Physica A 389 (2010) 5832–5837

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Human comment dynamics in on-line social systems

Ye Wu [a,b,e,*], Changsong Zhou [c], Maoying Chen [d], Jinghua Xiao [b], Jürgen Kurths [e,f,g]

[a] School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] Interdisciplinary Center for Dynamics of Complex Systems, University Potsdam, Am Neuen Palais 10, D-14469, Germany
[c] Department of Physics, Center for Nonlinear Studies, and The Beijing-Hong Kong-Singapore Joint Center for Nonlinear and Complex Systems (Hong Kong), Hong Kong Baptist University, Kowloon Tong, Hong Kong, China
[d] Tsinghua National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, PR China
[e] Potsdam Institute for Climate Impact Research (PIK), Telegraphenberg A31, 14473 Potsdam, Germany
[f] Institute of Physics, Humboldt University, 10099 Berlin, Germany
[g] Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, UK

ARTICLE INFO

Article history:
Received 28 April 2010
Received in revised form 27 July 2010
Available online 6 September 2010

Keywords:
Human dynamics
On-line social systems
Power-law distribution

ABSTRACT

Human comment is studied using data from 'tianya' which is one of the most popular on-line social systems in China. We found that the time interval between two consecutive comments on the same topic, called inter-event time, follows a power-law distribution. This result shows that there is no characteristic decay time on a topic. It allows for very long periods without comments that separate bursts of intensive comments. Furthermore, the frequency of a different ID commenting on a topic also follows a power-law distribution. It indicates that there are some "hubs" in the topic who lead the direction of the public opinion. Based on the personal comments habit, a model is introduced to explain these phenomena. The numerical simulations of the model fit well with the empirical results. Our findings are helpful for discovering regular patterns of human behavior in on-line society and the evolution of the public opinion on the virtual as well as real society.

1. Introduction

Recently, understanding the regularity in complex human dynamics has attracted more and more attention in various fields. The classical view has assumed that human activities are homogeneous Poisson processes [1,2]. Such processes have a well-known statistical property: the time interval between two consecutive events, called the inter-event time τ, follows an exponential distribution, $P(\tau) = \lambda e^{-\lambda\tau}$. Recent evidence [3–16] from various deliberate human activity patterns, including email and letter communications, library usage, broker trades, web browsing, etc., have shown that various human activities are non-Poissonian, with bursts of frequent actions separated by long periods of inactivity, leading to power-law heavy tails in the distribution of the inter-event time $P(\tau) = \tau^{-\gamma}$. These findings are very important in areas as diverse as disease spreading, resource allocation and emergency response, etc. Several mechanisms [3–16] proposed to explain the origin of bursts and heavy tails are limited in applications. More evidence about non-Poissonian human dynamics are needed. And the general origin of the heavy tails is still far from being clearly understood.

As an important part of modern life, human behavior on the Internet also attracts more and more research interest. Dezsõ et al. found that the time interval between consecutive visits by the same user to the site http://www.origi.hu follows a power-law distribution [17]. They also showed that the exponent characterizing the individual user's browsing patterns

* Corresponding author at: Interdisciplinary Center for Dynamics of Complex Systems, University Potsdam, Am Neuen Palais 10, D-14469, Germany.
E-mail addresses: wuye@pik-potsdam.de (Y. Wu), cszhou@hkbu.edu.hk (C. Zhou).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.08.049

Table 1
The detailed format of one topic.

User (ID)	Comment time	Topic
ID1	2010/02/12,12:10:01	Topic A
ID2	2010/02/12,12:11:05	Topic A
...	...	
ID3	2010/02/15,10:10:03	Topic A
ID2	2010/02/15,11:02:01	Topic A

determines the power-law decay in a document's visitation. Chmiel et al. [18] investigated flows of visitors migrating between different portal sub-pages. A model of portal surfing is developed where the browsing process corresponds to a self-attracting walk on weighted networks with a short memory. Grabowski [19] found that the distribution of human activity (e.g. the total number of books read or songs played in on-line social systems) has the form of a power-law. Zhou et al. focus on the origin of power-laws in rating of movies. Their results demonstrate a significant role of the activity of individuals on the society-level patterns of human behavior [20]. All these findings indicate that human behavior on the Internet is typically non-Poissonian. It is very interesting and important to further study the scaling about the human dynamics on the Internet.

