The Workshops of the Tenth International AAAI Conference on Web and Social Media
Wiki: Technical Report WS-16-17

What Can Wikipedia Tell Us About the Global or Local Character of Burstiness?

Yerali Gandica and **Renaud Lambiotte** and **Timoteo Carletti**
Department of Mathematics and Namur Center for Complex Systems - naXys,
University of Namur, rempart de la Vierge 8,
5000 Namur, Belgium.
Corresponding author: ygandica@gmail.com

Abstract

In this communication we take advantage of the global covering character of Wikipedia dataset to analyze the dependence of the usual coefficients used to measure burstiness respect to language. Analyzing separately the patterns for single editors over several pages, we show several characteristics of the super-editors in the WP written in English, Spanish, French and Portuguese. We report for the first time the Burstiness and Memory effect coefficients, separately for the 4 WP's, showing similitudes and differences for all the users respect to the super-editors, the exponent for their averaged inter-event activity and finally some statistical traces for their averaged monthly activity.

The digital media are an important component of our lives. Nowadays, digital records of human activity of different sorts are systematically stored and made accessible for academic research. Hence a huge amount of data became available on the past couple of decades, which allows for a quantitative study of human behaviour, opening progressively, the possibility to uncover some social patterns not detected so far (Barrat, Barthélemy, and Vespignani. 2008; Newman, Barabasi, and Watts. 2006).

The success of research in digital social patterns hinges on the access to high quality data. Even though the availability of recorded data and its accessibility are rapidly increasing, many data sets are not freely available for research. Wikipedia (WP) is an important exception, as not only is it considered a robust and trustworthy source of information (Giles 2005), but it is also easily accessible via the API (https://www.mediawiki.org/wiki/api:main page) or the different available dumps (http://wwm.phy.bme.hu/) by anyone with connection to internet.

In this communication we take advantage of the global covering character of Wikipedia (WP) dataset to answer the question: which of the usually used measurements for burstiness are global or have local dependence, in this case constrained to the language. Human bursty behaviour, is the mankind activity characterized by intervals of rapidly occurring events separated by long periods of inactivity (Barabási 2005). This phenomenon has been found to modulate several kind of human activities, such as sending letters, writing email messages, sending mobile SMS, making phone calls and web browsing, among others (Vázquez et al. 2006; Goh. and Barabási 2008; Wu et al. 2010; Malmgren et al. 2008; 2009; Ratkiewicz et al. 2010; Jo et al. 2012).

The main characteristics of a bursty behaviour is a power-law distribution of the inter-event activity, i.e, the interval of time between consecutive actions or events. The exponent of the power-law distributions have been reported as closely distributed around an universal value, which takes values of 1 in Web browsing, email, and library datasets, while $3/2$ for mail correspondence patterns (Vázquez et al. 2006). Under the premise of queuing process -when individuals execute tasks based on some perceived priority- as the origin of human burstiness, the change of the exponent value was suggested to depends on if there are or not limitations on the number of tasks an individual can handle in a finite time (Vázquez et al. 2006). For the case of Wikipedia (WP), the exponent for the averaged inter-event distribution over a sample of the 100 most active editors has been reported as 1.44 (Yasseri and Kertész 2013).

Another parameter based on the broad distribution of the inter-events, comparing the variance respect to the mean for the inter-events, has been defined as B parameter by Goh et al. (Goh. and Barabási 2008). In the same work the authors also defined the memory coefficient, M, to measure the probability to have short (large) inter-events followed by short (large) ones. In this work we report for the first time both values for Wikipedia data-set, showing similitudes and differences respect to the super-editors in the edition of WP written in 4 different languages. With this picture in mind, next we show the probability distribution function of the inter-events (in seconds) averaged over such super-editors. Then we show the averaged monthly activity and finally the averaged cumulative monthly activity for all of them, separately in the 4 studied WP.

