Appendix A: Detailed results per response variable for *Social influence network simulation design affects behavior of system-level entropy*

Michael J. Garee, Mario Ventresca

1. Appendix A: Detailed results per response variable

In this part, each of the six response variables are analyzed independently of one another with respect to Research Questions 1 and 2. In Section 1.1, we present detailed analysis of a single response variable with accompanying rationale for why the analysis was performed that way and explanation of some of the tests used. Sections 1.2-1.6 repeat that analysis with the remaining response variables but will omit the rationale and explanation elements for brevity; otherwise, all following sections will proceed identically. All readers are encouraged to read Section 1.1 to orient themselves to the processes and figures.

We use a full factorial design for the experiment, so every combination of the factor-levels is equally represented in the data used for the following analysis. Table 1 maps factor-level names to single letters that are used in many of the figures, in order to conserve space.

1.1. Response variable 1 - relative entropy, binning (RE-B)

Relative entropy, binning (RE-B) assigns agent opinion to one of a set of equal-width bins and computes the relative entropy of the resulting distribution $p(x)$ with respect to the uniform distribution $q(x)$, averaging across each agent and each replication to produce the trial-level response. Figure 1 shows the time series of RE-B for each trial and an associated kernel density

Email addresses: m.garee@gmail.com (Michael J. Garee), mventresca@purdue.edu (Mario Ventresca)

Table 1: In the following plots, factor-levels are mapped to single letters for spacing reasons.

	N	structure	influence model	error	activation
a	100	erdos_renyi_random(N)	standard_model	none	synchronous
b	1000	small_world(N, 0.0, 3)	similarity_bias	N(0, 0.05)	uniform
c	10000	small_world(N, 0.0, 10)	attractive_repulsive	N(0, 0.1)	random
d		small_world(N, 0.33, 3)	random_adoption	N(0, 0.2)	
e		small_world(N, 0.33, 10)	nonlinear		
f		small_world(N, 0.66, 3)			
g		small_world(N, 0.66, 10)			
h		scale_free(N, 1)			
i		scale_free(N, 3)			
j		scale_free(N, 5)			

estimate (KDE) for the final time step. This shows some groupings within trials—hinting at possible cluster analysis outcomes—and shows the data set as a whole to not be normally distributed—limiting the relevance of mean values and analysis tools like ANOVA.

Each trial undergoes an initial transient period before becoming monotonic. For some trials, RE-B approaches zero (the minimum bound for relative entropy) but does not reach it within the length of the simulation. The upper extreme appears to correspond to trials where individual agent opinion converged; we use fifty bins for RE-B, so the upper limit for relative entropy with respect to the uniform distribution occurs when its opinion takes on only a single value:

$$D_X(p \parallel q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} = 1 \log_2 \frac{1}{1/50} \approx 5.644. \tag{1}$$

1.1.1. Research Question 1: Which system design factors contribute most to system-level entropy?

For this research question, we explore the one-way sensitivity of each entropy response variable to changes in the levels of individual experimental

