Proceedings of the 34th Chinese Control Conference
July 28-30, 2015, Hangzhou, China

Modelling temporal patterns of news report

YOU Zhi-Qiang, ZHU Yan-Yan, HAN Xiao-Pu*, LÜ Linyuan

Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, P. R. China

E-mail: xp@hznu.edu.cn

Abstract: We study the news system and explore news report mechanism based on the corpus of Sina news website. Instead of using complete content of news, we utilize short keywords to represent original news and mainly focus on the report time series of news. Empirical analysis shows that the distributions of time intervals follow power-law distribution with exponential cutoff on both single-class level and aggregate level. In addition, we analyze the vitality of keywords to explain why the distributions of different kinds of news show difference in the exponential cutoff tail. On the basis of these findings, we propose a hybrid mechanism queuing model to reveal the hidden principle of news reports. The simulation results can well fit the empirical data. The present study may shed some light on the mechanism of temporal patterns of news report and human selection behaviors.

Key Words: News publication, Queuing model, Distribution of time interval, Power-law distribution, Hybrid mechanism

1 Introduction

The studies of information spreading on networks have attracted an increasing attention in the academic community [1–5]. The diffusion processes and dynamical properties are greatly influenced by two important factors, namely the role of spreaders and the property of disseminule (disseminule is a general term for information, rumor, behavior and so on)[6], which are key points to differentiate between information spreading and epidemic spreading [7–10]. People are usually the information spreaders and active to make decisions in information spreading, like approving or disapproving some news, while in epidemic spreading they are passive to be affected. Therefore, building the information spreading model is actually modeling the human selection behaviors in the spreading process [2].

Compared with the information spreading which has been extensively studied [11–17], less attention has been paid to the mechanism of information generation and publication. News is the communication of selected information on current events. As one of the major communication medias, it plays an important role in information spreading. News has the natural properties of timeliness and reliability. How does people select news for publication from a vast candidates? Due to lack of scientific evidence, it is hard to answer such question. Fortunately, the latest decade has seen a growing number of scientific achievements on the statistics and dynamics of human behaviors, such as human geography activities [18–21], mail system [22, 23], text message system [24], telecommunication [25], and language systems [26, 27], et al. Empirical investigations show that Darwin and Einstein correspondence patterns and today's electronic exchanges follow the same scaling laws by analyzing the distribution of interval time between mail reception and response [23]. Similarly, the pattern of short message activity also exhibits a heavy-tailed interevent time distribution [24]. Inspired by these works, we study the properties of news system in terms of the distribution of interevent time interval between two consecutive news publication, attempting to reveal the hidden mechanism of news publication.

In this paper, we use short keywords which are obtained by segmentating the news titles to represent original news, for example, keyword *gunshoot* represents one kind of news. Then, with the information of news publication time, by calculating the time interval between two consecutive report events of the same kind of news, we can get the interevent time sequence of each kind of news. Here, we treat each single kind of news as single-class level and all kinds of news together as aggregate level. Empirical analysis shows the probability density distributions for the time intervals follow power-law distribution with exponential cutoff both on single-class level and aggregate level. Based on the empirical results, we propose a news selection model with mixed mechanisms of strict and preferential priority strategies. The model result is essentially in agreement with the empirical data which means the suggested mechanisms can be used to explain the rule of news publication.

2 Dataset

The dataset used in this paper is the full-year news of Sina website in 2012, from January 1st to December 31st. Sina website is a major news provider in China which cov-

This work is supported by National Natural Science Foundation of China under Grant Nos. 11205040, 11205042, 11305043, and the research startup fund of Hangzhou Normal University.

ers the area of politics, military, culture, sports and so forth. In addition, it contains not only the domestic news but also the international news. This dataset is collected by Kaixu Zhang from Xiamen University, which can be downloaded from *http://zhangkaixu.github.io/resources.html*. The whole dataset has been separated into four zips quarterly. Each contains news for three months, e.g., Zip 'q1' contains the news from January 1st to March 31st, Zip 'q2' for the second quarter, and so on. The dataset has in total 250,000 news, about 250 million words. Each piece of news contains following attributes: URL, character encoding, title, keywords, description, the source media and content, all of which are in XML format. In particular, URL has the report time of news, accurate to minute. Note that the contents of news are in Chinese.