Our data sample for the WP editors consist of the four separated WP dumps (http://wwm.phy.bme.hu/): The one written in English (EN-WP), the Spanish one (ES-WP), the French WP (FR-WP) and the Portuguese one (PT-WP). All of them in the period of about 10 years ending in January 2010. The accessible data contain the whole editing history record for both pages and editors. For each entry the "light dump"has the WP page name, the edit time stamp and the identification of the editor who did the changes. We dis-

card entries associated to IP's and only consider editors who login before editing, so that the editor is univocally identified. Moreover only editors with more than 2000 edits are considered, in order to reduce the impact of outliers. This number is a good compromise to have enough reasonably active editors in this time span. The universe of the sample is 10473 editors in EN-WP, 1110 in ES-WP, 955 in FR-WP and 551 in PT-WP. Super editors are defined as the editors with a rank greater than 25% of the highest rank in that WP and with more than 1 year of editing activity. We define the rank for each editor, as her/his total number of editions divided by the total number of days, since the editor started to edit. The number of super-editors is 20 in EN-WP, 10 in ES-WP, 15 in FR-WP and 24 in PT-WP. We have checked there are neither WP-bots nor blocked editors in our super-editors list. In figure 1 we show the rank plot for all the editors with more than 2000 editions in semi filled area, for the 4 WP's. The darker area shows the super-editors. In the inset of each figure is shown a zoom for the better visualization of the super-editors zone.

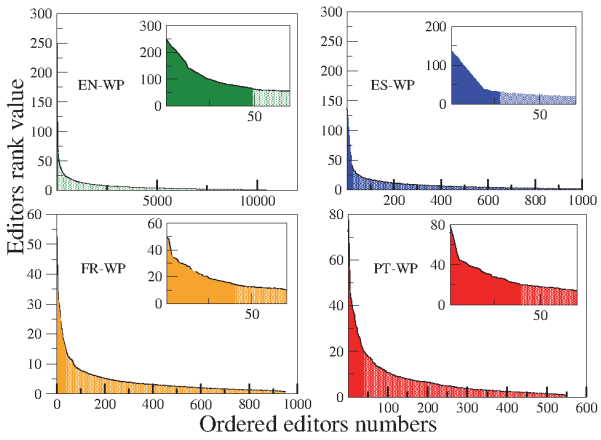

Figure 1: Rank plot for all the editors with more than 2000 editions in semi filled area, for the 4 WP's. The darker area shows the super-editors zone. In the inset of each figure is shown a zoom for the better visualization of the super-editors zone. Some editors in the darker zone were not analyzed as super-editors because they did not edit for more than one year.

In figure 2 (circles) is depicted the (M, B) value for all the considered editors in each WP. The Burstiness parameter, B, is defined for each editor as:

$$B_i \equiv \frac{\sigma - m}{\sigma + m}\,, \qquad (1)$$

where m and σ are, respectively, the mean and standard deviation of the set of values τ_n, which refer to difference in time between each pair of consecutive events. This parameter measures the magnitude of the variance in comparison to the mean, and goes from -1 ($\sigma \ll m$) for periodic signals to 1 ($\sigma \gg m$) for strongly bursty distributions. The memory coefficient M, for each editor is defined as:

$$M_i \equiv \frac{1}{n_\tau - 1} \sum_{j=1}^{n_\tau - 1} \frac{(\tau_j - m_\tau)(\tau_{j+1} - m'_\tau)}{\sigma_\tau \sigma'_\tau}\,, \qquad (2)$$

where n_τ is the total number of inter-event for that particular editor, and m_τ and σ_τ (m'_τ and σ'_τ) are, respectively, the average and the standard deviation of the set of all inter-event times except the last one (the first one). This coefficient is positive when short (large) inter-event times tend to be followed by short (large) inter-event times, negative in the opposite case and otherwise close to zero. We can see all the values are concentrated in high values of Burstiness, while small but positive Memory effect. In bigger white circles is represented the averaged values for (M, B) over all the editors, and in violet the averaged values over all the super-editors. Except for the case of the EN-WP, the other 3 WP's show the same (M,B) spatial distributions and interestingly, also the super-editors values (in squares) show the same distribution according to their proportion, in such as cases, super-editors behaviour are representative for normal editors, which are limited to have less data to be analyzed. However this is not the case for the EN-WP, where the super-editors are not representative of the whole community.

In fig. 3 is reported the probability distribution function of the inter-events (in seconds) averaged over all the super-editors in the 4 Wp's. The power-law exponents were