In this paper, based on the data collected from 'tianya' which is one of the most popular on-line social systems in China, we show that the inter-event time between two consecutive comments on the same topic follows a power-law distribution. Meanwhile, the distribution of the number of comments in the same topic from different users also follows a power-law. This means that there were some "hubs" in the topic who lead the direction of the public opinion. Furthermore, the power-law distribution of the inter-event time shows that there is no characteristic decay time on a topic. A topic may be ignored for a long time, and is revisited and intensively commented again sometimes. To obtain more insights into these observations of the human dynamics in on-line social systems, we propose a model based on the attraction mechanism. Our findings may be helpful to distinguish different types of public opinions in the virtual society in the future.

This paper is organized as follows: In Section 2, the original of the data and detailed information about the data are introduced. The statistic results are presented in Section 3. The model and numerical simulations are presented in Section 4. Finally, our conclusions are given in Section 5.

2. Data description

Our data were obtained from "tianya" (http://www.tianya.cn/), which is one of the most popular on-line social systems in China. Every user is assigned a different identity name (ID). An ID can build a topic, and all the IDs can comment on it. They can discuss different opinions and communicate on the topic. Until 2010/02/11, there were 33,296,350 IDs in "tianya", and there were about 200,000 IDs on average on-line at the same time. The topics and the public opinion in "tianya" reflect part of the public opinions in the real society in China. We randomly sample some topics which were commented more than 3000 times as our dataset. The types of the topics are different, from public news to personal stories, indicating that our results are general for different topics. The format of the data is shown in Table 1.

3. Statistic results

The inter-event time plays an important role in many human collective behaviors. For example, the power-law distribution of the inter-event time about sending two consecutive E-mails can advance the spreading of the computer virus [21]. The non-Poissonian distribution of arriving rate also has an impact on the classical queue theory [22,23]. Here, we focus on the inter-comment time τ on the same topic, i.e., the time interval between two consecutive comments made by any user on the same topic first. And then we also focus on the inter-comment time τ' made by a single user later on. Although the inter-comment time has some relationship with inter-visit time on a website [17], it is obviously different from inter-visit time because comment behavior is only a very little part of the web visiting behavior. Human comment dynamics on the web can reflect how individuals comment with each other while human dynamics of visiting a website cannot. Four different topics were taken as examples. Topic A is about "Various kinds of fanciful and fabulous flowers". Topic B and Topic C are about two social events that happened in China. Due to privacy, the detailed contents about these two topics are not listed here. Topic D is about "War against those addicted to the Internet: Sound for the freedom". The detailed information about the topics is given in Table 2. Here duration is the time interval between the topic creation and data collection. Total clicks, total comments and total number of IDs are counted during the whole duration.

The distribution of the inter-comment time is shown in Fig. 1. It is clearly seen that all the distributions are power-law, although the topics differ by content and popularity. The exponent varies for the different topics. These results show that the human comment process is non-Poissonian as the human dynamics of letter and E-mail communication, web browsing, on-line movie watching and broker trades. The heavy tail of the distribution allows for long periods of inactivity that separate bursts of intensive activity. Here, it also means there is no characteristic decay time in the human comment dynamics. A topic can be revisited, reactivated and commented upon frequently even after a very long time. For example, topic A was