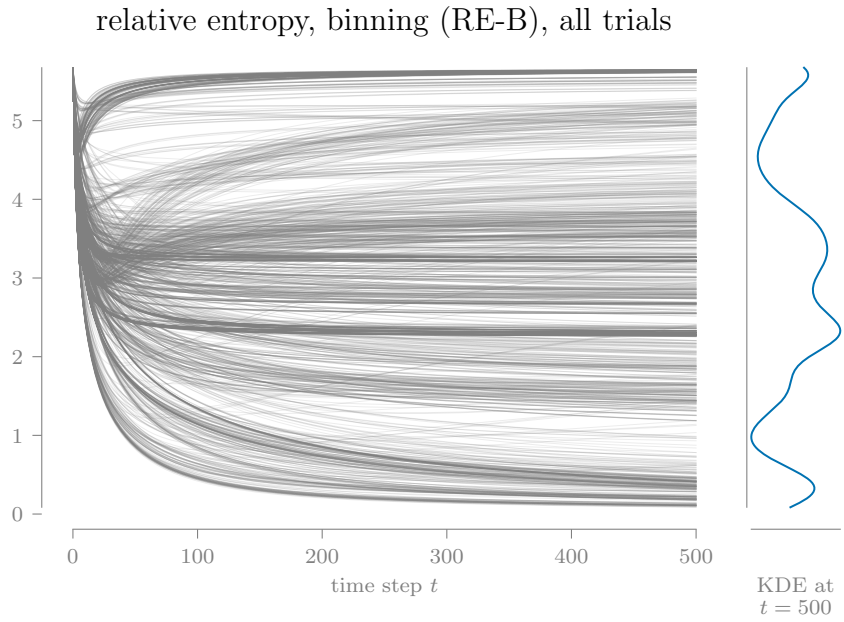

Figure 1: The time series of RE-B values for each trial are plotted on the same axes to reveal visual clusters at the extremes and center of the range. A gaussian kernel density estimate for the RE-B values at $t = 500$ shows the data to be somewhat tri-modal and also reinforces the visual clustering.

design factors. This exploration includes qualitative comparisons of RV distributions when the trial data is grouped by factor-levels and statistical tests for differences between levels. These methods support a subjective evaluation of whether an RV is sensitive to changes in the level of each design factor. In Table 2, we summarize the analysis results for the current response variable for this research question.

Table 2: The findings for Research Question 1 on RE-S are summarized to support the overall evaluation of each experimental design factor (final table row).

	Factor (number of levels)				
	N (3)	structure (10)	influence model (5)	error (4)	activation (3)
i.	*(Main effect plot) What differences are present between the response variable distributions for each level at the final time step?*				
	negligible	2 or 3 patterns	3 or 4 patterns	3 patterns	negligible
ii.	*(Grouped time series) What differences are present between the median response values over the duration of the simulation?*				
	negligible	overall similar shapes; small divide affected by density	3 patterns; nonlinear+standard, similarity+random paired	3 patterns; high+medium variance identical	negligible
iii.	*(K-W test) Does the Kruskal-Wallace test indicate statistical differences in the response variable between each level at the final time step? (i.e., is the p-value <0.05?)*				
	no	yes	yes	yes	no
iv.	*(M-W U test) How many pairs of levels are statistically different (p-value <0.05) according to the Mann-Whitney U test?*				
	0/3	28/45	10/10	5/6	0/3
*	*(Evaluation) Is the response variable sensitive to changes in the level for the factor?*				
	no	yes	yes	yes	no

Analysis for design of experiments (DoE) typically focuses on changes to mean responses. However, the non-normal distribution of the data (Figure 1) shows multiple modes, so the mean value is not salient. Instead, we look for main effects using the full distributions of the data when grouped by factor-level. Figure 2 presents these distributions for RE-B at the final time step, $t = 500$, using a half-violin plot. Differences between distributions among the levels for a single factor qualitatively show the effect each level has on the response. For example, the distributions for population size N are almost identical, so we infer that N is not important (i.e., does not have a significant effect on the response variable), at least over the range of levels used in the

