Replication files for "Predicting Network Events to Assess Goodness of Fit of Relational Event Models"

Purpose

The folders contain all replication files for the paper:

Brandenberger, Laurence (2018): "Predicting Network Events to Assess Goodness of Fit of Relational Event Models". Political Analysis.

All R scripts are extensively commented to facilitate replication. System requirements and package versions are documented in the following section. 8 R-files need to be run to replicate the analysis including generating the figures in the article and the SI Online. The 8 R-files are stored in the code.zip.

Setup and Runtime

Operating system

The analysis was performed on a macOS (High Sierra 10.13.6; x86_64-apple-darwin15.6.0; 8GB 1600 MHz DDR3, 2.3GHz Intel Core i7) using R version 3.4.3.

R-packages

The following R-packages (and verions) were used for the analysis:

- dplyr_0.7.4
- purrr_0.2.4
- readr_1.1.1
- tidyr_0.7.2
- tibble_1.3.4
- tidyverse_1.1.1
- statnet_2016.9
- statnet.common_4.0.0
- Rcpp_0.12.16
- ROCR_1.0-7
- ggplot2_2.2.1
- texreg_1.36.23
- survival_2.41-3
- rem_1.3.1

Runtime

Runtimes are listed for each R-file individually using the above mentioned operating system.

- 1_PredictArtificalSequence.R → 3.2 hours (fourCycleStat-function from the rem-package is run in parallel using 7 cores)
- 2_IllustrativeExample/0_remModels.R → 3.4 minutes
- 2_IllustrativeExample/1_Create_1000_sequences_fullmodel.R → one simulation round takes 34.1 minutes with the fourCycleStat-cpp-function run in parallel using 7 cores (runtime for 1000 simulation runs = approx. 568 hours).
- 2_IllustrativeExample/2_Create_1000_sequences_redmodel.R → one simulation round takes 13.1 minutes (no parallelization but with cpp-functions; runtime for 1000 simulations runs = approx. 218 hours).

- 2_IllustrativeExample/3_Create_1000_sequences_random.R → one simulation round takes 4.8 minutes (no parallelization; runtime for 1000 simulation runs = approx. 80 hours)
- 2_IllustrativeExample/4_eventTimingGraphs.R → 3.53 hours (no parallelization)
- 2_IllustrativeExample/5_matchingEvents.R → 35.0 minutes (no parallelization)
- 2_IllustrativeExample/6_networkChars.R → 2.1 hours

Contents

The repository is grouped into three main folders:

- code folder = includes all R-files that need to be run to reproduce the presented results (8 in total)
- input folder = includes the data set on which the example in the article builds
- output folder = includes all .RData and .pdf-figures produced using the R-scripts. Due to file size restrictions the output folder is grouped into 28 folders. Details follow:

input folder

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData
   - dt: original data set containing 6701 events (= Discourse Network Analysis statements)
   - dta: original counting process data with first 200 event days cut for the REM (from this, the REM is estimated)
   - dtc: full counting process data

Please note: this RData file (zipped) is password protected because the data is still under review for the substantive paper referenced in this article (Leifeld and Brandenberger 2018). As soon as the substantive paper is accepted for publication, this zip-file will be replaced with a non-password protected RData-file (should be sometime in 2019). If you'd like access to the data in the mean time, please write me an email: laurencebrandenberger@gmail.com or lbrandenberger@ethz.ch. Please indicate your name and affiliation and I will send you the password to unzip the file. Thank you for your patience.

code folder

1_PredictArtificalSequence.R

task

Simulation test for the prediction procedure. Closely documented in the SI Online. Here, output data and figure are generated in this R File.

input

None. Random data is generated in R.

output

1. output/fig3_PredictArtificialSequence.pdf
   - Figure 3 in the Article.
2. output/1_PredictArticifialSequence_DATA.RData
   - All objects generated in the simulation test.
   - dt: simulated event sequence. Variables:
     - sender: sender nodes
     - target: target nodes
     - eventTime: discrete time an event with a sender and target took place
     - eventDummy: 0/1 for events that occurred (1) and null events (0)
     - inertia_hl20: inertia statistic for each event in the sequence (halflife = 20)
     - senderActivity_hl20: activity statistic for each event in the sequence (halflife = 20)

  - targetPopularity_hl20: popularity statistic for each event in the sequence (halflife = 20)
  - fourCycleStat_hl20: four cycle statistic for each event in the sequence (halflife = 20)
  - survpr: Survivor probability for each event (null and true) in the sequence
- dtloop: data set containing REM coefficients from the simulated data (for 2000 strata, measured at 50 strata intervals)
  - eventStrata: ID of strata, measured at 50 intervals (i.e.: 51 = strata 1-51, strata = 100 is 52-100, etc.)
  - coef_inertia: coefficient for the inertia statistic calculated using a REM with strata < eventStrata
  - coef_activity: coefficient for the activity statistic calculated using a REM with strata < eventStrata
  - coef_popularity: coefficient for the popularity statistic calculated using a REM with strata < eventStrata
  - coef_fourcycle: coefficient for the four-cycle statistic calculated using a REM with strata < eventStrata
  - sig_inertia-sig_fourcycle: p-values for the respective REM coefficients.
  - BIC: BIC scores of the model (not informative since the data changes and BIC can only be compared to models with the same data)
  - R2: Mc-Faddens' R-squared for each REM.