We use the word segmentation tool called *PanGu Segment* to extract keywords of each news from its title and content. In order to avoid the influence of function words, we ignore the keywords with length less than 2. Since the title reflects the main point of news contents, to fetch the keywords from titles to represent news is more accurate and efficient than from the contents. Moreover, we have tested that the probability density distributions of content keywords and title keywords respectively and found that both distributions follow similar scaling law (see Fig. 1). Therefore, it is reasonable to use the keywords of news' title as abstract of its content. It should be noted that a piece of news' title may contain several different keywords, then every keyword represents one kind of news.

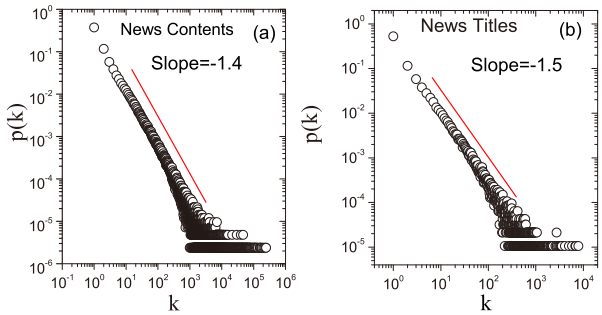

Fig. 1: Probability density distributions of keywords from news contents and titles respectively.

As we want to reveal the mechanism of news publication, we pay our attention to the publication time series naturally. Since each piece of news has the URL information with time stamp, we acquire the time information of each keyword, i.e. every category of news. Note that different keywords in the same piece of title get the same time stamp, which means the corresponding kinds of news were published at this time. Then we can have all kinds of news' time series by sorting whole keywords chronologically. Besides, to ensure the interevent time sequence is long enough, we only keep keywords which occur more than 500 times. Under these conditions, we get 331 keywords as hot news.

3 Empirical Analysis

According to the aforementioned way, on the scale of the minute, we acquire every keyword's report time series. In this section, we present the empirical results of the distributions of time interval between each two consecutive news reports both on single-class level and aggregate level. We use symbol τ to represent time interval. The results indicate that the publication behaviors of different news have some common characteristics. We observe that the distribution of interevent time between two consecutive news publication follows power-law distribution with exponential cutoff, but different keywords show different intensities in exponential cutoff tail.

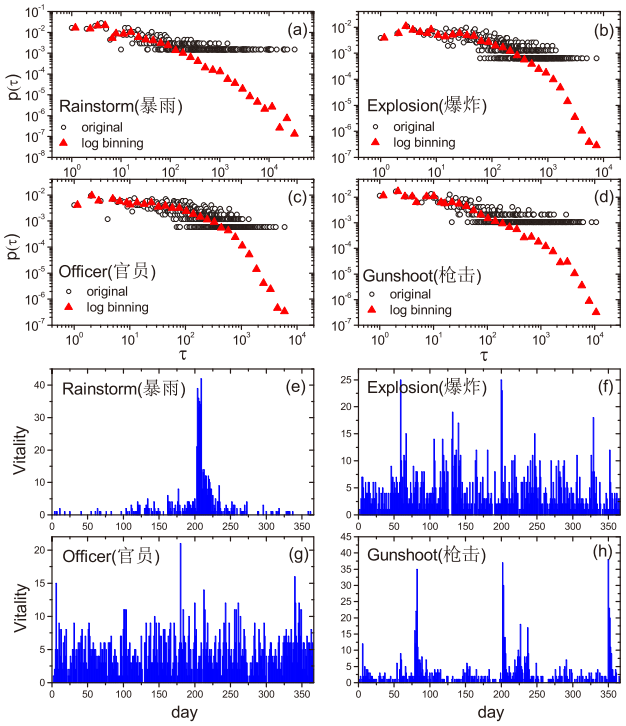

Fig. 2: (a)-(d) The distributions of interevent time intervals between two consecutive news publication and (e)-(h) the vitality of news.

