Electronic Supplementary Material for

Orderliness predicts academic performance: Behavioural analysis on campus lifestyle

Yi Cao[1], Jian Gao[1], Defu Lian[2], Zhihai Rong[1], Jiatu Shi[2], Qing Wang[3], Yifan Wu[1], Huaxiu Yao[4], and Tao Zhou[1,2,*]

1. *CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, China*
2. *Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China*
3. *Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, Center for Information in Medicine, University of Electronic Science and Technology of China, Chengdu 611731, China*
4. *College of Information Science and Technology, Pennsylvania State University, PA 16802, USA*

* Corresponding author. Email: zhutou@ustc.edu

September 06, 2018

This PDF file includes:

Supplementary Texts

Supplementary Figures

References

Supplementary Material

S1 Actual entropy and other measures on orderliness

The meaning of *orderliness* is twofold, say timing and order. Firstly, the happening times of the same kind of events should be close to each other, for example, taking breakfast at about 8:00 in the morning is more regular than taking breakfast between 7:00 and 9:00. Secondly, the temporal order of the different events should be regular. For instance, one may go to cafeterias following the order: breakfast→lunch→supper→breakfast→lunch→supper, which is more regular than breakfast→supper→lunch→ supper→breakfast→lunch.

We apply actual entropy[1,2] as it takes into account both ingredients above. Specifically, the actual entropy is defined as:

$$S_{\mathscr{E}} = \left(\frac{1}{n}\sum_{i=1}^{n}\Lambda_i\right)^{-1}\ln n, \tag{S1}$$

where n is the length of the sequence, and Λ_i represents the length of the shortest subsequence starting from i-th position of the binned event sequence $\mathscr{E}$, which has never appeared previously. Note that we set $\Lambda_i = n-i+2$ if such subsequence does not exist[2]. Smaller $S_{\mathscr{E}}$ suggests higher orderliness. For example, there are two consecutive sequences on taking showers with different happening times, {21:05, 21:13, 21:17, 21:28, 21:24, 21:15, 21:12, 21:08, 21:19, 21:03} and {21:05, 21:33, 21:13, 21:48, 21:40, 21:15, 21:42, 21:18, 21:49, 21:53}. The two binned sequences $\mathscr{E}$ are {43, 43, 43, 43, 43, 43, 43, 43, 43, 43} and {43, 44, 43, 44, 44, 43, 44, 43, 44, 44}, and the corresponding length sequences Λ are {1, 2, 3, 4, 5, 6, 5, 4, 3, 2} and {1, 1, 3, 2, 4, 6, 5, 4, 3, 2}, respectively. According to Eq. (S1), the actual entropy values are respectively 0.658 and 0.743, suggesting that the former sequence is more regular than the latter one. Next, we take meals as an example. Assume that two students take breakfast, lunch and supper at 7:30-8:00, 11:30-12:00 and 17:30-18:00, but one has every meal on campus, while the other sometimes has a dinner off-campus or does not take breakfast. The binned sequences $\mathscr{E}$ of the two students could be {16, 24, 36, 16, 24, 36, 16, 24, 36, 16, 24, 36, 16, 24, 36} and {16, 36, 16, 16, 24, 36, 36, 16, 24, 36, 16, 24, 24, 24, 36}, respectively. The corresponding length sequences Λ are {1, 1, 1, 4, 4, 4, 7, 7, 7, 7, 6, 5, 4, 3, 2} and {1, 1, 2, 2, 1, 2, 3, 4, 3, 4, 3, 2, 3, 3, 2}, and the actual entropy values are 0.645 and 1.128, respectively. This shows that the former sequence is more regular than the latter. As the sequences get longer, the differences become more remarkable.

