sociological science

Supplement to:

Liao, Tim F. 2021. “Using Sequence Analysis to Quantify How Strongly Life Courses Are Linked.” Sociological Science 8: 48-72.

Appendix: Random Seed Sensitivity Analysis

Sequence analysis can be sensitive to various factors, in particular, the choice of a distance measure. Studer and Ritschard (2016) showed that certain distance measures can be more sensitive to differences in timing, others to differences in order, and yet others to differences in duration. My empirical analysis also illustrated the different results based on three of the available distance measures. For a complete discussion of the whole set of distance measures and their sensitivity, see Studer and Ritschard (2016). Once a distance measure is chosen, however, analysis may still be sensitive to random seed choices. We will explore such sensitivity below.

The Issue of Random Seeds

Because the procedure introduced in this paper relies on randomization, it is only natural to question how random a given computation may be. Hidden in any statistical procedure that relies on randomization is the simple fact that random sample selection always starts with a certain random seed, often set to the system time of a computer. Iwaki et al. (2018) found that different random seeds may lead to different results, but how much would the choice of seeds affect results for the current procedure?

To answer that question, I sampled 300 seed numbers without replacement from the set of integers of 1 to 1,000,000. Then I used these seeds to generate the timing focused density plot of the LSOG data and present in Figure A1 these 300 density curves. I generated 300 density curves instead of a much larger number (e.g., 10,000) so that the resulting plots would not be too cluttered while still providing sufficient variation. The number 300 is already much greater than the 10 seeds used for taking the average results of the 10 runs performed by Sun, Liu, and Perc (2019).

In Figure A1, I highlighted three curves in each plot. The curves based on the seeds that generated the lowest and highest mean *U* and *V* value are colored red and purple, respectively, and the density curves based on mean *U* and *V* values are both colored green. All the *U* density curves are in the left plot, and the *V* density curves are in the right plot.

We can draw three general conclusions about the effects of random seeds on the computational results. First, there are differences depending on the random seeds used. There is a clear difference between the minimum and maximum *U* and *V* curves. In either the *U* or *V* curves, the minimum and maximum curves overlap little. Second, the *U* density curves are more clustered together than are their *V* counterparts. The stability of the *U* series in their range of values can be misleading, however, due to the wider range of *U* values. The shrinking effect of the *U* range is about 10 folds, suggested by the average highest density value of 0.1 for the *U* plot, compared to about 1.0 for the *V* plot. Furthermore, the *U* density curves are more normal looking than their *V* counterparts. This is due to the dichotomizing process in the generation of the *V* series. Finally, the mean *U* or *V* curve fits right in the middle of either of the two plots, showing its potential for summarizing all possible random seed computations. Yet, how much does the variation due to different random seeds really affect statistical analysis? I offer an answer to this question in the next subsection.

Assessing Sensitivity of the LSOG Analysis to Seed Choices

To see how sensitive our substantive results are to the choice of random seeds, I conducted a sensitivity analysis of the timing-focused regression model in Table 3 using the results of the 300 random seeds. There are four regression models of *U* and four regression models of *V* on the same independent variables as earlier. The four models use the original U or V as in Table 3 as well as the minimum, mean, and maximum values of U or V. The F-ratios and R^2s from the regressions are reported in Table A1.

I report only the overall modeling fitting statistics because virtually no substantive conclusions would be altered due to changes in coefficient estimates. Two general conclusions can be drawn from a comparison of Tables 3 and A1. First, results based on the dyadic distance U values are extremely consistent across the board (within three digits after the decimal point) compared with those based on the dyadic linkage degree V values. Second, and more important, the random seed variations did not change any of the significance tests in any models of U and V. This is assuring to know.

The sensitivity analysis shows that U is based on mean R_t and is relatively less sensitive to random seed choices; it is therefore preferred in empirical applications. Does that mean dyadic linkage (V) is useless? It has its own usefulness in addition to its ease for interpretation because of its normed range of [0,1]. Because it provides a randomization test for each U value for every observation, we can view V as a significance test or confidence probability for U computations. Figure A2 presents the relationship between U and V for results using the mean values of the runs of the 300 seeds based on Hamming distance.

The plot shows that confidence probability V is a monotonically increasing nonlinear function of dyadic Hamming distance U. It increases fast for the earlier range of U but slows down after about $U = 5$. In the current example, 90% confidence is reached when $U \geq 7$, and 95% confidence is reached when $U \geq 10$, with 45 dyads (or about 10% of the sample) satisfying the 90% condition and 22 dyads (or 5% of the sample) satisfying the 95% condition, as confirmed by the 90% and 95% confidence probabilities provided by the V statistics.