We choose four typical keywords to display their properties, namely *rainstorm* (暴雨), *explosion* (爆炸), *officer* (官员) and *gunshoot* (枪击). Other hot keywords are

governed by similar non-Poisson statistics. As shown in Fig. 2(a-d), black circles represent the real data distribution and the red triangles are the corresponding logarithmic binning results. The interevent distributions are power-law with exponential cutoff which can be well fitted by $P(\tau) = A\tau^{-\alpha}e^{-\beta\tau}$, where the exponent α is between 0.25 and 0.45 and the exponent β ranges from 5.0×10^{-4} to 4.5×10^{-3} for different keywords respectively. The power-law interevent distribution indicates that the news tends to happen frequently during a short time, while sometimes appears after a long period of time. The cutoff reveals some randomness may occur during news publication process. In Table 1, we show the two exponents of nine hot keywords.

Table 1: The fitting results of distributions of news' interevent time series and the corresponding Burstness B and Memory M. The corresponding Chinese characters are presented in the brackets.

News	α	β	B	M
American (美国)	0.45	0.00354	0.21	0.17
Beijing (北京)	0.25	0.00220	0.21	0.11
Diaoyu Islands (钓鱼岛)	0.50	0.00100	0.55	0.23
President (总统)	0.45	0.00250	0.20	0.11
Earthquake (地震)	0.35	0.00200	0.25	0.21
Rainstorm (暴雨)	0.30	0.00050	0.54	0.43
Explosion (爆炸)	0.35	0.00200	0.21	0.15
Officer (官员)	0.25	0.00200	0.15	0.07
Gunshoot (枪击)	0.30	0.00300	0.30	0.24

Generally speaking, in the news report process, a kind of news is often consecutively reported due to seasonal outbreak like rainstorm or the enthusiasm of the masses for breaking news like earthquake. These kinds of news often need repeated publication to fulfill the audience's expectations. Along with the changes of weather or the reduction of the masses' enthusiasm, the recently reported news will become less popular.

At the same time, although the similar behaviors of different kinds of news have been observed, there are still differences laying on the intensity in the cutoff tail. Compared with news *explosion* and *officer*, news *rainstorm* and *gunshoot* show weaker tendency in exponential cutoff tail. In order to explain the different intensities, we further analyse the vitality, burstness and memory [28] of news. For a given news, we define the publication times per day as its vitality. Fig. 2(e-h) display the distributions of the four keywords' vitalities. News *rainstorm* shows strong burstness between June and July and is relatively silent among other months. It is because the widespread flooding often occurs during that season. We also notice that *gunshoot* displays strong periodicity and is active in day 80, 200, 350. Comparatively speaking, *explosion* and *officer* exhibit much strong randomness which leads to more apparent exponential cutoff tail. And burstness B is a statistics term referring to the intermittent increases and decreases in activity or frequency of an event. According to Ref. [28], B is defined as:

$$B = \frac{c_v - 1}{c_v + 1}, \tag{1}$$

where c_v is the coefficient of variation of time intervals tau. Burstness value is between -1 and 1. B equals 1 for the most bursty series, 0 for the poisson distribution, and -1 when the series is completely periodic signal. The memory value M measures the correlation of consecutive inter-event time intervals. M is defined as:

$$M = \frac{1}{n_\tau - 1}\sum_{i-1}^{n_\tau - 1}\frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1\sigma_2}, \tag{2}$$

where n_τ is the number of interevent times measured from the signal and $m_1(m_2)$ and $\sigma_1(\sigma_2)$ respectively are sample mean and sample standard deviation of τ_i's (τ_{i+1}'s), $(i = 1, \cdots, n_\tau - 1)$. M is bounded in the range [-1,1], $M > 0$ for memory while $M < 0$ for anti-memory. We determine the series displays memory when a short(long) interevent time often follows a short(long) one, and it shows anti-memory when repeated pattern hardly exists. We present the burstness and memory values of some keywords in Table 1. News *rainstorm* and *gunshoot* show stronger burstness and memory than *explosion* and *officer*.

Meanwhile, we investigate all keywords' time intervals between two consecutive news reports with same key words. The distribution is presented by purple squares in Fig. 4(a). The pattern of news publication at aggregate level also displays power-law distribution with exponential cutoff, which can be well fitted by $y = a(x + b)^{-c}$, where $a = 280$, $b = 190$ and $c = 2.05$, presented by the solid line in Fig. 4(a).