2_IllustrativeExample/0_remModels.R

task

Load the pension data used in Leifeld and Brandenberger (2018). It contains the full event sequence as well as the REM-data set (also called counting-process data set) with all the REM-statistics. Run two models: full model contains four-cycle statistics (fit1) and the reduced model (fit2).

input

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData

output

1. output/fig4_RemRegressionplot.pdf
   - Figure 4 in the article. Regression plot for the full and reduced relational event model.
2. output/table2.csv
   - Table 2 in the article. Correlation table for survivor probabilities.

2_IllustrativeExample/1_Create_1000_sequences_fullmodel.R

task

Run 1000 simulated event sequences using the full relational event model (i.e., including four-cycle learning effects). Caution: these simulations are computationally intense. Only run a portion of it to replicate.

intput

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData

output

1. output/2_1_Create_1000_sequences_fullmodeloutput/seedXX_dtsim_eventsonly700-800.RData
   - full event sequence of simulated events between event time 700 and 800, includes simulated true events only.
2. output/2_1_Create_1000_sequences_fullmodeloutput/seedXX_dtsim700-800.RData
   - full event sequence of simulated events between event time 700 and 800, includes simulated true events and simulated null-events.

Please note: Due to file size restriction of 2.5GB (and uploads stalling at 1.5GB), the 1000 simulations are divided into multiple output-folders (output2-output17) on the PA Dataverse.

2_IllustrativeExample/2_Create_1000_sequences_redmodel.R

task

Run 1000 simulated event sequences using the reduced relational event model (i.e., excluding four-cycle learning effects).

intput

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData

output

Same as script "2_IllustrativeExample/1_Create_1000_sequences_fullmodel.R". All simulation runs are recorded in the folder: output/2_2_Create_1000_sequences_redmodel Please note: Due to file size restriction of 2.5GB (and uploads stalling at 1.5GB), the 1000 simulations are divided into multiple output-folders (output18-output25) on the PA Dataverse.

2_IllustrativeExample/3_Create_1000_sequences_random.R

task

1000 random event sequences are generated as a baseline.

intput

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData

output

Same as script "2_IllustrativeExample/1_Create_1000_sequences_fullmodel.R". All simulation runs are recorded in the folder: output/2_3_Create_1000_sequences_random Please note: Due to file size restriction of 2.5GB (and uploads stalling at 1.5GB), the 1000 simulations are divided into three output-folders (output26-output28) on the PA Dataverse.

2_IllustrativeExample/4_eventTimingGraphs.R

task

Calculate the time difference between the simulated true event and the original true event. Generate graphs.

input

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData
2. Simulation data from the following three folders:
    - output/2_1_Create_1000_sequences_fullmodeloutput
    - output/2_2_Create_1000_sequences_redmodel
    - output/2_3_Create_1000_sequences_random

output

1. output/fig5_timeApartFullModel.pdf
    - Figure 5 in the article (top row)
2. output/fig5_timeApartRedModel.pdf
    - Figure 5 in the article (bottom row)
3. output/fig6_timeApartFullRedRandModel.pdf
    - Figure 6 in the article
4. output/si_fig6_timeApartFullRedRandModel_nsim100.pdf

- Part of Figure 6 in the SI Online

5. output/si_fig6_timeApartFullRedRandModel_nsim200.pdf
   - Part of Figure 6 in the SI Online
6. output/si_fig6_timeApartFullRedRandModel_nsim500.pdf
   - Part of Figure 6 in the SI Online
7. output/2_4_eventTimingDataSim.RData
   - aggregated data sets from the 1000 simulations
   - dtsimALL: full data frame of all simulated sequences using the full model. Variables:
     - organization: actor or organization name
     - category: cateogry or policy solution
     - agreement: dummy variable on agreement with policy solution
     - gov: dummy variable on actor type of organization (1 = governmental actor)
     - start.strata: ID of stratum
     - event.seq.ord: ordered event sequence
     - inertia.hl20, activity.hl20, popularity.hl20, fourCycPos, fourCycNeg: REM statistics calculated on the events
     - tie: dependent variable: dummy variable of event occurrence (1 = event occurred)
     - survpr: survivor probability of this event to the next strata
     - simrun: simulation seed used for the prediction
     - timeApart: time difference between the simulated true event and the original true event
   - dtsimredALL: full data frame of all simulated sequences using the reduced model.
     - same as dtsimALL but without fourCycPos and fourCycNeg
   - dtsimrandALL: full data frame of all simulated sequences generated randomly.
     - same as dtsimALL but without any endoegnous network statistics

