Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education. New Zealand: Asia-Pacific Society for Computers in Education

Impact of Both Prior Knowledge and Acquaintanceship on Collaboration and Performance: A Pair Program Tracing and Debugging Eye-Tracking Experiment

Maureen VILLAMOR[a,b*] & Ma. Mercedes RODRIGO[a]
[a]*Ateneo de Manila University, Philippines*
[b]*University of Southeastern Philippines, Davao City, Philippines*
*maui@usep.edu.ph

Abstract: We compared the collaboration of pairs whose composition was based on both prior knowledge and degree of acquaintanceship as they traced and debugged fragments of code. We performed a cross-recurrence quantification analysis (CRQA) to build cross-recurrence plots using the eye tracking data and computed for the CRQA metrics, such as recurrence rate (RR), determinism (DET), entropy (ENTR), and laminarity (LAM) using the CRP toolbox for MATLAB. Findings revealed that high prior knowledge pairs who were poorly acquainted (BH/PA) performed better among categories despite having collaborated the least. This confirmed the findings of prior studies that skilled strangers perform best. Mixed prior knowledge pairs who were highly acquainted (M/HA) collaborated the most but their familiarity did not translate to better performance. The results of this study could contribute to the learning sciences and pedagogy. If we know what makes collaboration successful as measured through their performance, we can design interventions that could facilitate the process of creating programming pairs who can collaborate and perform better.

Keywords: Eye-tracking, collaboration, cross-recurrence quantification analysis

1. Introduction

Prior knowledge from previous courses can influence student achievement (Hailikari, Katajavuori and Lindblom-Ylanne, 2008) so it is a known fact that students with high prior knowledge outperform students with low prior knowledge in problem solving tasks. It can also be assumed that when students with high prior knowledge are paired or grouped together in collaborative learning situations, they will perform better than pairs or groups consisting of low prior knowledge students. However, is this always the case if we put friendship into the picture? Previous studies had looked into the impact of friendship on collaborative and competitive performance. They tested whether the quality of social interaction between friends as opposed to non-friends influence collaborative success. Findings have shown that groups composed of friends may perform better as a direct result of their collaboration history (Rittenbruch and McEwan, 2009) or may increase commitment to goals of the group (Jehn and Shah, 1997), which could contribute to more successful collaboration.

However, there is also evidence that friendships could diminish performance because friends have a tendency to focus more on socializing than the group task (Shah and Jehn, 1993). Friendship may even be immaterial for effective collaboration. Groups composed of skilled strangers will perform best because highly skilled individuals may already know from experience how to work well with other experts, and hence, are easily adaptable to the actions of their group mates (Shah and Jehn, 1993).

Pair programming is a collaborative work arrangement where two programmers execute different programming activities together. It has become a well-known pedagogical practice for teaching introductory programming as it has shown that students who are involved in pair programming produce better quality of code, are more confident with their solutions, and are more likely to succeed and persevere in their programming courses compared to solo programmers (Murphy et al., 2010).

In recent years, dual eye-tracking in the context of pair programming has been explored to study joint attention in collaborative learning situations. Two eye-trackers, for instance, can be synchronized for studying the gaze of two individuals collaborating to solve a problem and for understanding how gaze and speech are coupled (Pietinen et al., 2008).

Cross-recurrence quantification analysis (Marwan and Kurths, 2002) is used to quantify how frequently two systems exhibit similar patterns of change in time. It produces a cross-recurrence plot (CRP), which has been used to analyze the coordination of gaze patterns between individuals and can be used to measure how much and when two subjects look at the same spot (Nüssli, 2011).

This paper used CRQA to characterize collaboration of pairs of novice programmers in the act of tracing fragments of code and debugging in a remote pair programming setup. Specifically, it attempted to answer the following research questions: What characterizes collaboration between pairs of participants who (a) both have high prior knowledge that are highly or poorly acquainted, (b) both have low prior knowledge that are highly or poorly acquainted, and (c) have a high and low prior knowledge that are highly or poorly acquainted?

Our prior work focused on characterizing collaboration of pairs based on prior knowledge and acquaintanceship separately using CRQA and seeking for existing patterns. This paper attempts to substantiate the findings from our previous studies by investigating the impact of both prior knowledge and acquaintanceship on the collaboration and performance of the pairs of novice programmers while they traced and debugged codes.