4 The Model

The process of journalists choosing news from a number of candidates is very similar to the decision-based queuing process. Thus, on the basis of the empirical findings above and the early work of Barabási in 2005 [18], we propose a hybrid mechanism queuing model considering the timeliness to simulate the news publication behavior. The three core principles of this model are as follows: (i) Strict priority. Under this rule, we choose the news which has

the highest importance from candidates, like breaking news *earthquake*; (ii) Preferential selection [29, 30]. This mechanism allows us to choose news contents in proportion to their degree of importance. The more important a piece of news is, the higher possibility it will be chosen. Under this condition, news with less importance still has a chance to be selected; (iii) Information redundancy and strong timeliness. Redundancy refers to the number of candidate news is much larger than the amount of publication news. Timeliness means the reported news is always selected from the set of latest candidates. Therefore, news that can not be reported in time will lose their significance for publication.

We define C categories of news with each category endowed with a fixed weight ω ranging from 0 to 1, to represent the importance of news in this category. This is because the importance of specific category will not change markedly, for instance, *gunshoot* and *president* always have absolute priority. The detailed model rules are as follows:

(i) At time $t = 0$, we initialize a news candidates list with length L by randomly choosing L pieces of news from C categories of news. Each news has a fixed weight ω which is determined by its category and has been set before.

(ii) With probability q, the model will use strict priority mechanism to select news with the largest ω from all candidates. Otherwise, the model will follow the preferential selection with probability $1 - q$, where a news i will be selected with probability $\Omega_i = \omega_i / \Sigma\omega_i$.

(iii) Once a news has been selected, it will be removed from the candidates list, and simultaneously a new one randomly chosen from C categories will be added into the list. We record the time of news removal and birth.

(iv) The news that exist in the candidates list for over $L/20$ time steps will be removed directly and replaced by a new one. This mechanism is necessary due to the consideration of timeliness of news selection, namely the news will be meaningless if it has not been reported for a long period of time. In reality, people will barely be interested in the old stories.

The whole iterations of this simulation process are $24 \times 60 \times 366$ steps which is in agreement with the total minutes of year 2012. After the completion of model iteration, we extract each category of news' interevent time series to investigate the distribution of interevent time series at the aggregate level.

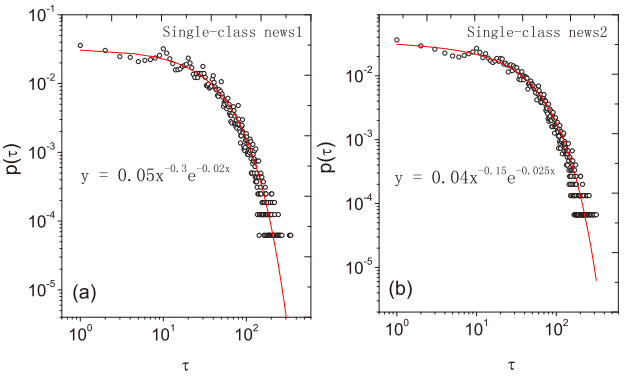

Fig. 3: Distributions of interevent time at single-class level.

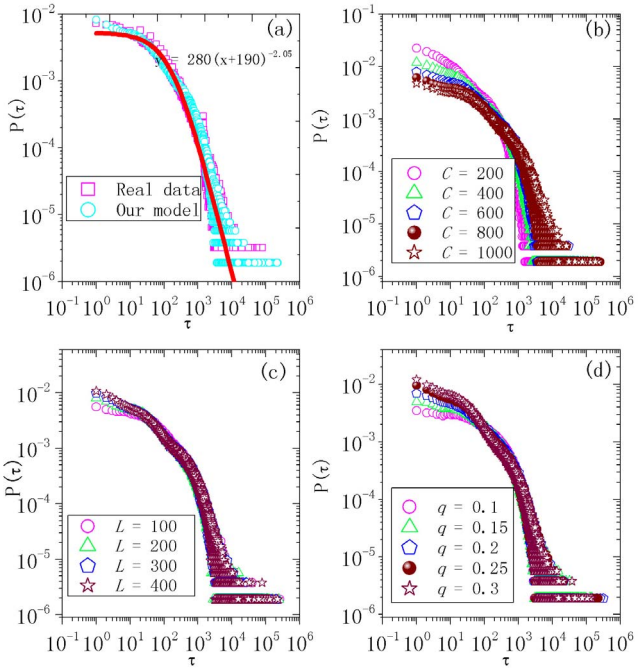

Fig. 4: Distributions of interevent time. (a) The comparison between the empirical analysis and result of our model at the aggregate level. (c)-(d) The parameters' effects on simulation results of the model.