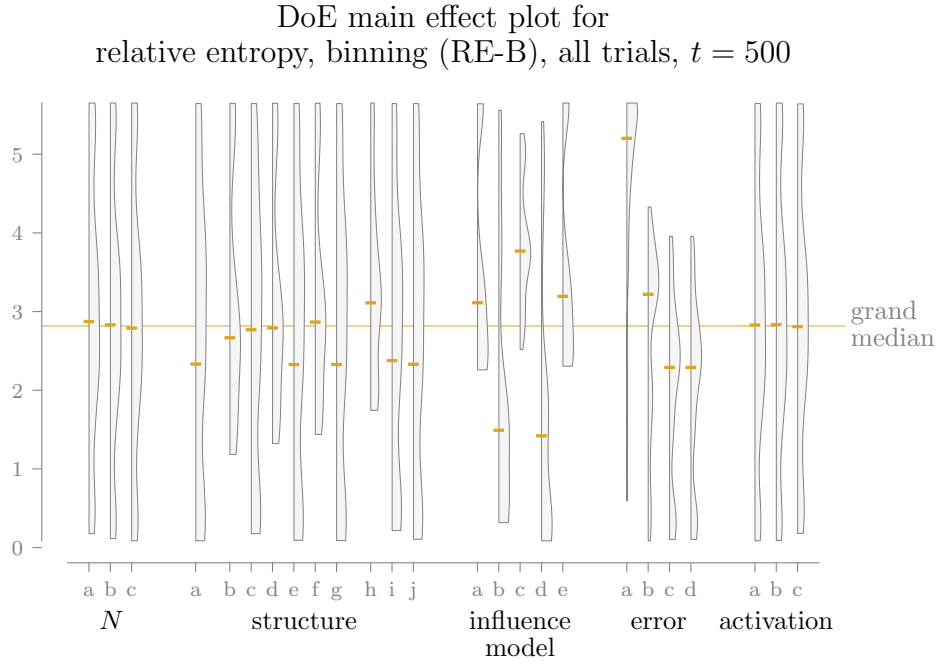

Figure 2: Each half-violin of this design of experiments (DoE) main effect plot represents the distribution of RE-B at $t = 500$ for all trials with the corresponding level on the horizontal axis, and its median is indicated with a horizontal dash; the grand median is shown for reference. For the network structure factor, horizontal space separates the three model families (Erdős-Rényi random, small world, and scale-free). Refer to Table 1 to identify all factor-levels, and see text for further discussion.

This plot suggests that N and activation are unimportant to the response variable, while zero error (level a) leads to significantly different outcomes than the other error distributions.

experimental design; the same holds for the agent activation regime. On the other hand, strong differences between distributions are visible for the influence model and error distribution, marking these as important to RE-B. Structure models appear to fall into at least two groups of similar response distributions, and these groups resemble the higher/lower density networks (identified in Appendix B).

Figure 2 uses data for only a single time step. To reveal the effect of time on the response variable, Figures 3-7 plot the medians of the grouped data over the full length of the simulation. The two inner quartiles (25th to 75th percentiles) are indicated by the shaded regions around each line.

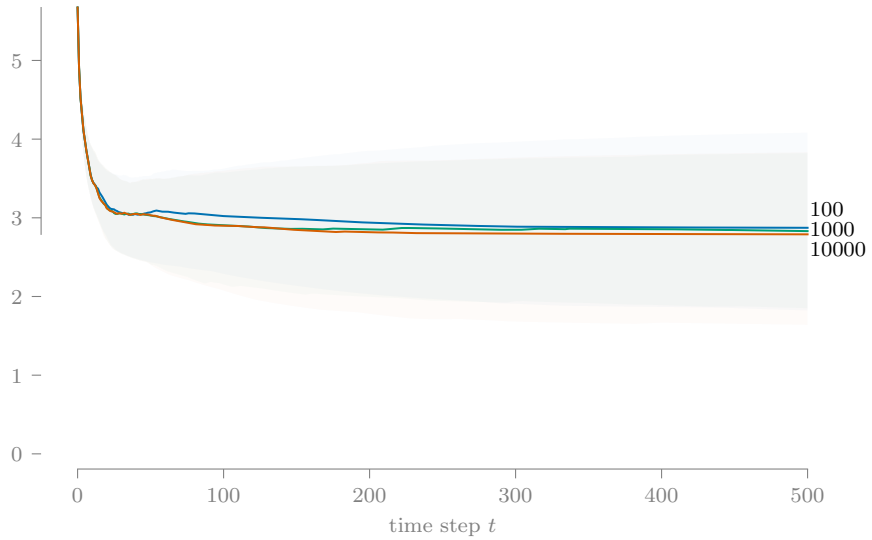

Figure 3: All trials are grouped by factor-level as in Figure 2 and the groups' median response value over time is plotted. Shaded regions around each line enclose the 25th to 75th percentiles of the data. For population size N, these regions almost entirely overlap, which reinforces the low importance of N shown in the DoE main effect plot.

Overall, these figures reinforce the similarities and differences observed in the main effect plot. They also show that the (median) response values are fairly stable over time, after an initial transient, so any observations made at $t = 500$ should be informative about the system over a longer period of time.

Thus far, we have use qualitative approaches to show the effect of varying individual design factors. Now, we turn to statistical measures. ANOVA is the classical tool for analyzing the results of a designed experiment, but it assumes the data to be normally distributed, which is not the case here (Figure 1). Further, Levene's test shows that three of the five design factors violate ANOVA's assumption of homogeneity of variance. While ANOVA is robust to violations of these assumptions, we instead adopt a non-parametric approach to measuring differences between factor-levels.