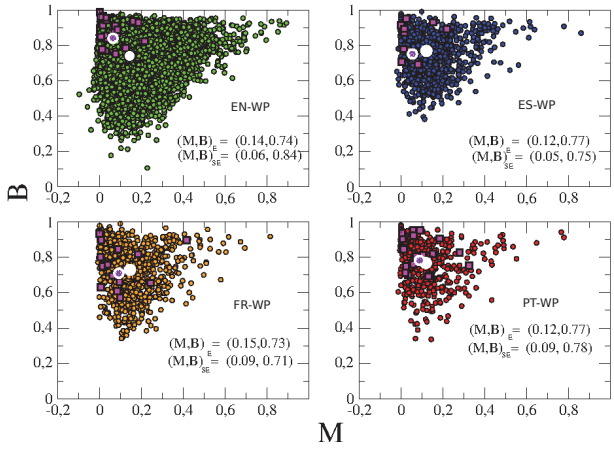

Figure 2: In circles, the scatter for the (M, B) value of each editor in the 4 WP's. Pink squares show the (M, B) values for the super-editors in each set. In bigger white circles is represented the averaged values for (M, B) over all the editors, and in violet the averaged values over all the super-editors. In each plot is indicated the resulting (M,B) values averaged over the whole community, $(M, B)_E$, and over all the super editors, $(M, B)_{SE}$. Except for the case of the EN-WP, the other 3 WP's show the same (M,B) spatial distributions and interestingly, also the super-editors values (in squares) show the same distribution according to their proportion.

found -1.73 ± 0.02 for the EN-WP, -1.73 ± 0.02 for the ES-WP, -1.43 ± 0.01 for the FR-WP and -1.63 ± 0.01 for the PT-WP. The exponent for the same probability distribution but averaged over a sample of the 100 most active editors over all the WP's has previouly reported as 1.44. (Yasseri and Kertész 2013). Although the value for the exponents coincide between the EN and ES-WP, there was not any similarity between them when the Burstiness and Memory coefficients were analyzed (fig. 2).

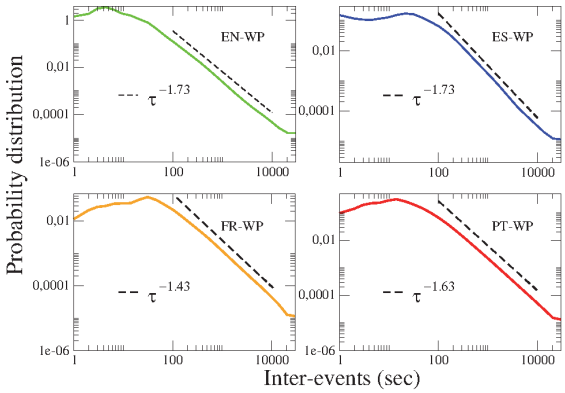

Figure 3: Probability distribution function of the inter-events (in seconds) averaged over all the super-editors in the EN-WP (green), ES-WP (blue), FR-WP (yellow) and PT-WP (red). The best fitted power law function is showed for each WP.

In order to analyze the effect that the diversity of behaviours could have over the previous averaged values, in figure 4 is plotted, with dots, the averaged monthly activity for all the super-editors in the 4 WP's and in shading the standard deviation. The standard deviation for the EN-WP achieves the 97% of the mean, while the ES-WP achieves the 94%, different from the FR and PT-WP which only achieve to the 68% and 74, 5%, respectively. These results seem to indicate a possible relation between the value of the exponents and the variability (standard deviation) of behaviours in each WP community.

Finally, in order to have a normalized point of view to compare the 4 WP's, we show in figure 5 the normalized average cumulative monthly activity for all the super-editors in the 4 WP's. The standard deviation is also shown in each month. We can only remark a different behaviour for the EN-WP. This could be caused for the broadly geolocalized community who writes in the English WP. In lesser degree the ES-WP shows different behaviour for the last months of the year.

To summarize, in this work we use Wikipedia data-sets to study the dependence of the usual coefficients used to measure burstiness with language. Moreover, the behaviour of the super-editors was analyzed with respect to the whole population in each set. We studied the 4 WP data-sets written in English, Spanish, French and

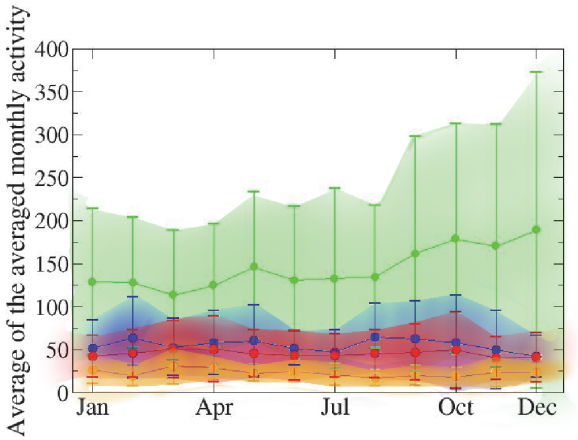

Figure 4: Averaged monthly activity in dots and in shadow the standard deviation for all the super-editors in the EN-WP in green (light gray), ES-WP in blue (gray in the middle part), FR-WP in yellow (gray in the lower part) and PT-WP in red (dark gray).