2. Methods

The study was conducted in three private universities in the Philippines. Students aged 18-23 years old who were in their 2nd year to 4th year level in college and had taken the college-level fundamental programming course were recruited to participate in this study. Twenty four (24) pairs of participants were asked to read fragments of code with known bugs and then identify the location of each bug. For a detailed description of the structure of the study, see Villamor and Rodrigo (2017).

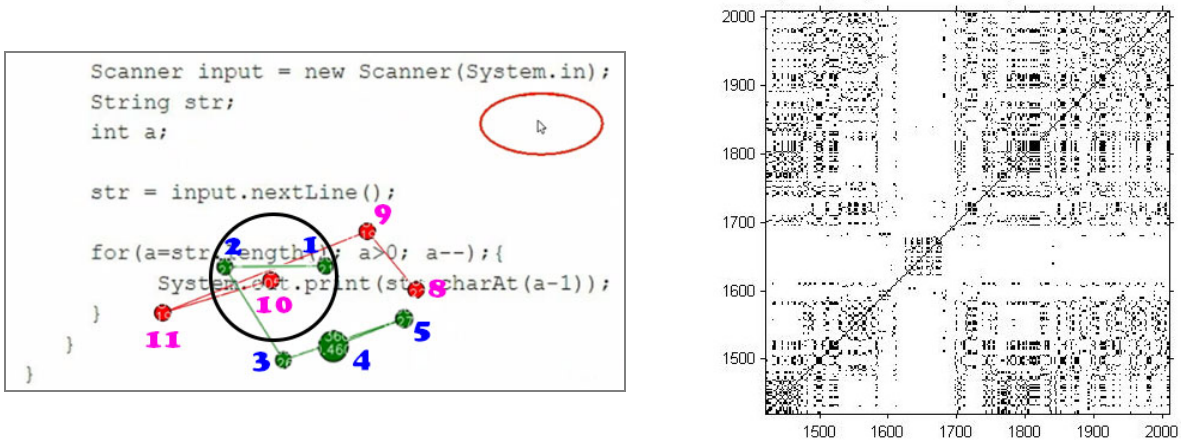
To conduct a cross-recurrence analysis, an N x N matrix called cross-recurrence plot (CRP) is built, which is essentially a representation of the time coupling between two time series. The horizontal axis represents time for the first collaborator (*C1*) and the vertical axis represents time for the second collaborator (*C2*). Recurrence occurs when the distance between the fixations of the two collaborators has to be lower than a given radius. In Figure 1.a, let us assume that the numbered red and green dots are from the fixation sequences of *C1* and *C2*, respectively. Given a certain radius bounded by the black bordered circle shown in the figure, fixation pairs (1, 10) and (2, 10) are considered recurrent since their distances fall within a certain radius.

If fixations i and j are recurrent, they are represented as a black point (pixel) in the plot (see Figure 1.b). Hence, a point in the plot indicates that the states of the two systems for their respective times are recurrent. If two collaborators uninterruptedly looked at two different spots on the screen for the entire interaction, the resulting CRP would be completely blank (white space in Figure 1.b). If the two collaborators looked at the same spot on the screen continuously, the plot would show only a dark line on the diagonal. Points exactly on the diagonal of the plot correspond to synchronous recurrence, such as, collaborators look at the same target at exactly the same time.

CRQA defines several measures that can be assessed along the diagonal and vertical dimensions. For the diagonal dimension, we have: recurrence rate (RR), determinism (DET), average (L) and maximal length (LMAX) of diagonal structures, and entropy (ENTR). For the vertical dimension, we have: laminarity (LAM) and trapping time (TT). The definitions of these metrics can be found in Marwan and Kurths (2002).

The number of fixations per slide that contained the actual program were segregated and saved on separate files. A CRP was constructed for each pair for every program using the CRP toolbox for MATLAB (Marwan and Kurths, 2002), and CRQA was performed to get the RR, DET, ENTR, and LAM for each of the 12 programs. The CRQA metrics L, LMAX, and TT were not included due to the page limit restriction. For this data, no further embedding was done (Iwanski and Bradley, 1998) and the delay was also set equal to one since no points were time delayed (Webber and Zbilut, 2005). The radius was set to 5% of the maximal phase space diameter (Schinkel, Dimigen and Marwan, 2008).

Pearson's correlation was performed to determine relationships between the categories' performance task score and CRQA metrics based on both prior knowledge and acquaintanceship. ANOVA was performed twice: (1) comparing the CRQA metric means per program, and (2) comparing the CRQA metric means of the overall task (12 programs). Tukey post hoc tests at 0.05 level of significance were performed to determine which relationships were significant.