The Kruskal-Wallace test lets us statistically determine if varying the level of each factor has a significant effect on the response variable. Based on the p-values from the Kruskal-Wallace test (Table 3) on RE-B at $t = 500$, when

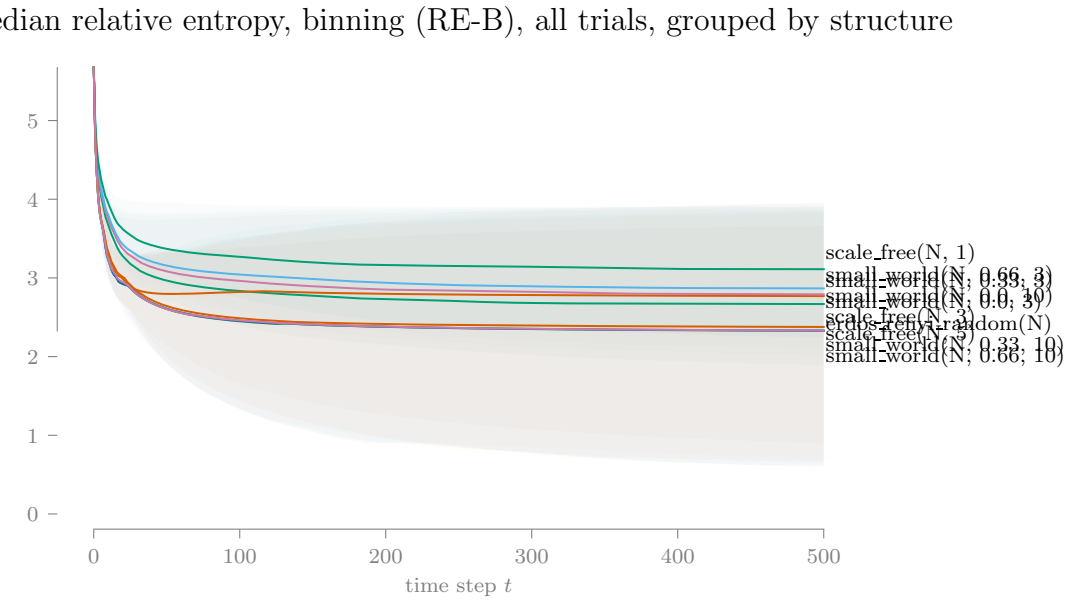

Figure 4: Following Figure 3 in design, this plot shows some differentiation between network models but many are very similar. The line labels are allowed to overlap to reinforce the small differences between factor-levels. The scale free $(N, 1)$ model does stand out above the rest.

median relative entropy, binning (RE-B), all trials, grouped by influence model

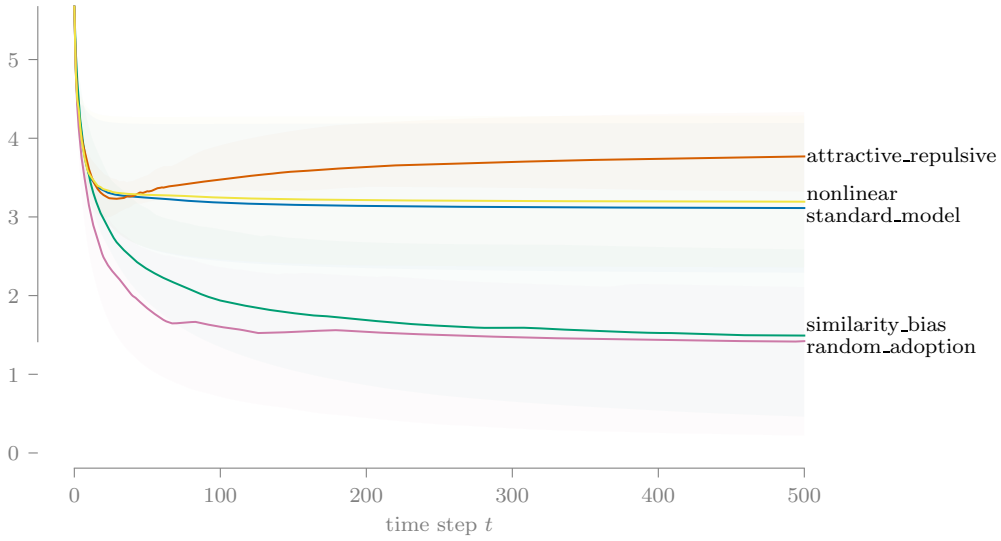

Figure 5: Some grouping is apparent based on the level of the influence model. One explanation is that nonlinear and the standard model both use weighted averages of neighbor opinion, while similarity bias and random adoption interact with (at most) one neighbor at a time.