Besides the actual entropy, we come up with a few classic metrics to explain why the temporal order of a behavioural sequence is important and why these metrics are inappropriate to quantify the orderliness. Information entropy[3] is the most frequently used metric to measure the regularity. Mathematically, the information entropy is defined as

$$E = -\sum_{i=1}^{N} p_i \log_2 p_i, \tag{S2}$$

where N is the number of different kinds of elements (here $N = 48$), and p_i denotes the normalized frequency of the i-th element, and thus

$$\sum_{i=1}^{N} p_i = 1. \tag{S3}$$

Larger E means higher orderliness. However, this method cannot distinguish sequences of different temporal order. For example, the above two students with different meal times are assigned exactly the same information entropy ($E = 1.585$), since the probabilities of the normalized appearance frequencies of 16, 24 and 36 are all 1/3.

Analogously, the well-known Simpson index[4] also fails to distinguish the differences when measuring the temporal order. The Simpson index is initially used to measure the diversity of entities when they are classified into different types. Here, we extend it to represent the regularity level of a given behavioural sequence. Formally, the Simpson index is defined as

$$D = \frac{\sum_{i=1}^{N} r_i(r_i-1)}{n(n-1)}, \tag{S4}$$

where r_i is the number of appearances of the i-th element, say

$$\sum_{i=1}^{N} r_i = n. \tag{S5}$$

A sequence has higher orderliness if D is larger. Considering the above two students with different meal times, the Simpson index values based on Eq. (S4) are all the same ($D = 0.286$).

In a word, the above two classic metrics (information entropy and Simpson index) are inappropriate to measure the orderliness since they only consider the number of the events but ignore the temporal order of these events.

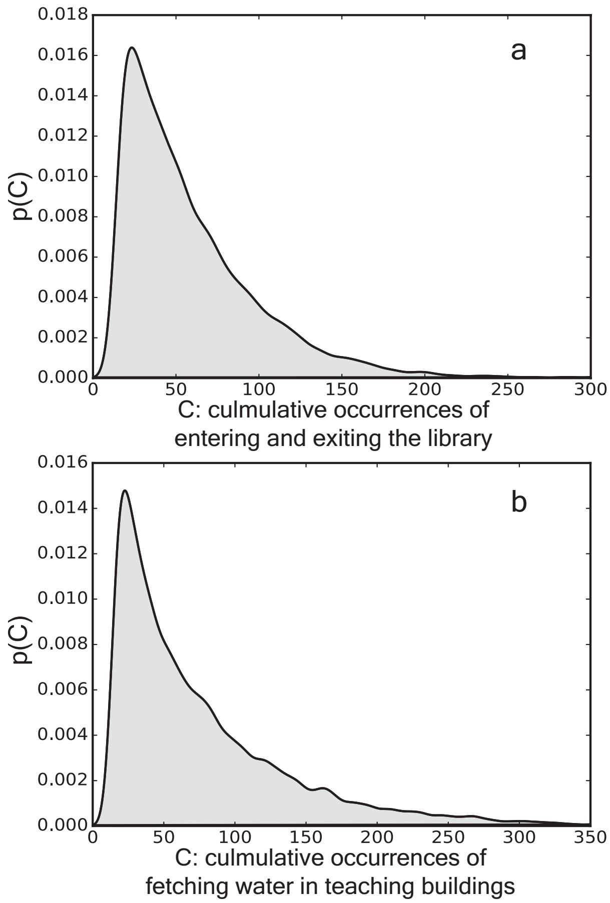

Figure S1. The distributions of two diligence metrics. Distributions $p(C)$ of students for entering/exiting the library (**a**) and fetching water in teaching buildings (**b**). The broad distributions of cumulative occurrences ensure that students with different diligence levels are distinguishable from each other.