(a) (b)

Figure 1. (a) An Illustration of Recurrent Fixations and (b) An Example of a Cross-Recurrence Plot.

3. Results and Discussion

A student has high prior knowledge if his/her program comprehension test result was equal to or greater than the median score; otherwise, the student has low prior knowledge. The post-test pair evaluation was used to assess the degree of acquaintanceship of the pairs. Pairs were highly acquainted if their average post-test survey rating is within 3.6 to 5; otherwise, they were poorly acquainted.

Of the 23 pairs (one pair was discarded), there were six (6) both high prior knowledge pairs who were highly acquainted (BH/HA), two (2) both high prior knowledge students who were poorly acquainted (BH/PA), eight (8) were mixed prior knowledge pairs who were highly acquainted (M/HA), and three (3) were mixed prior knowledge pairs who were poorly acquainted (M/PA). The remaining four (4) pairs who both had low prior knowledge students were highly acquainted so they were not included in this analysis. The CRQA metrics for every program in these categories were averaged separately to get the aggregated CRQA metrics. Incidences of high and low values of the CRQA metrics were examined to find the differences among the categories. A value is high if it is equal to or greater than the mean plus one standard deviation; and low, otherwise. Table 1 shows the descriptive values of all the aggregated CRQA metrics per program and the ANOVA results per program and overall task.

3.1 Recurrence Rate (RR)

Half of the RR's in the BH/PA pairs were low and the rest were average. The M/HA pairs had RR's ranging from average to high. Both the BH/HA and M/PA pairs had one high RR each and the rest were average (see Table 1 for high and low RR). This suggests that the BH/PA pairs collaborated the least while the M/HA pairs collaborated the most. The extent of collaboration of the BH/HA and M/PA pairs in terms of RR were comparable.

One possible explanation for this is because it might be difficult for pairs with both high prior knowledge students who are not familiar with each other to open up for collaboration due to differences in ideas or plainly because secure people are already confident being on their own so they do not really feel the need to collaborate with others. We speculate that this might possibly be the reason for the poor RR result of the BH/PA pairs.

Sharma, Jermman and Nüssli (2012) describe convergent and divergent phases of collaboration. A convergent episode is one in which collaborators look at the same part of the program in a manner that is reflected by fixations less than a given threshold. A divergent episode is one in which collaborators look at the different parts of the program, which happens when participants try to build

their own understanding of the program. This implies that participants are attempting to build their own understanding of the program. We hypothesize that this is probably what happened when the pairs tried to work independently. Their divergent episodes caused their RR's to drop. However, despite of lower RR turnout, the BH/PA pairs performed the best among categories confirming the findings of prior studies that skilled strangers perform best.

Table 1: Descriptive values and ANOVA results of each CRQA metric

CRQA Metric	Mean	SD	Min	Max	Low <=	High >=	ANOVA Per Program		ANOVA Overall Task	
							F(3,44)	p-value	F(3,15)	p-value
RR	0.09	0.03	0.03	0.18	0.06	0.12	9.269	0.000	4.700	0.017
DET	0.32	0.08	0.14	0.48	0.24	0.39	14.659	0.000	5.500	0.009
ENTR	0.56	0.18	0.17	0.87	0.38	0.74	8.246	0.000	4.636	0.017
LAM	0.36	0.12	0.10	0.59	0.25	0.48	12.046	0.000	5.034	0.013

As for the M/HA pairs, since one would probably try to dominate, possibly the high prior knowledge student, while the other submits, collaboration is expected to be smooth, most especially if the two are already familiar with each other to begin with. This could be the reason for the M/HA pairs' higher incidences of RR. The BH/HA and M/PA pairs' RR's and average performance scores were comparable. Their RR's and average performance scores were in the middle.

A significant high negative relationship was only found between the M/HA pairs' average performance score and RR ($r = -0.731$, $p = 0.039$) indicating that as their scores increased, their RR's significantly decreased. The M/HA pairs had the lowest average performance score but with the highest average RR among the categories. This implies that their being familiar and comfortable with their partners and despite of having collaborated better did not warrant a better performance. Perhaps they just spent a great deal amount of time chatting or socializing. ANOVA and post hoc test results revealed that the RR's of BH/PA and M/HA as well as M/HA and M/PA were statistically significant while others were not (see Table 1).

3.2 Determinism (DET)

The DET values of the BH/PA pairs were from low to average, the M/HA pairs had average to high DET, the BH/HA pairs had one high DET and majority were above the mean, the M/PA had one high DET, two low, and the rest were average (see Table 1 for high and low DET).

