[62] Sinatra, R., Condorelli, D. and Latora, V., Networks of motifs from sequences of symbols, *Phys. Rev. Lett.* **105** (2010) 178702.

[63] Skowron, M., Affect listeners: Acquisition of affective states by means of conversational systems, *Lect. Notes Comput. Sci.* **5967** (2010) 169–181.
[64] Skowron, M., Pirker, H., Rank, S., Paltoglou, G. and Gobron, S., No peanuts! Affective cues for the virtual bartender, in *Proc. 24th Int. FLAIRS Conf.*, eds. Murray, R. Ch. and McCarthy, P. M. (AIII Press, Palm Beach, 2011), pp. 117–122.
[65] Skowron, M., Rank, S., Theunis, M. and Sienkiewicz, J., The good, the bad and the neutral: Affective profile in dialog system-user communication, *Lect. Notes Comput. Sci.* **6974** (2011) 337–346.
[66] Sobkowicz, P. and Sobkowicz, A., Dynamics of hate based Internet user networks, *Eur. Phys. J. B* **73** (2010) 633–643.
[67] Song, Ch., Qu, Z., Blumm, N. and Barabási, A.-L., Limits of predictability in human mobility, *Science* **327** (2010) 1018–1021.
[68] Spencer, H., *First Principles* (Appleton, New York, 1864).
[69] Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., van den Broeck, W., Régis, C., Lina, B. and Vanhems, P., High-resolution measurements of face-to-face contact patterns in a primary school, *PLoS ONE* **6** (2011) e23176.
[70] Stivers, T., Enfield, N. J., Brown, P., Englert, Ch., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E. and Levinson, S. C., Universals and cultural variation in turn-taking in conversation, *Proc. Natl. Acad. Sci. USA* **106** (2009) 10587–10592.
[71] Suler, J., The online disinhibition effect, *CyberPsychol. Behav.* **7** (2004) 321–326.
[72] Takaguchi, T., Nakamura, M., Sato, N., Yano, K. and Masuda, N., Predictability of conversation partners, *Phys. Rev. X* **1** (2011) 011008.
[73] van Dijk, T. A., Cognitive context models and discourse, in *Language Structure, Discourse and the Access to Consciousness*, eds. Stamenow, M. (Benjamins, Amsterdam, 1997), pp. 189–226.
[74] Weroński, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M. and Hołyst, J. A., Emotional analysis of blogs and forums data, *Acta Phys. Pol. A* **121** (2012) B-128–B-132.
[75] Williams, J. D., Poupart, P. and Young, S., Partially observable Markov decision processes with continuous observations for dialogue management, in *Proc. 6th SigDial Workshop on Discourse and Dialogue*, eds. Dybkjær, L. and Minker, W. (ACL, East Stroudsburg, 2005), pp. 25–34.
[76] Williams, R. J., Biology, methodology or chance? The degree distributions of bipartite ecological networks, *PLoS ONE* **6** (2011) e17645.
[77] Wu, Y., Zhou, Ch., Xiao, J., Kurths, J. and Schellnhuber, H. J., Evidence for a bimodal distribution in human communication, *Proc. Natl. Acad. Sci. USA* **107** (2010) 18803–18808.
[78] Xenikos, D. G., Modeling human dialogue — The case of group communications in trunked mobile telephony, *Physica A* **388** (2009) 4910–4918.