Affinity Networks as a Tool for Assessing Writing Processes: A Novel Method Utilizing Pause-Based Visibility Graphs

Davi Alves Oliveira[1,4], **Erica dos Santos Rodrigues**[2], **Hernane Borges de Barros Pereira**[1,3,4]

[1] Universidade do Estado da Bahia (UNEB)
41.150-000 – Salvador – BA – Brazil

[2]Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
22.451-900 – Rio de Janeiro – RJ – Brazil

[3]Centro Universitário Senai Cimatec
41.650-010 – Salvador – BA – Brazil

[4]Programa de Pós-Graduação em Difusão do Conhecimento
(UFBA, UNEB, IFBA, UEFS, LNCC, Senai Cimatec),
40.110-100 – Salvador – BA – Brazil

davi.alvesoliveira@gmail.com, ericasr@puc-rio.br,

hbbpereira@gmail.com

Abstract. *This study proposes utilizing affinity networks to characterize writing profiles by examining pausing behavior during writing tasks. Data from 34 participants was collected using Inputlog. Pauses were used to create visibility graphs, and the resulting graph metrics were used to build an affinity network. The network was then analyzed to group writers based on their pausing behavior. Results indicate that this method can be used to identify different writing profiles of writers with distinct proficiency in text production. Future research should investigate if changes in pausing behavior can improved writing proficiency.*

1. Introduction

Numerous methods exist for analyzing writing proficiency. One such method involves using keystroke logging to capture writers' behaviors. Specifically, pauses captured during keystroke logging may indicate moments of higher cognitive load, such as during planning or revision [Medimorec and Risko 2017]. Comparing the pausing behavior of writers at different levels of proficiency can provide valuable insights into writing cognitive processes (e.g., [Shen and Chen 2021]) and inform the development of effective writing instruction strategies.

One challenge in analyzing pausing behavior during writing tasks is the vast number of data points collected from each participant. We propose using visibility graphs [Melo et al. 2017] to analyze pausing behavior during writing, as this approach allows for consideration of all data points and provides a technique for describing pausing behavior in detail. The visibility graphs are created from the unevenly-spaced time series formed by the consecutive pauses between writing-related events.

Moreover, we propose utilizing metrics derived from the visibility networks as 'genes' for each writer's 'chromosomes' to construct an affinity network [Monteiro et al. 2014, Monteiro et al. 2015]. In affinity networks, vertices are connected based on their degree of similarity. Thus, writers with similar pausing behavior are connected, forming communities based on their shared characteristics.

The objectives of this study are (1) to demonstrate how visibility graphs, generated from unevenly-spaced time series of writing-related pauses, capture differences in pausing behavior among different writers and (2) to investigate how the identification of communities in an affinity network suggests differences in writing profiles among individuals.

2. Related Works

To the best of our knowledge, no published study has employed visibility graphs or affinity networks in investigating writing processes. By contrast, several studies have focused on analyzing pauses during the writing process to better understand the cognitive processes involved. A brief review of the literature was conducted by using the string '("pauses" OR "pausing behavior") AND "writing"' in the Google Scholar search. Articles were selected from their titles and abstracts according to their relevance for the discussion of pausing behavior in the investigation of writing and the use of keystroke logging data as part of the method. Some of these studies are summarized in sequence.

[Alves et al. 2008] investigated the cognitive processes related to writing and found that planning and revising were predominantly activated during pauses. [Medimorec and Risko 2017] studied pauses during writing tasks using keystroke logging data and discovered that they "... are related to various aspects of writing, regardless of transcription fluency and genre" (p. 1267). They also identified that the location of pauses is significant, with pauses more frequent at paragraph boundaries than at sentences or word boundaries, indicating that pauses are linked to cognitive processes such as sentence and text planning, as well as lexical access. [Guo et al. 2018] investigated pauses occurring between the typing of two characters of a word and found that it is significantly associated with human scores. [Conijn et al. 2019] compared features from keystroke logging data across different writing tasks with different associated cognitive loads and discovered significant differences, suggesting that these features capture differences in cognitive demands. Finally, [Valenzuela and Castillo 2022] investigated differences in pausing behavior across communicative purposes (to persuade and to inform) and reading media (print and digital) at three stages of writing (beginning, middle and end) and found that pausing behavior does not change across reading media, but the communicative purpose interacts with the stage of writing, with longer pauses observed at the end of the writing process when the communicative purpose was to persuade compared to inform. Overall, these studies provide valuable insights into the relationship between pausing behavior and cognitive processes during the writing process. However, a full systematic review of the literature could shed more light on these discussions and, thus, is a suggestion for future studies.