The BH/PA pairs had the least shared identical scanpaths and the M/HA pairs shared the most identical scanpaths. Possible explanation for this is the same in RR. High prior knowledge pairs might not feel the need to converge frequently but they still performed better nonetheless. It is assumed for pairs who are highly acquainted where one is a leader and the other a follower to have more matching scanpaths because most likely the follower would follow what the leader is aiming at on the screen. This could be the case of the M/HA pairs having better DET results compared to others, but their high DET turnout was the exact opposite of their performance. The M/HA's average performance score was the lowest, possibly indicating that the pairs had engaged in more off-task behaviors. It could also be that the M/HA pairs had disagreed more frequently.

A significant negative high relationship only existed between the M/HA pairs' performance score and DET ($r = -0.789$, $p = 0.02$), denoting that as more bugs were found, their matching scanpaths dropped. This implies that the M/HA pairs tried to work on their own but their scores did not improve. This could also be the case of more divergent episodes. The BH/PA pairs had the highest average score but with the lowest DET, possibly conveying that the BH/PA pairs collaborated less but their scores were still the highest among the categories. Their divergent episodes trying to build program understanding on their own could have contributed to their better performance.

The BH/HA pairs did not perform better as the BH/PA pairs. Though their average RR was the second highest among the categories, their average score was in the bottom two. This indicates that their being familiar and at ease with each other could have affected their performance. They might have spent more time on off-task situations or chatting rather than on the main task. The M/PA pairs average performance score was in the top two but their average RR was in the bottom two, signifying that their

being unfamiliar with each other caused them to engage in more divergent episodes and in doing so, their performance scores increased. ANOVA and post hoc tests showed that BH/HA and BH/PA, BH/PA and M/HA were statistically significant while others were not (see Table 1).

3.3 Entropy (ENTR)

The BH/PA pairs had the least complicated scanpaths as half of its ENTR values were low, whereas the M/HA pairs had the most complicated scanpaths because of its ENTR values which ranged from average to high. Refer to Table 1 for high and low ENTR value. The BH/PA pairs had the highest average score but with the lowest average ENTR among the categories. This could probably mean that their predictable but proven and tested debugging strategies, which implied more consistent or steadier scanpaths, could have contributed to their high performance scores. Their being unfamiliar was irrelevant since they still performed the best among the categories.

The M/HA pairs had the lowest average score but with the highest average ENTR. This could mean that they possibly did shotgun debugging which made their scanpaths the most complex since they just tended to look anywhere on the screen and this made their average performance score the lowest. Their being highly acquainted causing them to spend more on off-task behaviors could have contributed to their poor performance scores.

There were no significant correlations between the performance score and ENTR in all categories. However, ANOVA per program and post hoc tests showed significant differences between BH/HA and BH/PA, BH/PA and M/HA, and M/PA and M/HA. ANOVA overall revealed significant differences only between the BH/PA and M/HA pairs (see Table 1).

3.4 Laminarity (LAM)

Half of the BH/PA pairs' LAM values were low and the rest were average. The BH/HA, M/HA, and M/PA pairs only had average to high LAM with the M/HA pairs having more high LAM compared to other categories. See Table 1 for high and low LAM values.

This could mean that the BH/PA pairs did not spend as much time as the other categories on certain programs or regions of code, and hence, they transitioned faster to other slides and were faster in terms of debugging. This is an indication of better program comprehension. This was confirmed in their average slide switches between the program specification and actual program and also their average fixation points and fixation duration, which were the lowest among the categories. Their average performance score was also the highest among the categories. Their degree of acquaintanceship was irrelevant in relationship to their performance scores.

The M/HA pairs tended to spend the most time on certain regions of the code but this did not translate to better scores since their average score was the lowest. It is possible that in those moments, they were just chatting and were not really concerned with finding the bugs in the programs. Their being acquainted might have caused their poor performance scores possibly because of too much spending in off-task behaviors and socializing. The BH/HA pairs despite being both skilled did not perform well as the BH/PA pairs. Their familiarity might have caused their weak performance focusing more on off-task behaviors and socialization. Their highest average fixation points and fixation duration as well as their average high LAM next to the M/HA pairs indicated that they might have socialized more.