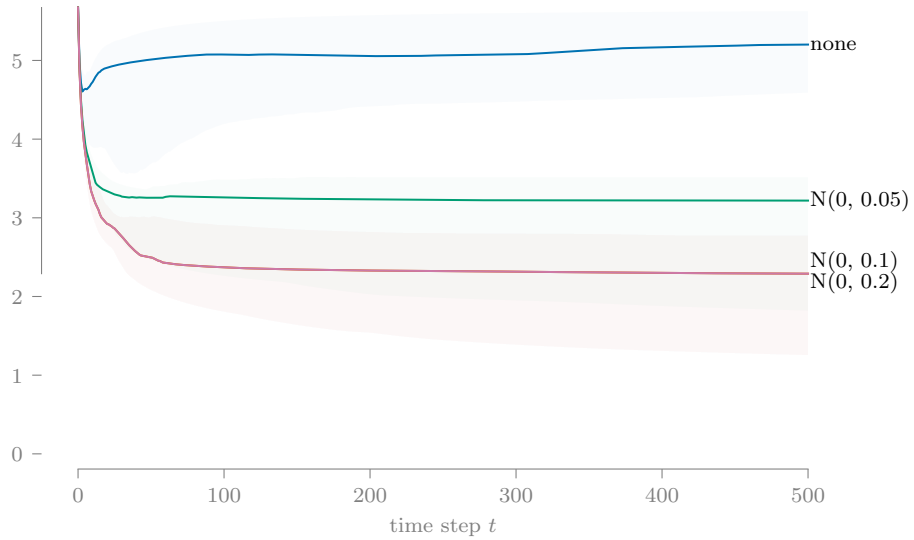

Figure 6: With the influence error distribution, we observe clear separations in the response variable, again reinforcing the findings from Figure 2.

the data is split into levels for both population size N and activation regime, the data appears to come from the same population. Practically, this suggests that varying these factors—over the levels specified in our experiment—does not have a significant effect on the response variable. This agrees with what we observe in the previous figures.

Although the Kruskal-Wallace test indicates that differences exist among levels for the structure, influence model, and error factors, it does not identify where those differences are. For that, we use the Mann-Whitney U test on each pair of levels for a factor. A significant (< 0.05) p-value indicates a statistical difference between the tested pair of factor-levels. Figure 8 aggregates the results. (We include N and activation regime in the results to show consistency between the Mann-Whitney U test results and Kruskal-Wallace.)

Some interesting observations can be made about what pairwise results are/are not significant. For example, in the network structure models, structures with higher density test as similar to nearly all other higher density structures, while structures with lower density test as different from most other low density structures (see Appendix B for density values). Also, this