Portuguese. Studying the B and M coefficients, we found that, in all the WP's, the editors are characterized by high values of Burstiness, but small and positive Memory effects. Super-editor behaviours are representative of each complete community in ES-WP, FR-WP and PT-WP but not in EN-WP, where their wider stamp data cannot be analyzed as representative of the whole set. In this case super-editors have higher values of burstiness and lower memory effects. One possible explanation is that the majority of such EN-WP super-editors could be WP administrators and for that reason never enter in edit wars. The tendency to continuously edit during edit wars can increase the probability to have short inter-events followed by short ones, and hence to increase the Memory effects. On the other hand, edit wars can trigger very short inter-events, which increases the B value. The relation between burstiness and memory effects on the one hand and edit wars on the other is a natural extension for the present work, to be studied in the future.

The exponent values for the super-editors inter-event distribution, evaluated separately for each WP, were also reported, and the highest values were found for the EN and ES WP's. The diversity at the time scale of months along the year was found higher for the same WP's, which seems to suggest a relation between the power-law exponents and the diversity of behaviours among the editors in this time scale. Finally, we found the EN-WP showed the most dissimilar patterns in the monthly activity along the year, which could be a natural consequence of the broadly geo-localized community who writes in the English WP.

Acknowledgements

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimisation), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The

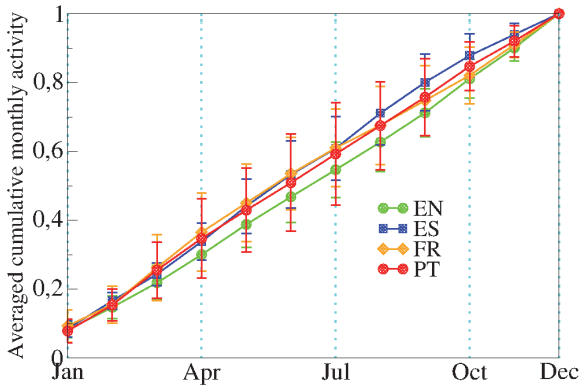

Figure 5: Normalized averaged cumulative monthly activity for all the super-editors in the 4 WP's. The standard deviation is also shown in each month. The plot shows different behaviour for the EN-WP. This could be caused for the broadly geolocalized community that writes in the English WP.

authors thank Julieta Barba Gandica by her voluntary work in detecting some of the editors' location by internet.

References

Barabási, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435:207.

Barrat, A.; Barthélemy, M.; and Vespignani., A. 2008. *Dynamical Processes on Complex Networks.* Cambridge.

Giles, J. 2005. Internet encyclopaedias go head to head. *Nature* 438:900.

Goh., K.-I., and Barabási, A.-L. 2008. Burstiness and memory in complex systems. *EPL* 81:48002.

Jo, H.-H.; Karsai, M.; Kertesz, J.; and Kaski, K. 2012. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics* 14:013055.

Malmgren, R. D.; Stouffer, D. B.; Motter, A. E.; and Amaral, L. A. N. 2008. Poissonian explanation for heavy tails in e-mail communication. *PNAS* 105(47):18153.

Malmgren, R. D.; Stouffer, D. B.; Campanharo, A. S. L. O.; and Amaral, L. A. N. 2009. On universality in human correspondence activity. *Science* 325:1696.

Newman, M.; Barabasi, A.-L.; and Watts., D. 2006. *The Structure and Dynamics of NETWORKS*. Princeton University Press.

Ratkiewicz, J.; Fortunato, S.; Flammini, A.; Menczer, F.; and Vespignani, A. 2010. Characterizing and modeling the dynamics of online popularity. *PRL* 105:158701.

Vázquez, A.; Oliveira, J. G.; Dezsö, Z.; Goh, K.-I.; and Barabási, K. A.-L. 2006. Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E* 73:036127.

Wu, Y.; Zhou, C.; Xiao, J.; Kurths, J.; and Schellnhuber, H. J. 2010. Evidence for a bimodal distribution in human communication. *PNAS* 107:18803.

Yasseri, T., and Kertész, J. 2013. Value production in a collaborative environment sociophysical studies of wikipedia. *J Stat Phys* 151:414439.