2_IllustrativeExample/5_matchingEvents.R

task

Calculate the precision of the simulated true events versus the original true events. Predictions are evaluated with a time tolerance of 0, 5 and 10 event days. Additionally, the average tolerance is calculated to correctly predict 50 percent of true events.

input

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData
2. Simulation data from the following three folders:
   - output/2_1_Create_1000_sequences_fullmodeloutput
   - output/2_2_Create_1000_sequences_redmodel
   - output/2_3_Create_1000_sequences_random

output

1. output/fig7_matchingEvents.pdf
   - Figure 7 in the article
2. output/fig8_matchingEvents_50cutoff.pdf
   - Figure 8 in the article
3. output/si_fig1_eventRate_overfulltime.pdf
   - Figure 1 in the SI Online
4. output/si_fig2_eventRate_overSimTime.pdf
   - Figure 2 in the SI Online
5. output/si_fig7_matchingEvents_nsim.pdf
   - Figure 7 in the SI Online
6. output/si_fig8_matchingEvents_50cutoff_nsim.pdf
   - Figure 8 in the SI Online
7. output/2_5_matchingEventsAgg.RData

- outputs of matching events
- dtpercfullALL
  - variable: variable with string indicating what is matched
  - perc: percentage share of correctly matched events
  - simrun: which simulation (full/red/rand + seed) generated the percentage
- dtpercredALL
  - variable: variable with string indicating what is matched
  - perc: percentage share of correctly matched events
  - simrun: which simulation (full/red/rand + seed) generated the percentage
- dtpercrandALL
  - variable: variable with string indicating what is matched
  - perc: percentage share of correctly matched events
  - simrun: which simulation (full/red/rand + seed) generated the percentage
- dtcutoffs
  - cutoffs: value at which temporal tolerance has to be set to get 50% correct matches
  - models: string variable: full model, reduced model, random sequence

8. output/si_table1.csv
   - Table 1 in the SI Online

2_IllustrativeExample/6_networkChars.R

task

Simulated true events are aggregated into network snapshots and compared to the snapshot of the original event sequence within the same prediction period.

input

1. input/PensionData_REM_Leifeld_Brandenberger_2018.RData
2. Simulation data from the following three folders:
   - output/2_1_Create_1000_sequences_fullmodeloutput
   - output/2_2_Create_1000_sequences_redmodel
   - output/2_3_Create_1000_sequences_random

output

1. output/fig9_networkChars_fullModel_nooutlier.pdf
   - Figure 9 in the article (top row)
2. output/fig9_networkChars_redModel_nooutlier.pdf
   - Figure 9 in the article (bottom row)
3. output/si_fig3_networkChars_fullModel.pdf
   - Figure 3 in the SI Online
4. output/si_fig4_networkChars_redModel.pdf
   - Figure 4 in the SI Online
5. output/si_fig5_networkChars_randModel.pdf
   - Figure 5 in the SI Online
6. output/si_fig9_networkChars_fullModel_nooutlier_100nsim.pdf, output/si_fig9_networkChars_fullModel_nooutlier_
   output/si_fig9_networkChars_fullModel_nooutlier_500nsim.pdf
   - All three make up Figure 9 in the SI Online
7. output/2_6_networkChars.RData
   - data frame with network statistics for each simulation round.
   - nwstats
     - char: name of network measure
     - variable: original value from which freq is calculated (freq = variable/sum(variable))
     - type: original sequence vs. simulation
     - typeseed: "original" for the original sequence and the seed number for the simulations