S2 Diligence

Diligence is another high-level behavioural character that stands for how people take efforts to strive for achievements. It is considered as a class of high-level features that is directly correlated to academic performance. Considering the difficulties in quantifying diligence due to the lack of ground truth, we roughly estimate diligence based on two behaviors: entering/exiting the library, and fetching water in teaching buildings. Specifically, we use a student's cumulative occurrences of entering/exiting the library and fetching water as a rough estimate of his/her diligence. Normally, borrowing books and self-studying are the most common purposes of a student to go to the library, while attending professional courses is the most common purpose of being at the teaching buildings. However, unlike the library, the teaching buildings have no entry terminals or check-in devices. Hence we use fetching water as a proxy behavior with high frequency for study. For each behavior, we use the cumulative occurrences to estimate the level of diligence. We present the distributions of diligence metrics (Library and Water) in Fig. S1. The two diligence metrics are both broadly distributed, suggesting that the two metrics are good to distinguish students with different levels of striving for achievements. Next, we present the correlation between diligence and GPA in Fig. S2, where both metrics and GPA are regularized by *Z-score*[5]. The Spearman's rank correlation coefficient[6] r is applied to quantify the correlation between regularized diligence metrics and regularized GPAs. As shown in Fig. S2, academic performance is vitally and positively correlated to diligence for both entering/exiting the library ($r = 0.251$; $p < 0.0001$) and fetching water in teaching buildings ($r = 0.291$; $p < 0.0001$).

S3 Relationship between behavioural features and academic performance

We present the scatter chart of relations between regularized behavioural features and regularized GPA in Fig. S3 (a: taking showers; b: having meals; c: entry-exit library; d: fetching water). We found that the four behavioural features are all significantly correlated to GPA with the correlations being about 0.2. The Spearman's rank correlation coefficients for diligence

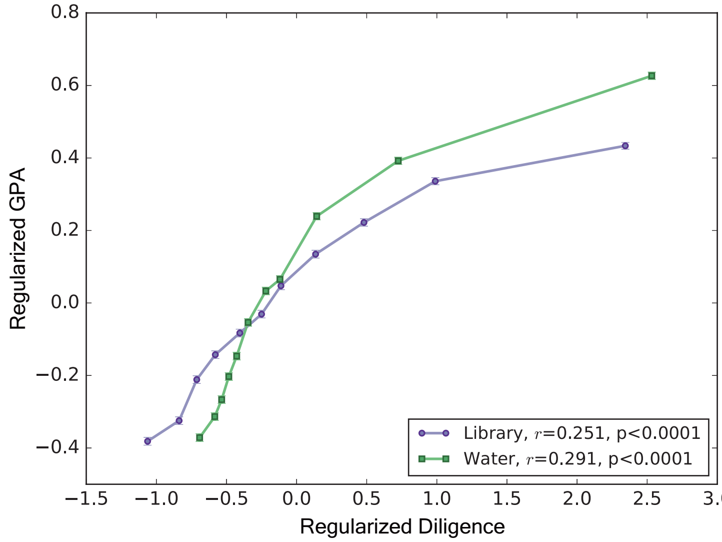

Figure S2. Relationship between diligence and academic performance for entering/exiting the library (purple circles) and fetching water in teaching buildings (green squares). Binned statistics are used to aggregate the data points, where the regularized diligence is divided into 11 bins, each of which contains the same number of data points. The mean value of data points in each bin is presented, with error bar denoting the standard error of the regularized GPA. The Spearman's rank correlation coefficients for GPA-Library ($r = 0.251$; $p < 0.0001$) and GPA-Water ($r = 0.291$; $p < 0.0001$) suggest the statistical significance.

metrics (Library and Water) are stronger than those for orderliness metrics (Shower and Meal), while eyeballing of the data suggests the opposite. We additionally calculated the corresponding Pearson correlation coefficients as an robustness check, and the result showed that correlations for diligence metrics remain stronger than those for orderliness metrics. The visual discrepancy may be because the data points are dispersive. One may notice that the diligence metrics seem to have lower bounds. The reason is that diligence metrics are directly calculated based on the total number of behavioural records. The lower bounds in the number of behavioural records (specifically, the times of entrying/exiting the library and fetching water cannot be negative) lead to the lower bounds of diligence metrics after regularized by *Z-score*.