3. Materials and Methods

We obtained keystroke data from 34 university students at PUC-Rio using Inputlog [Leijten and Van Waes 2013]. The data collection was part of a project that received ap-

proval from the PUC-Rio Ethical Review Board under opinion number 2016-09. The participants were 17 writing tutors (undergraduate students with high writing skills) and 17 undergraduate students with writing difficulties. They were asked to write a text on the theme "Money and happiness" during a 50min writing session. We used the XML files from Inputlog general analysis output, one file for each participant, each one containing from 1683 to 5814 events ($M = 3377.09, SD = 1116.56$). The events logged by Inputlog are mouse (e.g., clicks and movement) and keyboard events (e.g., keystrokes).

Time series were created from the pause duration of the events by using the order of the event and its duration. This method results in unevenly-spaced time series in which the duration between events is abstracted and the order of the event is instead used as the event index [Aris et al. 2005].

3.1. Visibility Graphs

A graph $G = (V, E)$ is a mathematical structure formed by two finite sets, with V being the set of vertices and E the set of edges formed by binary relations over V [Gross et al. 2019, p. 2]. In this study, graphs are not allowed to have self-loops or multiple edges. A visibility graph is a graph constructed from a time series according to the visibility criteria defined by [Lacasa et al. 2008] and described by [Melo et al. 2017]. To test the criteria, each point in a time series is considered a vertex in the visibility graph. Then, two vertices V_a and V_b are connected if $\frac{v_b - v_c}{x_b - x_c} > \frac{v_b - v_a}{x_b - x_a}$, where x_a and x_b are the respective positions of V_a and V_b in the time series, v_a and v_b are their respective values and V_c is a third vertex with position x_c between them, and value v_c.

The time series of pause times were converted to visibility graphs and for each one we calculated the following metrics: (1) **Modularity** (Q) [Newman 2004], which is given by $Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$ (we used the Louvain method [Blondel et al. 2008] to maximise the modularity); (2) **Number of communities** (N_c), found with the Louvain method [Blondel et al. 2008]; (3) **Average degree** ($\langle k \rangle$), given by $\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i$, where k_i is the degree of vertex i, and n is the number of vertices in the network; (4) **Density** (Δ), given by $\Delta = \frac{2m}{n(n-1)}$, where m is the number of edges and n is the number of vertices in the network; and (5) **Clustering coefficient** ($Clust$). Considering Γ_i the subgraph formed by the neighbors of vertex i and their connecting edges, n_{Γ_i} the number of vertices in this subgraph (i.e., the number of neighbors of i), m_{Γ_i} the number of edges connecting them, and n the number of vertices in the network, $Clust$ is given by $Clust = \frac{1}{n} \sum_{i=1}^{n} \frac{2m_{\Gamma_i}}{n_{\Gamma_i}(n_{\Gamma_i} - 1)}$.

3.2. Affinity Networks

We represent affinity networks as weighted graphs in which vertices correspond to entities and edges indicate their similarity. In this study, the entities are writers and their similarity is based on pausing behavior during a writing task. We constructed an affinity network of writers with the following steps. Each vertex, representing a writer, was associated with a single chromosome with five genes, which are the metrics obtained from the visibility graphs (Q, N_c, $\langle k \rangle$, Δ and $Clust$) rounded to two significant figures. Two vertices in the network were connected if they shared at least one common gene.

4. Results and Discussion

As shown in Figure 1, the affinity network of the participants consisted of three communities that indicate different pausing behaviors. The two largest communities have unequal

proportions of tutors and students: tutors make up 67% of the community shown in green and only 31% of the other, shown in red. However, this difference is not statistically significant according to a Chi squared test with $\alpha = .5$ ($X^2(1, N = 28) = 3.59, p = .058$).