The M/PA pairs performed well next to the BH/PA pairs and their average LAM value was the second lowest. This indicates that the M/PA pairs preferred to engage in more divergent episodes because of their unfamiliarity with their partners. In doing so, their performance scores improved. ANOVA and post hoc tests showed significant differences between BH/HA and BH/PA, BH/PA and M/HA, and M/PA and M/HA (see Table 1).

4. Summary and Conclusion

This paper compared the collaboration and performance of pairs consisting of two individuals who may have different or same level of prior knowledge and who may be highly or poorly acquainted given the task of program tracing and debugging. High-performing pairs who are poorly acquainted (BH/PA)

collaborated the least (low RR and DET) but performed the best. This confirmed the findings of prior studies that skilled strangers perform best. Mixed prior knowledge pairs who are highly acquainted (M/HA) collaborated the most (high RR and DET) but they performed weakly implying that their familiarity and being comfortable with each other did not warrant a better performance.

The BH/PA pairs' least complicated scanpaths (low ENTR) translated to better performance. The M/HA pairs' more complicated scanpaths (high ENTR) possibly indicated the use of trial-and-error debugging strategies, which resulted to a weak performance. The BH/PA pairs spent the least amount of time on certain regions of code (low LAM). This indicated better comprehension and better performance. The M/HA pairs had the highest LAM but with the lowest average performance score. High performing pairs who are highly acquainted (BH/HA), despite of being both skilled, did not perform well as the BH/PA pairs. Mixed prior knowledge pairs who are poorly acquainted (M/PA) performed well next to the BH/PA pairs but their LAM values were among the lowest suggesting that they preferred to engage more in divergent episodes which translated to better scores.

These preliminary findings confirmed that friendships or strong familiarity with partners could detract from performing well and are linked to reduced productivity because they have a tendency to spend less time focusing on the task and instead spend more time socializing. This also confirmed that friendship is irrelevant to better performance when groups are composed of highly skilled individuals. Further results of this study could help educators in teaching introductory computer programming where attrition rate is known to be high. Collaborative learning tasks such as pair programming could be strengthened by pairing students who would most likely to collaborate the most and perform at its best. Future work will look at their discourse data and triangulate it with the eye-tracking data to further validate the results of this study.

References

Hailikari, T., Katajavuori, N., & Lindblom-Ylanne, S. (2008). The Relevance of Prior Knowledge in Learning and Instructional Design. *American Journal of Pharmaceutical Education*, *72*(5). Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2630138/

Iwanski, J. S., & Bradley, E. (1998). Recurrence plots of experimental data: To embed or not to embed?. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *8*(4), 861-871.

Jehn, K. A., & Shah, P. P. (1997). Interpersonal relationships and task performance: An examination of mediation processes in friendship and acquaintance groups. *Journal of Personality and Social Psychology*, *72*(4), 775.

Marwan, N., & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, *302*(5), 299-307.

Murphy, L., Fitzgerald, S., Hanks, B., & McCauley, R. (2010, August). Pair debugging: a transactive discourse analysis. In *Proceedings of the Sixth international workshop on Computing education research* (pp. 51-58). ACM.

Nüssli, M. A. (2011). *Dual eye-tracking methods for the study of remote collaborative problem solving* (Doctoral dissertation, École Polytechnique Fédérale de Lausanne).

Pietinen, S., Bednarik, R., Glotova, T., Tenhunen, V., & Tukiainen, M. (2008, March). A method to study visual attention aspects of collaboration: eye-tracking pair programmers simultaneously. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 39-42). ACM.

Rittenbruch, M., & McEwan, G. (2009). An historical reflection of awareness in collaboration. In *Awareness Systems* (pp. 3-48). Springer London.

Schinkel, S., Dimigen, O., & Marwan, N. (2008). Selection of recurrence threshold for signal detection. *The European Physical Journal-Special Topics*, *164*(1), 45-53.

Shah, P. P., & Jehn, K. A. (1993). Do friends perform better than acquaintances? The interaction of friendship, conflict, and task. *Group decision and negotiation*, *2*(2), 149-165.

Sharma, K., Jermann, P., Nüssli, M. A., & Dillenbourg, P. (2012). Gaze Evidence for different activities in program understanding. In *24th Annual conference of Psychology of Programming Interest Group* (No. EPFL-CONF-184006).

Villamor, M.M. and Rodrigo, M.M.T. (2017, June). Characterizing Collaboration in the Pair Program Tracing and Debugging Eye-Tracking Experiment: A Preliminary Analysis. In *Proceedings of the 10th International Conference on Educational Data Mining*, (pp. 174-179).

Webber Jr, C. L., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences*, 26-94.