S4 Inter and intra correlations between behavioural features

Figure S4 reports the scatter chart of the inter correlations for the four orderliness-diligence feature pairs. Results indicate that there is no significant correlation between orderliness and diligence. In contrast, the intra correlations between the two orderliness features and between the two diligence features are all positive and significant as shown in Fig. S5. These correlations suggest the robustness of the indices for orderliness and diligence. While eyeballing of the correlation between orderliness (Meal) and orderliness (Shower) looks much stronger than the correlation between diligence (Water) and diligence (Library), the Spearman's rank correlation coefficients show the opposite. The visual discrepancy may due to the dispersity of the data points. Finally, we summarize these correlations in Fig. S6, which clearly indicates that the intra correlations are all positive and significant, while the inter correlations are all close to 0.

References

1. Kontoyiannis, I., Algoet, P. H., Suhov, Y. M. & Wyner, A. J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory* **44**, 1319-1327 (1998).
2. Xu, P., Yin, L., Yue, Z. & Zhou, T. On predictability of time series. arXiv: 1806.03876 (2018).
3. Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379-423 (1948).
4. Simpson, E. H. Measurement of diversity. *Nature* **163**, 688 (1949).
5. Kreyszig, E. Advanced engineering mathematics (John Wiley & Sons, Hoboken, New Jersey) (2010).
6. Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology* **15**, 72-101 (1904).

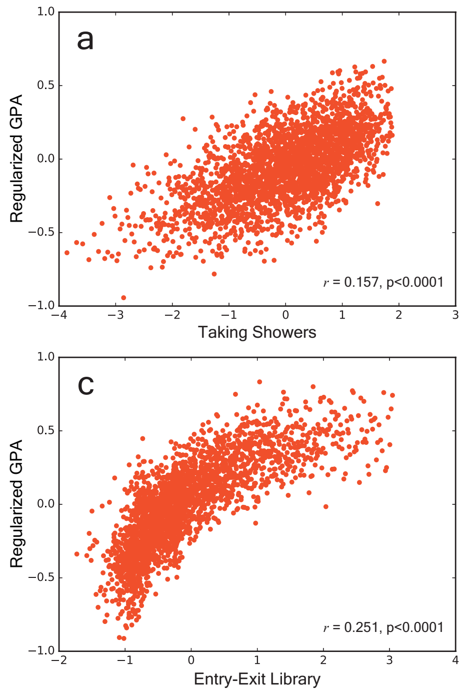

Figure S3. Relations between behavioural features and academic performance. (a) Correlation between regularized orderliness (Shower) and regularized GPA. (b) Correlation between regularized orderliness (Meal) and regularized GPA. (c) Correlation between regularized diligence (Library) and regularized GPA. (d) Correlation between regularized diligence (Water) and regularized GPA.

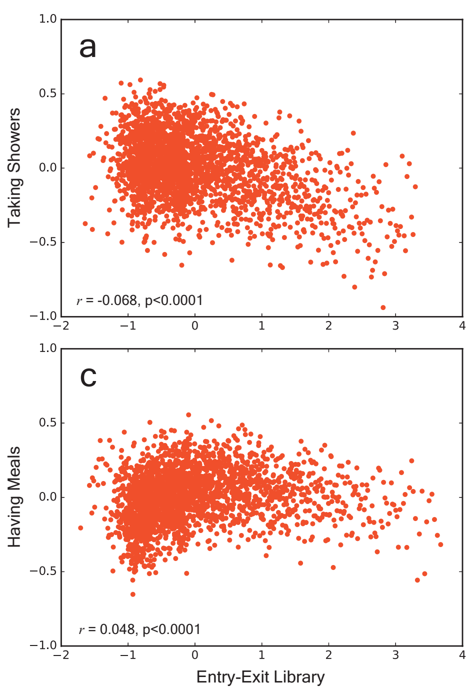

Figure S4. Inter correlations between behavioural features. (a) Correlation between regularized orderliness (Shower) and regularized diligence (Library). (b) Correlation between regularized orderliness (Shower) and regularized diligence (Water). (c) Correlation between regularized orderliness (Meal) and regularized diligence (Library). (d) Correlation between regularized orderliness (Meal) and regularized diligence (Water).