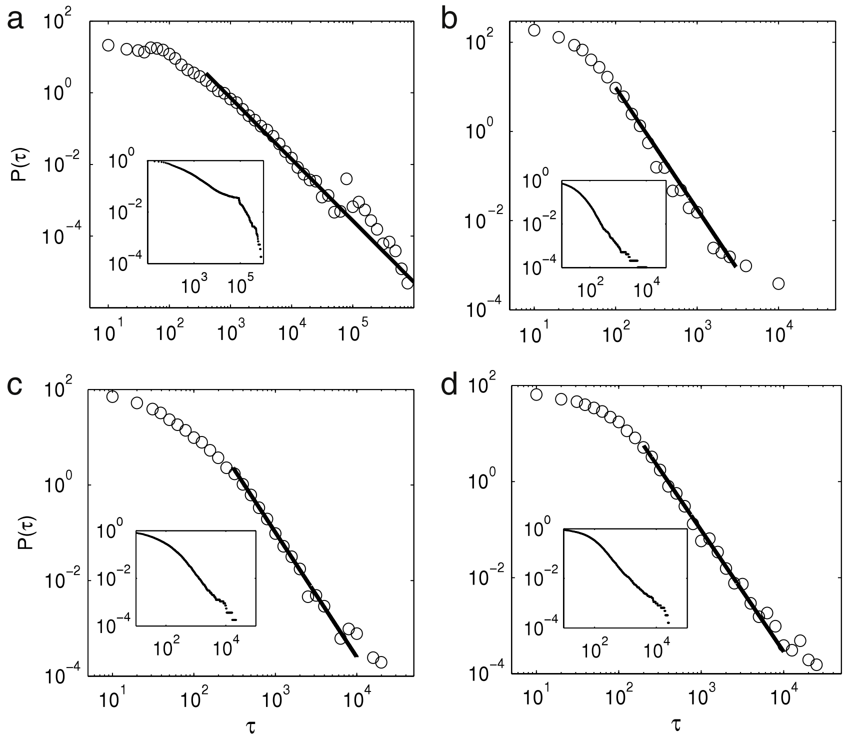

Fig. 1. The inter-comment time distribution of four different topics. The black line is the fitting function. In the inset we plot the cumulative probability distribution. (a) Topic A, slope $\gamma = 1.32 \pm 0.06$, (b) Topic B, slope $\gamma = 2.83 \pm 0.30$, (c) Topic C, slope $\gamma = 2.67 \pm 0.09$, (d) Topic D, slope $\gamma = 2.65 \pm 0.14$.

Table 2
The detailed information about four randomly selected topics.

	A	B	C	D
Total clicks	792,429	618,885	223,961	524,512
Total comments	5,549	9,822	5,757	7,186
Total numbers of IDs	3,965	1,760	3,959	2,663
Duration(s)	24,989,467	464,957	829,408	1,082,256

created 8 months ago and it is still often commented upon now. A large population would read the topic and their opinions may be influenced by it.

There were thousands of IDs taking part in the discussion on one topic (Table 2). It is an interesting question to ask: do they contribute to the topic equally or are there some ‘‘hub'' IDs who are more important in the topic? In order to answer this question, we study the number of times n_i an ID participates in a topic. The distributions are shown in Fig. 2. It is evident that they can be described by power-law $P(n_i) \propto n_i^{-\alpha}$. The exponents are slightly different but all in the range $2 < \alpha < 3$. The power-law distributions mean that most of the users only comment once or very few times in a topic. But there are also some users who comment many times in a topic. We may call these types of users as the audience and actors, respectively. The actors play a crucial role in leading the direction of the public opinion formation of the topic followed by the audience. This could be useful for some commercial applications. For example, the topic could be about some negative news of a company, which maybe created by its competitor and the ‘‘hubs'' could belong to or depend on this competitor. In such a case, the guidance of the opinion by the hubs will be harmful for this company.

4. Model and simulation

We propose a model in order to get a better understanding of our empirical observations from Section 3. Based on our intuitive experience on comment habit, we can see that the number of the comments grows one by one. After the topic was created by a user *a*, some other users *b* and *c* et al. will comment on it later on. Then *a* would respond to their comments. *b* and *c* may come back to respond to the response to their original comments, and the process continues. Comment behavior can be regarded as a kind of communication. Someone who comments on a topic before would come back to read the response of others to his comment. And he would comment again with higher probability than other IDs. This discussion indicates