References

Iwaki, Asako, Takahiro Maeda, Nobuyuki Morikawa, Shusuke Takemura, and Hirouiku Fujiwara. 2018. “Effects of Random 3D Upper Crustal Heterogeneity on Long-Period (≥1 s) Ground-Motion Simulations.” *Earth Planets and Space* 70: Article number 156.

Studer, Matthias and Gibert Ritschard. 2016. “What Matters in Differences between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(2): 481–511.

Sun, Xiaojuan, Zhaofan Liu, and Matjaž Perc. 2019. “Effects of Coupling Strength and Network Typology on Signal Detection in Small-World Neural Networks.” *Nonlinear Dynamics* 96: 2145–2155.

Figure A1: Density Plots of Dyadic Distance and Degree of Dyadic Linked Lives (*U* and *V*), LOSG, Using 300 Unique Seeds, State Event Subsequence Method (*N* = 461)

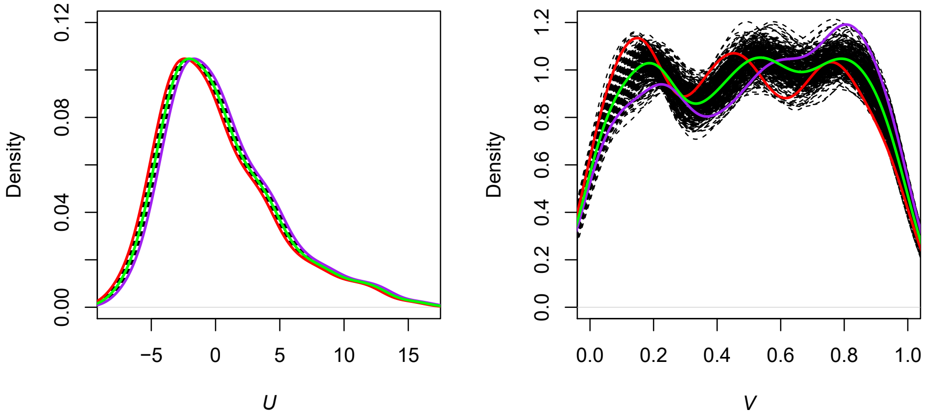

Figure A2: Plot of Dyadic Distance and Dyadic Linkage (Mean U_i by Mean V_i Indicating Conference Probability)

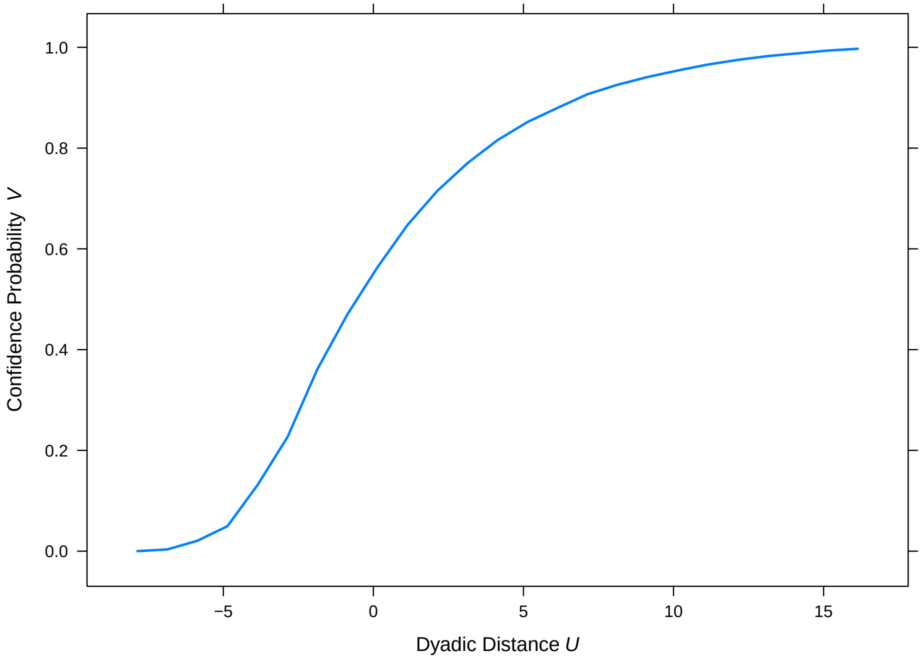

Table A1: Sensitivity Analysis of *U* and *V*: Regression *F* Ratio and R^2 Statistics Correcting for Dyadic Clustering (N = 391)

	Timing *U*	Min. *U*	Mean *U*	Max. *U*
F-statistic	12.304	12.304	12.304	12.304
R^2	0.201	0.201	0.201	0.201
	Timing *V*	Min. *V*	Mean *V*	Max. *V*
F-statistic	16.887	16.745	17.052	17.212
R^2	0.259	0.258	0.263	0.268