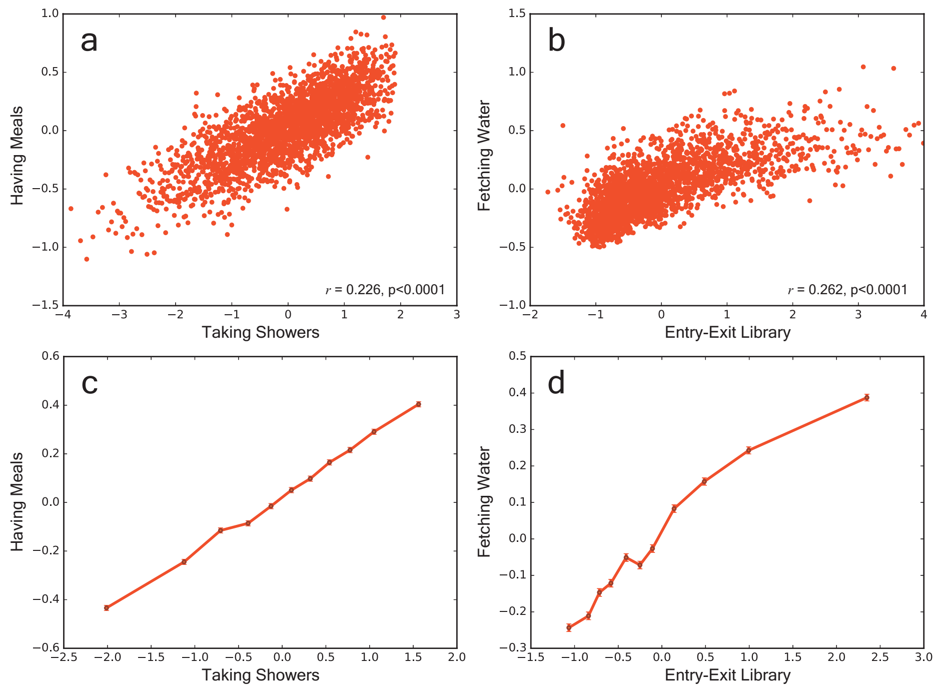

Figure S5. Intra correlations between behavioural features. (a) Correlation between orderliness (Meal) and orderliness (Shower). (b) Correlation between diligence (Water) and diligence (Library). All features are regularized via Z-score. (c) Binned statistics for panel a. (d) Binned statistics for panel b. Error bars correspond to the standard errors.

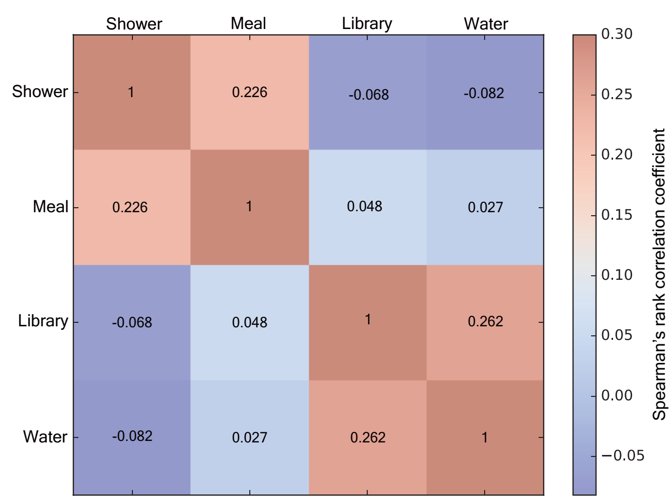

Figure S6. Correlations between each pair of behavioural features. Shower and Meal are the two orderliness features, while Library and Water are the two diligence features. The color in each square denotes the corresponding Spearman's rank correlation coefficient.