5 Results and analysis

Fig. 3 shows the distributions of interevent time series at single-class level simulated by our model. We select two kinds of news with longest publication time series to demonstrate the model results. The model parameters are set as $C = 600$, $L = 200$ and $q = 0.22$. Fig. 3 display the single-class news' distributions of interevent time intervals follow power-law with exponential cutoff form, which can be fitted by $P(\tau) = A\tau^{-\alpha}e^{-\beta\tau}$. The results of single-class news are in agreement with the empirical analysis.

Fig. 4(a) compares the distributions of interevent time series between model result (circles) and empirical investi-

gation (squares) at the aggregate level. The red line is the fitting function $y = 280(x + 190)^{-2.05}$. The model result essentially accords with the empirical analysis.

Furthermore, we analyze the influence of different parameters on model results, as shown in Fig. 4(b-d). Fig. 4(b) displays the influence caused by the news category C. We fix $L = 200$ and $q = 0.22$, then select $C = 200, 400, 600, 800, 1000$, respectively. With the increasing of C, the distribution shows an downward trend when τ is less than 300, and has a significant right shift when τ is larger than 500. Fig. 4(c) presents the influence of L, where C and q are fixed as 600 and 0.22, respectively. We analyze the results with L changing from 100 to 400 with step 100. Only minor differences are observed in the region $[1, 10]$. The bigger the L is, the higher the proportion of small τ is. However, the tail part of the distribution remains almost the same. Fig. 4(d) shows the impacts of probability q, which has similar effect as parameter L. Differences are observed in the region $[1, 100]$ and the tail parts keep unchanged.

The growth of C indicates that the number of candidates becomes larger. Since in the candidate list new added news will be randomly selected from C categories, then with the increasing of C, the chosen probability of each category will decrease accordingly, leading to the decrease of the co-occurrence probability of news in same category in the candidates list. As a result, it is less likely to frequently choose the news in the same category in a short time. Overall, all kinds of news' time interval τ between two consecutive news publication and its frequency $P(\tau)$ tend to be larger with the increasing of C. The growth of L increases the co-occurrence probability of news in same category in the candidates list, thus there is a great possibility to have same category of news published within a short time. However as our model has a bias that news with more weight has higher priority to be published, the news with larger ω will be affected more significantly than others. Additionally, as there is a small number of news with large ω, the frequency of small τ increases slightly. Besides, we observe that parameter C has a greater impact than L on the model results. Moreover, the increase of the selection probability q makes it more likely that news in the same category with larger ω has higher chance to be chosen frequently, resulting in the rise of the frequency of small τ. In summary, news with high weight is more sensitive to all parameters' change (corresponds to the head part), while news with low weight seems unaffected markedly by L and q (corresponds to the tail part), but are influenced greatly by the number of news category C.

6 Discussion

Since news as one of mainstream media plays essential role in information spreading, we pay our attention to news publication process. Empirical analysis showed that the distribution of interevent time intervals between two consecutive news publication, both on single-class level and aggregate level, follows power-law with exponential cutoff. To uncover the underlying rule governing news generation, inspired by the decision-based queuing model presented by *Barabasi* in 2005 [18], we proposed a queue-based news selection model with hybrid mechanisms which considers the strict and preferential priority mechanisms, as well as timeliness of news selection. The model produces rich non-Poisson characteristics of interevent time intervals and the simulation result can fit the real data very well.

Note that although the present model assumes the time intervals of each kind of news' advent are homogeneous, the burstness of the time intervals of news' publication still exists due to the preferential selection of news. Our studies contribute to better understanding the mechanism of news report and human selection behaviors, and may shed some light on similar studies of other media's selection rules, such as magazines and movies.

References

[1] G. Miritello, E. Moro, and R. Lara, Dynamical strength of social ties in information spreading, *Physical Review E*, 83(4), 2011: 045102.

[2] L. Lü, D.-B. Chen, and T. Zhou, The small world yields the most effective information spreading, *New Journal of Physics*, 13(12), 2011: 123005.

[3] C. Dutta, et al. On the complexity of information spreading in dynamic networks, *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 2013.