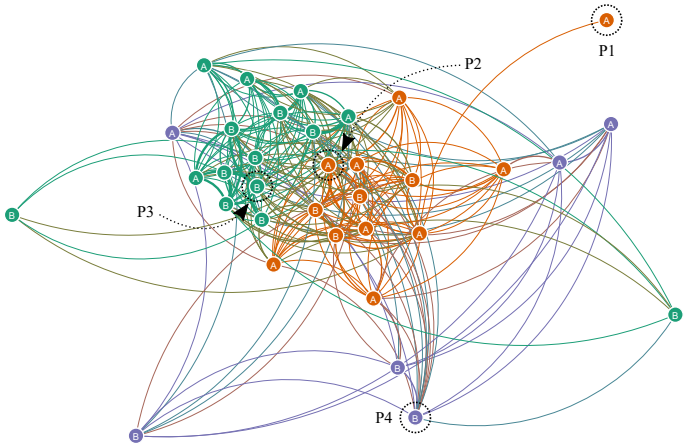

Figure 1. Affinity network of writers who participated in this study. The vertices with label 'A' correspond to the students and those with label 'B' correspond to the tutors. The colors indicate the communities found. P1 to P4 were selected to further analysis.

From the degree distribution of the network, we identified one outlier vertex with degree 1. This vertex, labeled P1 in Figure 2, was selected for further analysis along with the vertices with the highest degree in each community: P2 (the highest-degree vertex in the network), P3 and P4. Table 1 shows their chromosome values.

Table 1. Chromosome values for participants P1, P2, P3 and P4

Participant	Group	Q	N_c	$\langle k \rangle$	Δ	$Clust$
$P1$	A	0.71	12	15.0	0.0087	0.85
$P2$	A	0.85	18	10.0	0.0032	0.82
$P3$	B	0.88	16	11.0	0.0032	0.83
$P4$	B	0.87	19	10.0	0.0024	0.81

Figure 2 illustrates the time series and visibility graphs used to compose the chromosomes values in Table 1. Considering the results in [Valenzuela and Castillo 2022], we considered the beginning of the writing process as comprised of events with normalized event index in the interval $\left[0, \frac{1}{3}\right)$, the middle in the interval $\left[\frac{1}{3}, \frac{2}{3}\right)$ and the end in the interval $\left[\frac{2}{3}, 1\right]$. Participant P1's time series has shorter pauses in the beginning and middle segments and one long pause in the end segment, resulting in a denser visibility graph with fewer communities. Participant P2's time series, on the other hand, has longer pauses in the beginning and middle segments and shorter pauses in the end segment. This suggests, for example, that P1 revised only at the end of the writing process while P2 revised at earlier stages. Both P3 and P4's time series have longer pauses in all three segments, but P3's data has more pauses with normalized duration greater than .25, for example. This difference may reflect different revision strategies as well as sentence planning or lexical access processes.

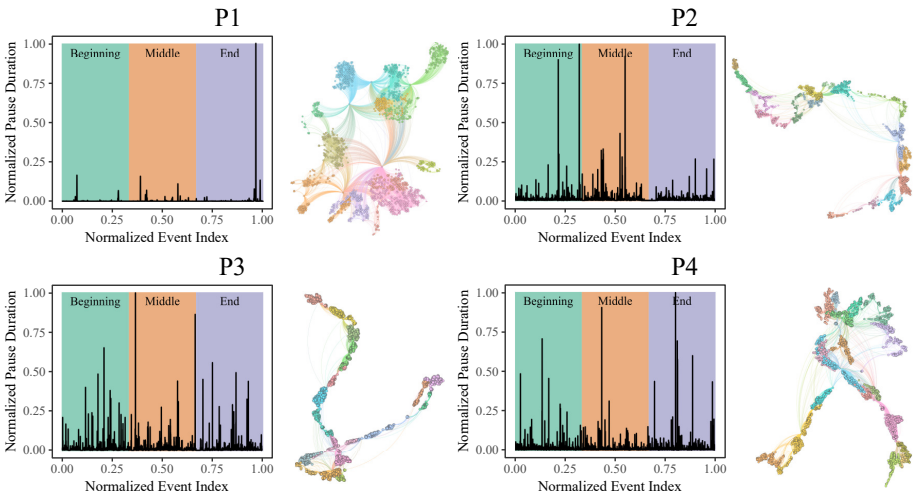

Figure 2. Pausing behavior of four writers represented by time series and visibility graphs.