  - variableRange: x-axis
  - freq: y-axis

output folder

1. output/1_PredictArticifialSequence_DATA.RData
   - dt and dtloop data frames
   - generated in the R-script 1_PredictArtificalSequence.R
2. output/2_1_Create_1000_sequences_fullmodeloutput
   - 1000 simulated event sequences are generated from the full model
   - R-script: 2_IllustrativeExample/1_Create_1000_sequences_fullmodel.R
   - 18GB data; subset into output2-17.zip archives
3. output/2_2_Create_1000_sequences_redmodel
   - 1000 simulated event sequences are generated from the reduced model
   - R-script: 2_IllustrativeExample/2_Create_1000_sequences_redmodel.R
   - 9GB data; subset into output18-25.zip archives
4. output/2_3_Create_1000_sequences_random
   - 1000 random event sequences are generated
   - R-script: 2_IllustrativeExample/3_Create_1000_sequences_random.R
   - 4GB data; subset into output26-28.zip archives
5. output/2_4_eventTimingDataSim.RData
   - output of first goodness-of-fit test: temporal prediction error
   - R-script: 2_IllustrativeExample/4_eventTimingGraphs.R
6. output/2_5_matchingEventsAgg.RData
   - output of second goodness-of-fit test: precision with tolerance
   - R-script: 2_IllustrativeExample/5_matchingEvents.R
7. output/2_6_networkChars.RData
   - output of the third goodness-of-fit test: similar network characteristics
   - R-script: 2_IllustrativeExample/6_networkChars.R
8. output/fig3_PredictArtificialSequence.pdf
   - Figure 3 in the article.
   - R-script: 1_PredictArtificalSequence.R
9. output/fig4_RemRegressionplot.pdf
   - Figure 4 in the article
   - R-script: 0_REM_Cpp_Functions.R
10. output/fig5_timeApartFullModel.pdf
    - Figure 5 in the article (top row)
    - R-script: 2_IllustrativeExample/4_eventTimingGraphs.R
11. output/fig5_timeApartRedModel.pdf
    - Figure 5 in the article (bottom row)
    - R-script: 2_IllustrativeExample/4_eventTimingGraphs.R
12. output/fig6_timeApartFullRedRandModel.pdf
    - Figure 6 in the article
    - R-script: 2_IllustrativeExample/4_eventTimingGraphs.R
13. output/fig7_matchingEvents.pdf
    - Figure 7 in the article
    - R-script: 2_IllustrativeExample/5_matchingEvents.R
14. output/fig8_matchingEvents_50cutoff.pdf
    - Figure 8 in the article
    - R-script: 2_IllustrativeExample/5_matchingEvents.R
15. output/fig9_networkChars_fullModel_nooutlier.pdf
    - Figure 9 in the article (top row)
    - R-script: 2_IllustrativeExample/6_networkChars.R

16. output/fig9_networkChars_redModel_nooutlier.pdf
    - Figure 9 in the article (bottom row)
    - R-script: 2_IllustrativeExample/6_networkChars.R
17. output/si_fig1_eventRate_overfulltime.pdf
    - Figure 1 in the SI Online
    - R-script: 2_IllustrativeExample/5_matchingEvents.R
18. output/si_fig2_eventRate_overSimTime.pdf
    - Figure 2 in the SI Online
    - R-script: 2_IllustrativeExample/5_matchingEvents.R
19. output/si_fig3_networkChars_fullModel.pdf
    - Figure 3 in the SI Online
    - R-script: 2_IllustrativeExample/6_networkChars.R
20. output/si_fig4_networkChars_redModel.pdf
    - Figure 4 in the SI Online
    - R-script: 2_IllustrativeExample/6_networkChars.R
21. output/si_fig5_networkChars_randModel.pdf
    - Figure 5 in the SI Online
    - R-script: 2_IllustrativeExample/6_networkChars.R
22. output/si_fig6_timeApartFullRedRandModel_nsim100.pdf, output/si_fig6_timeApartFullRedRandModel_nsim200.pdf
    output/si_fig6_timeApartFullRedRandModel_nsim500.pdf
    - All three figures make up Figure 6 in the SI online
    - R-script: 2_IllustrativeExample/4_eventTimingGraphs.R
23. output/si_fig7_matchingEvents_nsim.pdf
    - Figure 7 in the SI Online
    - R-script: 2_IllustrativeExample/5_matchingEvents.R
24. output/si_fig8_matchingEvents_50cutoff_nsim.pdf
    - Figure 8 in the SI Online
    - R-script: 2_IllustrativeExample/5_matchingEvents.R
25. output/si_fig9_networkChars_fullModel_nooutlier_100nsim.pdf, output/si_fig9_networkChars_fullModel_nooutlier_
    output/si_fig9_networkChars_fullModel_nooutlier_500nsim.pdf
    - All three make up Figure 9 in the SI Online
    - R-script: 2_IllustrativeExample/6_networkChars.R
26. output/table2.csv
    - Table 2 in the article. Correlation table for survivor probabilities.
    - R-script: 2_IllustrativeExample/0_remModels.R
27. output/si_table1.csv
    - Table 1 in the SI Online.
    - R-script: 2_IllustrativeExample/5_matchingEvents.R

Contact

Analysis was performed by Laurence Brandenberger. Please direct any questions to Laurence Brandenberger (laurencebrandenberger@gmail.com).