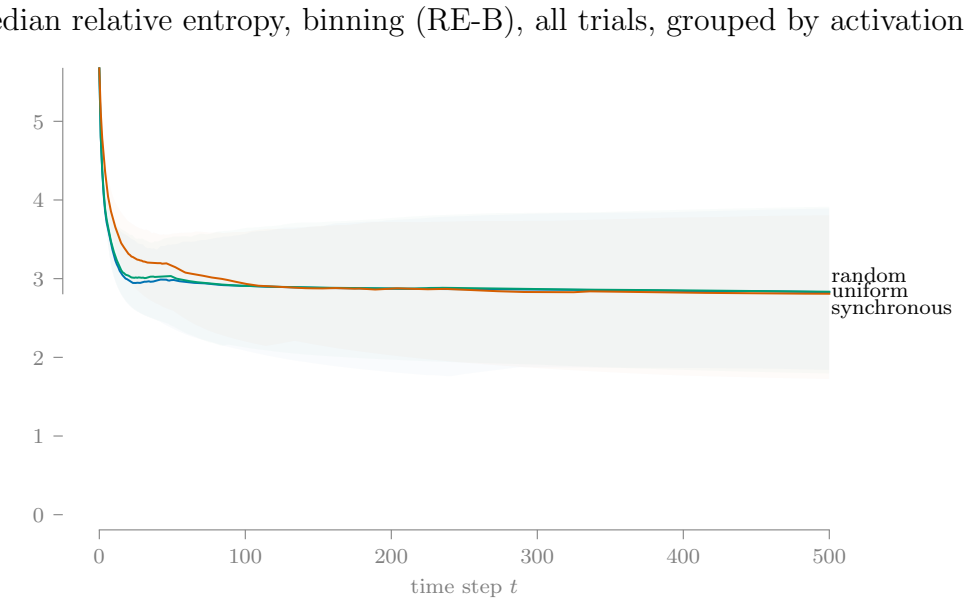

Figure 7: As with population size N, the agent activation regime appears to be unimportant to RE-B, although some differences are present in the early time steps.

Table 3: The Kruskal-Wallace test is ran on trial-level RE-B values at $t = 500$ to test if changing the level for a factor has a statistical effect on the response value. The asterisks indicate that population size N and activation regime show no significant impact on RE-B.

factor	K-W test stat	p-value
N	2.05	* 3.57e-01
structure	29.37	5.60e-04
influence model	622.07	2.58e-133
error	892.12	4.51e-193
activation	0.13	* 9.33e-01

test finds a significant difference between the standard model (level a) and nonlinear (level e) influence models, which is not apparent from the DoE main effect or grouped time series plots.

Overall, for the RE-B response variable, varying population size N and the agent activation regime has no significant effect; influence model error terms set to zero leads to higher relative entropy than normally distributed, zero mean error, and that effect varies slightly with the variance of the distribution; differences in the response due to changing the network structure model are slight, though tree-like structures have a higher minimum response than dense graphs; and influence models that cause agents to interact with one neighbor at a time lead to lower response values than those that interact with all neighbors at once.

1.1.2. Research Question 2: How is system design related to the response space of entropy time-series values?

Cluster analysis has three elements: clustering algorithm, distance (or dissimilarity) measure, and evaluation criteria of the results. Few guidelines exist for designing a cluster analysis *a priori*, so we used an assortment of options to search for meaningful clusters in our entropy time-series data. For the distance (or dissimilarity) measure, we used dynamic time warping (DTW) and Pearson's correlation coefficient between each pair of trial-level time series for the response variable. Then, we used fourteen clustering methods provided by the R library NbClust.[1] Each clustering method suggested an optimal number of clusters (which we bounded between two and twelve, inclusive), then the number of clusters with the most "votes" was passed to an agglomerative hierarchical clustering method to assign each trial to a cluster. This process was applied to both distance measures.

Cluster assignments are summarized in the following figures. DTW for RE-B produced four clusters (Figure 9), while Pearson's correlation produced two clusters (Figure 10). With respect to the time series plots, DTW led to rather differentiated clusters, while Pearson's correlation produced highly homogeneous clusters.

However, clusters differentiated in the response variable space are not necessarily meaningful. In Figures 11 and 12, we conduct a "census" of

[1] NbClust provides more than fourteen methods, but ten produced errors due to the dimensions of the data, and several were omitted due to their overwhelming computation time.