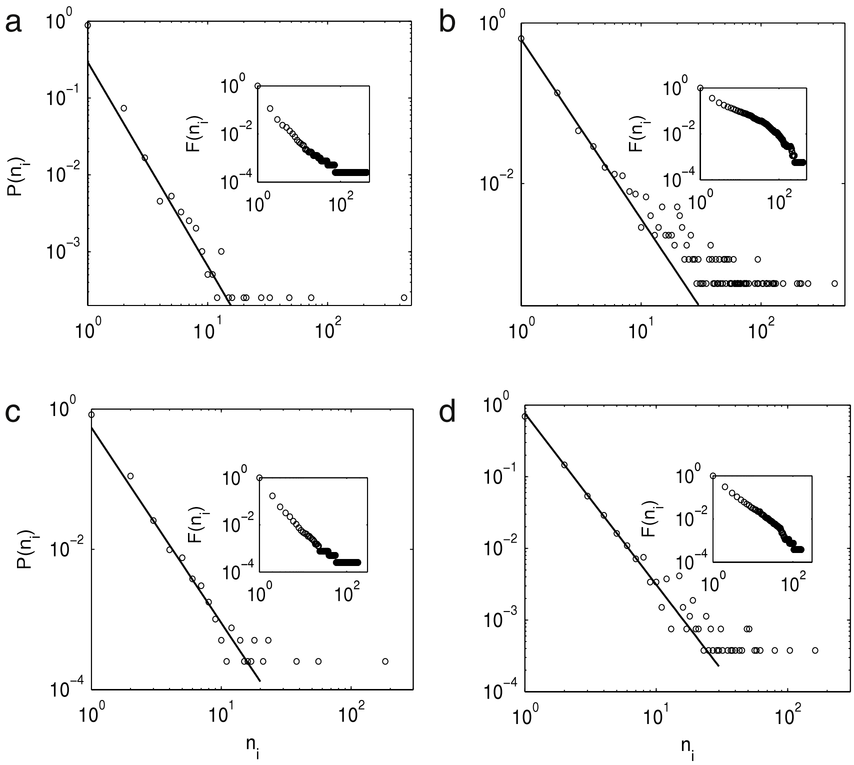

Fig. 2. The distribution of the frequency of different IDs commenting on a topic. The black line is the fitting function. In the inset we plot the cumulative probability distribution. (a) A, slope $\alpha = 2.65$, (b) B, slope $\alpha = 2.24$, (c) C, slope $\alpha = 2.78$, (d) D, slope $\alpha = 2.39$.

that the more times we participate in a topic, the higher is the probability that we make comments again. So our model is defined by the following scheme:

Step 1. Growing. A topic is created at $t = 0$. There is an ID commenting on the topic at each time step in the following time. In other words, the number of the comments increases one by one. This is a simplification of the real situation where the time needed to make a comment varies due to various length of the comments.

Step 2. Comment habit. A new comment is created with probability P by a new ID who has never commented in the topic in the past time and with probability $1 - P$ by other old IDs. The old IDs do not contribute equally to the creation of a new comment. Rather, the probability that a new comment is created by ID i is a function of the topic attraction of this ID: $\Pi(i) = \frac{A_i(t)}{\Sigma A_i(t)}$, where $A_i(t)$ is the attraction of user i at time t reflected by the number of comments $n_i(t)$ made by the ID in the past, i.e., $A_i(t) = A(0) + n_i(t)$. Here $A(0)$ represents the initial attraction, which is different for different topics.

Mathematically, the model is similar to growing networks in Ref. [24], where an existing node is selected to be connected to a new node with a probability depending on the degree k_i as $\Pi(i) = \frac{(B+k_i(t))}{\Sigma(B+k_i(t))}$. Basing on the analysis of the growing network in this work, we obtain that the distribution of n_i is a power-law $P(n_i) \propto (n_i)^{-\alpha}$at a large enough time t and with the exponent $\alpha = 2 + A(0)$.

To compare our model with data, let us take topic C in our data as an example. Here $\alpha(c) = 2.78$, and we take $A(0) = \alpha(c) - 2 = 0.78$ in the model simulation. There were 5757 comments and totally 3959 IDs made comments in topic C. As a result, there are 3959 comments which are made by new IDs when they first join the topic. Thus, we can estimate that $P = 3959/5757 = 0.687$. We then simulate the model with $A(0)$ and P obtained in this way. The results are shown in Fig. 3(a). The distribution of the simulation is indeed a power-law with a similar exponent as found from the data.

Comparing the model to the empirical data in Fig. 2, we note that there are a few IDs with extremely large n_i in the data. Based on our experience with on-line comments, this could result from the fact that some users, for example, the author and his strong supporters, have a different behavior in the topic. They comment on the topic with higher probability than others. This could be accounted in the model with larger $A(0)$ for these users. We test this by assuming that there was a user whose $A(0) = 35$ which is much larger than others, so that most of the comments by the old users of the topic in the early stage are from this particular user. The result is shown in Fig. 3(b). We can see that indeed there is an ID whose n_i is much larger than the others, as observed in the data. Comparison of the results averaged over many realizations of simulations in the insets of Fig. 3(a) and (b) shows clearly that the probability to generate large number of comments $n_i \sim 10^2$ is significantly enhanced when introducing a large value $A(0)$. So the difference between the data and the power-law fitting can be accounted by different initial attraction of some special users. In reality, $A(0)$ may follow a heterogeneous distribution and its effects will