[4] C. Liu, and Z.-K. Zhang. Information spreading on dynamic social networks, *Communications in Nonlinear Science and Numerical Simulation*, 19(4), 2014: 896-904.

[5] A. Clementi, R. Silvestri, and L. Trevisan, Information spreading in dynamic graphs, *Proceedings of the 2012 ACM symposium on Principles of distributed computing*, ACM, 2012.

[6] M. Zheng, L. Lü, and M. Zhao, Spreading in online social networks: The role of social reinforcement, *Physical Review E*, 88(1), 2013: 012818.

[7] T. Zhou, Z.Q. Fu, and B.-H. Wang, Epidemic dynamics on complex networks, *Progress in Natural Science*, 16(5), 2006: 452-457.

[8] N. Madar, et al, Immunization and epidemic dynamics in complex networks, *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2), 2004: 269-276.

[9] T. Gross, C.J.D. D'Lima, and B. Blasius. Epidemic dynamics on an adaptive network, *Physical Review Letters*, 96(20), 2006: 208701.

[10] M. Barthelemy, et al, Dynamical patterns of epidemic outbreaks in complex heterogeneous networks, *Journal of Theoretical Biology*, 235(2), 2005: 275-288.

[11] B. Haeupler, Simple, fast and deterministic gossip and rumor spreading, *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 2013.

[12] M. Nekovee, et al, Theory of rumour spreading in complex social networks, *Physica A: Statistical Mechanics and its Applications*, 3741, 2007: 457-470.

[13] P.G.Lind,et al, Spreading gossip in social networks, *Physical Review E*, 76(3), 2007: 036117.

[14] A. Guille, et al, Information diffusion in online social networks: A survey, *ACM SIGMOD Record*, 42(1), 2013: 17-28.

[15] E. Bakshy, et al, The role of social networks in information diffusion, *Proceedings of the 21st international conference on World Wide Web*, ACM, 2012.

[16] J. Yang, and J. Leskovec, Modeling information diffusion in implicit networks, *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, IEEE, 2010.

[17] J. Zhao, J. Wu, and K. Xu, Weak ties: Subtle role of information diffusion in online social networks, *Physical Review E*, 82(1), 2010: 016105.

[18] A.-L. Barabási, The origin of bursts and heavy tails in human dynamics, *Nature*, 435(7039), 2005: 207-211.

[19] M.C. Gonzalez, C.A. Hidalgo, and A.-L. Barabasi, Understanding individual human mobility patterns, *Nature*, 453(7196), 2008: 779-782.

[20] C. Song, et al, Limits of predictability in human mobility, *Science*, 3275968, 2010: 1018-1021.

[21] D. Brockmann, L. Hufnagel, and T. Geisel, The scaling laws of human travel, *Nature*, 4397075, 2006: 462-465.

[22] R.D. Malmgren, et al, A Poissonian explanation for heavy tails in e-mail communication, *Proceedings of the National Academy of Sciences*, 105(47), 2008: 18153-18158.

[23] J.G. Oliveira, and A.L Barabasi, Human dynamics: Darwin and Einstein correspondence patterns, *Nature*, 437(7063), 2005: 1251-1251.

[24] H. Wei, X.-P. Han, T. Zhou, and B.-H. Wang, Heavy-tailed statistics in short-message communication, *Chinese Physics Letters*, 26(2), 2009: 028902.

[25] J. Candia, et al, Uncovering individual and collective human dynamics from mobile phone records, *Journal of Physics A: Mathematical and Theoretical*, 41(22), 2008: 224015.

[26] L. Lü, Z.-K. Zhang, and T. Zhou, Deviation of Zipf's and Heaps' Laws in Human Languages with Limited Dictionary Sizes, *Scientific Reports*, 3, 2013: 1082.

[27] L. Lü, Z.-K. Zhang, and T. Zhou, Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems, *PLoS ONE*, 5(12), 2010: e14139.

[28] K-I. Goh, and A-L. Barabasi. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4), 2008: 48002.

[29] D. Jong, K. Alan, Analysis of the behavior of a class of genetic adaptive systems, 1975.

[30] E.A. Jones, and W.T. Joines, Design of Yagi-Uda antennas using genetic algorithms, *Antennas and Propagation, IEEE Transactions on*, 45(9), 1997: 1386-1392.