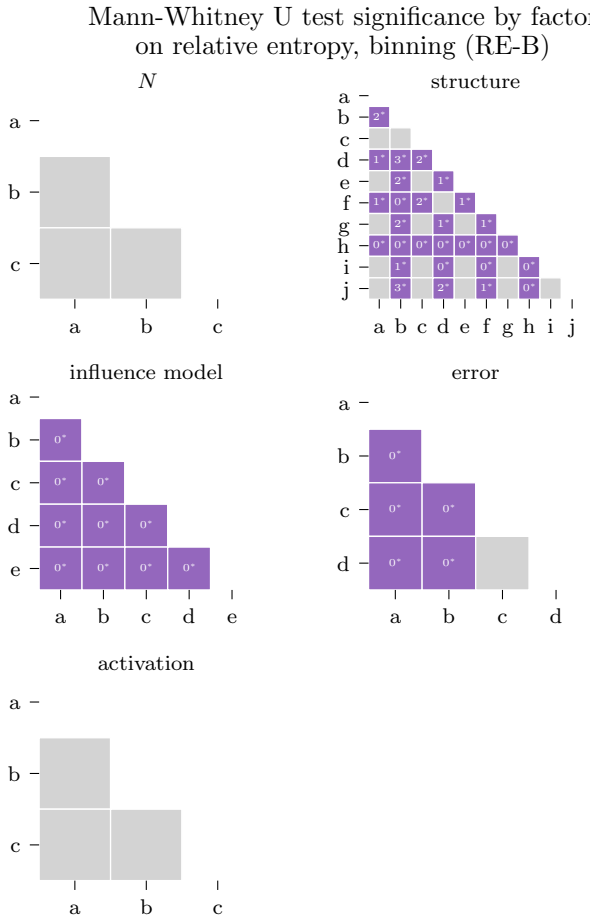

Figure 8: We use the Mann-Whitney U test for post-hoc comparison testing to determine which levels are statistically different within each factor. The numbers in cells for the pairs with a significant test statistic (< 0.05) express the p-value as a percentage (e.g. 3* means $0.03 \leq$ p-value < 0.04). The non-significant results for N and activation are consistent with the previous findings. The results for influence model are somewhat unexpected since similar pairs of models appear in Figures 2 and 5. See text for further discussion.

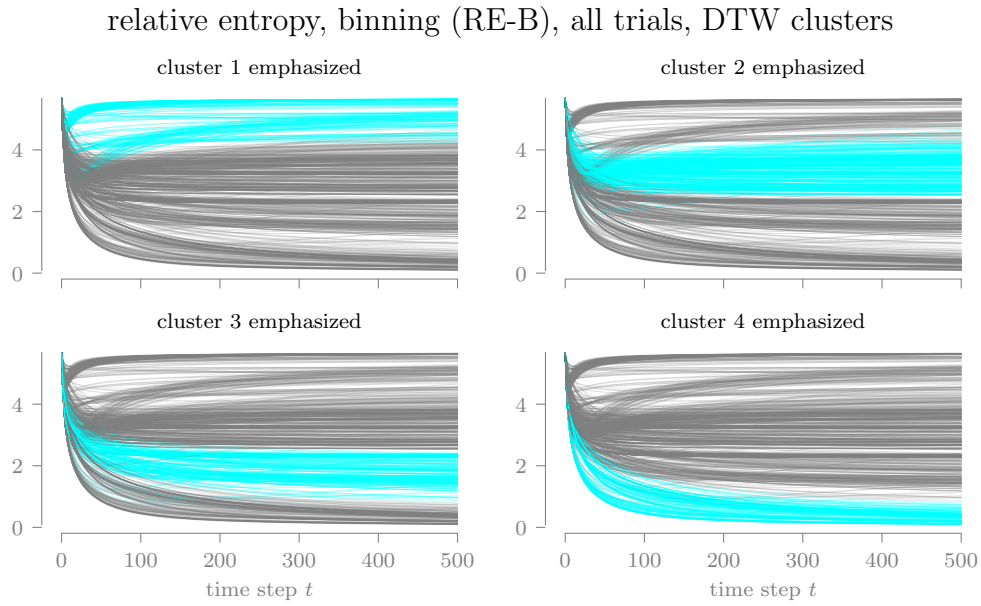

Figure 9: Using dynamic time warping (DTW) as the distance measure between pairs of response variable time series, the consensus method produces four clusters, each highlighted here using the original time series plot (Figure 1). The densely grouped nature of these clusters suggest a high level of cluster quality.

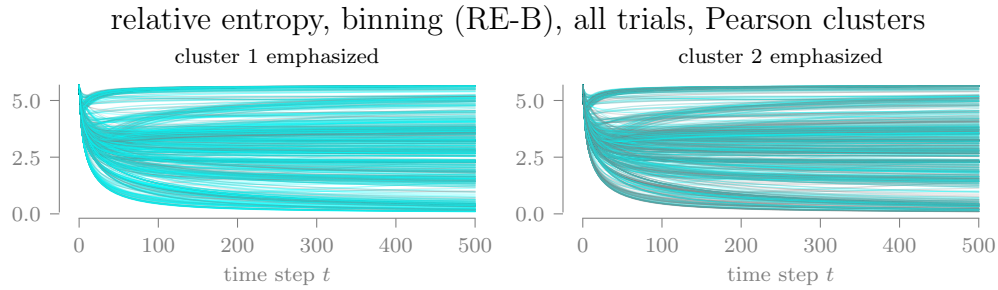

Figure 10: Using Pearson's correlation as the distance measure, the consensus method produces only two clusters. The results are less visually satisfying than DTW (Figure 9) and may indicate less meaningful clusters.