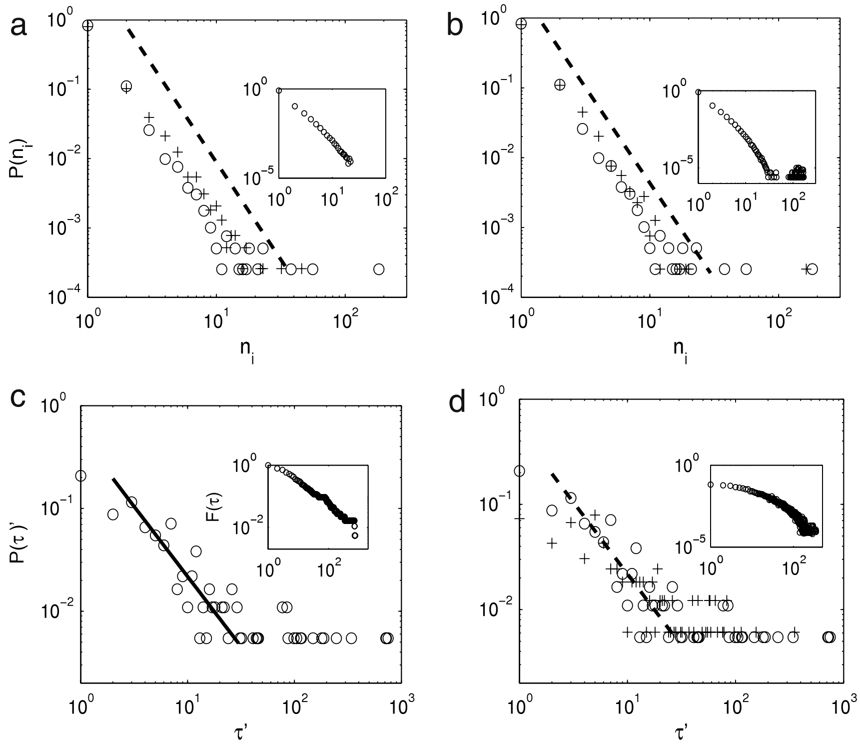

Fig. 3. The simulation result of the model (+) compared to data (○). (a–b) The distribution of frequency of different IDs commenting on a topic. Model result is from one random realization of the simulation generating the same number of comments as in the data. The slope of the dash line is the same as in Fig. 2(c). The Parameters are (a) $A(0) = 0.78$, $P = 0.687$, (b) $A(0) = 0.78$, $P = 0.687$, $A(0)_1 = 35$. The results averaged over 100 realizations with the same event size are shown in the insets. (c–d) Inter-comment time distribution of the most active ID. (c) Data, slope of the line is $\beta = 1.36$. The inset shows the cumulative probability distribution. (d) Result from one random realization of model simulation (+) compared to data (○), the slope of the dashed line is the same as in Fig. 3(c). The inset shows a non-Poisson broad distribution from the result averaged over 100 realizations of model simulations.

be analyzed in more detail in the future. These types of users with the largest $A(0)$ are the "hubs" who could deliberately try to lead public opinion formation of the topic.

It is very hard to know exactly the time used for a comment because the size of the comments can be very different. Therefore the model cannot simulate the distribution of the inter-event time directly. But we can do it in another way. We assume one comment takes one time step, i.e. the number of the comments between two comments can be taken as the time interval between the two comments. For example, the time interval between Ith and Jth comment was $\tau' = J - I$. In this way, we can study the inter-event time between two consecutive comments made by the same ID. We analyze the most active user to have good statistics. Furthermore, it is important to understand his behavior as an "opinion leader". We find that his inter-event time distribution is also a power-law, as shown in Fig. 3(c). Meanwhile, the corresponding result of the model is compared to the data in Fig. 3(d). It is interesting that our model reproduces a broad distribution, even though it does not fit precisely to the distribution from the data. This is likely due to many other complications that may not be captured by our simple model. It clearly implies that individual user's behavior is non-Poissonian. He/She may not take part in the discussion for a long time and may make comments in the topic frequently in a short time. From the analysis above, we can see that while simple, our model can well describe most important features in the human comment dynamics in on-line social systems.