5. Final Remarks

This study demonstrated how visibility graphs can capture differences in pausing behavior among writers and how affinity networks can indicate differences in writing proficiency based on pausing behavior. The results suggest that the metrics from visibility graphs may be effective in distinguishing different writers. Moreover, the communities in the affinity network implied that tutors and students had different pausing patterns that may relate to their writing proficiency, but this difference was not statistically significant and requires further investigation. This study was limited to pausing behavior and future studies could include more genes and chromosomes with additional writer characteristics to create more comprehensive writing profiles. Additionally, this study considered a single writing task with a particular group of participants. The results must be replicated with different writing tasks and different participants in order to be validated appropriately. Also, future studies could analyse writing process and product data together to better examine the link between pausing behavior and writing proficiency.

6. Acknowledgments

The second author acknowledges the support of the Productivity in Research Scholarship - CNPq 311422/2019-5.

References

Alves, R. A., Castro, S. L., and Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology*, 43(6):969–979.

Aris, A., Shneiderman, B., Plaisant, C., Shmueli, G., and Jank, W. (2005). Representing Unevenly-Spaced Time Series Data for Visualization and Interactive Exploration. In Costabile, M. F. and Paternò, F., editors, *Human-Computer Interaction - INTERACT 2005*, volume 3585 of *Lecture Notes in Computer Science*, pages 835–846. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

Conijn, R., Roeser, J., and van Zaanen, M. (2019). Understanding the keystroke log: the effect of writing task on keystroke features. *Reading and Writing*, 32(9):2353–2374.

Gross, J. L., Yellen, J., and Anderson, M. (2019). *Graph Theory and its applications*. Textbooks in Mathematics. CRC Press, Boca Raton, 3 edition.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., and Bennett, R. E. (2018). Modeling Basic Writing Processes From Keystroke Logs. *Journal of Educational Measurement*, 55(2):194–216.

Lacasa, L., Luque, B., Ballesteros, F., Luque, J., and Nuño, J. C. (2008). From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences of the United States of America*, 105(13):4972–4975.

Leijten, M. and Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, 30(3):358–392.

Medimorec, S. and Risko, E. F. (2017). Pauses in written composition: on the importance of where writers pause. *Reading and Writing*, 30(6):1267–1285.

Melo, D. d. F. P., Fadigas, I. d. S., and de Barros Pereira, H. B. (2017). Categorisation of polyphonic musical signals by using modularity community detection in audio-associated visibility network. *Applied Network Science*, 2(1):1–15.

Monteiro, R. L., Fontoura, J. R., Carneiro, T. K., Moret, M. A., and Pereira, H. B. (2014). Evolution based on chromosome affinity from a network perspective. *Physica A: Statistical Mechanics and its Applications*, 403:276–283.

Monteiro, R. L. S., Carneiro, T. K. G., Andrade, L. P. C. S., Fadigas, I. d. S., and Pereira, H. B. d. B. (2015). An affinity-based evolutionary model of the diffusion of knowledge. *Obrta Digital*, 1(9):44–57.

Newman, M. E. (2004). Analysis of weighted networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 70(5):9.

Shen, C. and Chen, N. (2021). Profiling the pausing behaviour of EFL learners in real-time computer-aided writing: a multi-method case study. *Asian-Pacific Journal of Second and Foreign Language Education*, 6(1):1–26.

Valenzuela, Á. and Castillo, R. D. (2022). The effect of communicative purpose and reading medium on pauses during different phases of the textualization process. *Reading and Writing*, 36(4):881–908.