5. Conclusion

In this paper, we present clearly new evidence that human comment behavior in on-line social systems, a different type of interacting human dynamics, is non-Poissonian. The inter-comment time follows a power-law distribution as many other human dynamics. A model based on personal attraction was introduced to explain the human comment behavior. Numerical simulations of the model fit well with the empirical results. Our work would be useful to understand human comment behavior in realistic society, for example, human discussion behavior in a meeting, group communications in trunked mobile telephone [25]. We expect that quantitative understanding of the human comment dynamics, when combined with additional content analysis, will open a new perspective on how to distinguish fraudulent public opinion from realistic opinion.

It is important to note that various human activities are likely interacting with each other in an intricate way. For example, the human comment behavior on a topic is expected to be related to the behavior of web site visits [17] or the interval between clicks on the comments. When commenting on a particular topic, some special groups of words could be used more frequently. This could lead to a long-tailed inter-event distribution of words, which is observed recently [26]. Following the same line of thinking, the complex structures in language in other sources [27] could be related to some other human activities. Studying the interplay among various human activities is therefore an interesting topic for future investigations.

Acknowledgements

Y. Wu thanks the financial support by China Scholarship Council. C.S. Zhou is supported by Hong Kong Baptist University. M. Chen thanks the partial support by NSFC project (No. 60804046). J.H. Xiao thanks the National Basic Research Program of China (No. 2010CB923200). J. Kurths acknowledges the support by ECONS (Leibniz Society).

References

[1] F.A. Haight, Handbook of the Poisson Distribution, Wiley, New York, 1967.
[2] P. Call, Center Staffing, The Call Center School Press, Lebanon, 2003, Tennessee.
[3] A.L. Barabási, Nature 435 (2005) 207.
[4] J.G. Oliveira, A.L. Barabási, Nature 437 (2005) 12511.
[5] A. Vázquez, J.G. Oliveira, et al., Phys. Rev. E 73 (2006) 036127.
[6] U. Harder, M. Paczuski, Phys. A 361 (2006) 329.
[7] J.P. Eckmann, E. Moses, D. Sergi, Proc. Natl. Acad. Sci. USA 101 (2004) 14333.
[8] J. Candia, et al., J. Phys. A 41 (2007) 2240156.
[9] R.D. Malmgen, D.B. Stouffer, A.E. Motter, L.A.N. Amaral, Proc. Natl. Acad. Sci. USA 105 (2008) 18153.
[10] W. Hong, T. Zhou, B.H. Wang, Chin. Phys. Lett. 26 (2009) 028902.
[11] X.P. Han, T. Zhou, B.H. Wang, New. J. Phys. 7 (2008) 0 73010.
[12] A. Vázquez, Phys. A 373 (2007) 747.
[13] A. Grabowski, Phys. A 385 (2007) 363.
[14] F.Z. Wang, S.J. Shieh, S. Havlin, H.E. Stanley, Phys. Rev. E 79 (2009) 056109.
[15] D. Rybskia, S.V. Buldyrevb, S. Havlin, et al., Proc. Natl. Acad. Sci. USA 106 (2009) 12640.
[16] A. Carbone, H.E. Stanley, Phys. A 384 (2007) 21.
[17] Dezsõ, Z. Almaas, et al., Phys. Rev. E 73 (2006) 066132.
[18] A. Chmiel, K. Kowalska, J.A. Holyst, Phys. Rev. E 80 (2009) 066122.
[19] A. Grabowski, Eur. Phys. J. B 69 (2009) 603.
[20] T. Zhou, H.A.T. Kiet, B.J. Kim, B.H. Wang, P. Holme, Europhys. Lett. 82 (2008) 28002.
[21] A. Vázquez, A. Rácz, A. Lukács, A.L. Barabási, Phys. Rev. Lett. 98 (2007) 158702.
[22] D. Gross, C.M. Harris, Fundamentals of Queueing Theory, John Wiley & Sons, New York, 1985.
[23] G.F. Newell, Applications of Queueing Theory, Chapman and Hall, London & New York, 1982.
[24] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhim, Phys. Rev. Lett. 85 (2000) 4633.
[25] D.G. Xenikos, Phys. A 388 (2009) 4910.
[26] E.G. Altmann, J.B. Pierrehumbert, A.E. Motter, PLoS One 4 (11) (2009) e7678.
[27] A. Corral, R. Ferrer-i-Cancho, G. Boleda, A. Diaz-Guilera, 2009. arXiv:0901.2924v1 [physics.soc-ph].