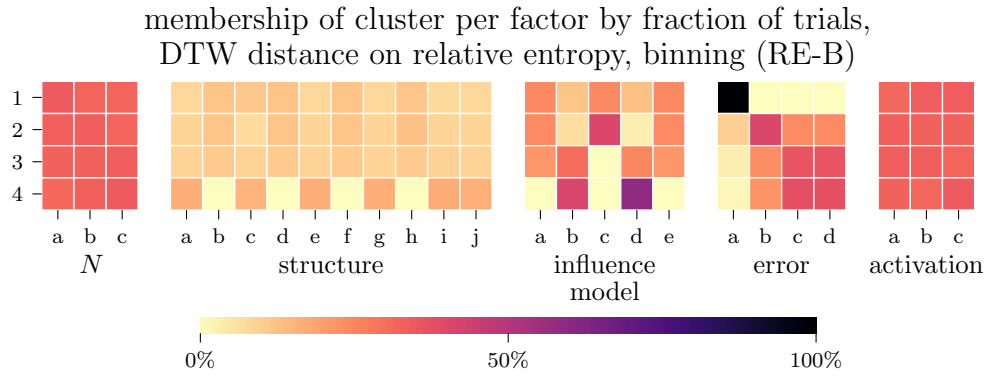

Figure 11: For each cluster produced through DTW, the trials assigned to the cluster are grouped by factor-level in order to find the percentage of a cluster associated with each factor-level. Within a cluster (row), the percentages for a single factor sum to one; within a factor-level (column), there are no such constraints. For example, all trials assigned to cluster 1 use error level a (no error), and trials assigned to cluster 4 use either influence model b or d.

the trials assigned to each cluster, with respect to the experimental design factors. For DTW, cluster 1 contains exclusively trials with no influence error term (error level a), though other clusters do contain a small number of such trials, as well. Cluster 4 is strong in the similarity bias and random adoption influence models (levels b and d), perhaps because those models have agents interact with a single neighbor at a time, while the other models have agents interact with all neighbors at once. Interestingly, cluster 4 is also very weak in lower-density network structures (levels b, d, f, and h). The levels for population size N and activation regime are uniformly distributed within each cluster, further reinforcing the earlier evidence that those factors are not important to RE-B.

In summary, the variation in system design studied here can produce meaningful clusters, with respect to the experimental design factors, in the response space for RE-B. This effect is achieved when using dynamic time warping as the distance measure, but not when using Pearson's correlation coefficient.

1.2. Response variable 2 - mutual information, binning (MI-B)

Mutual information, binning (MI-B) assigns agent opinion to one of a set of equal-width bins and computes the mutual information between each

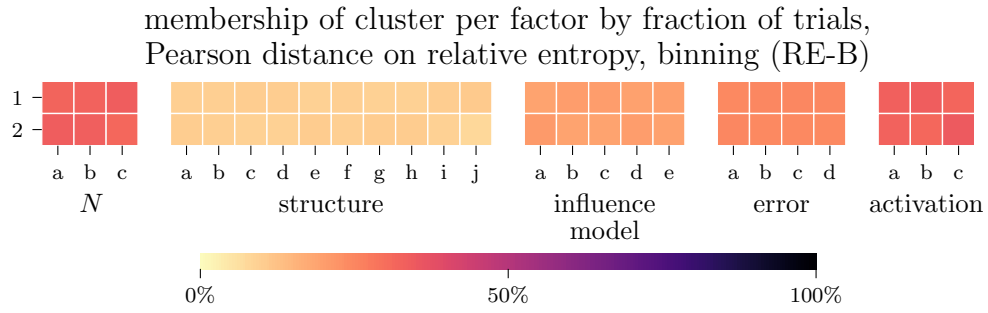

Figure 12: Clusters produced through Pearson's correlation are completely undifferentiated.

agent-neighbor pair, averages across the neighbors for each agent, and then averages across each agent and each replication to produce the trial-level response. Figure 13 shows the time series of MI-B for each trial and an associated kernel density estimate (KDE) for the final time step. This shows a grouping near zero mutual information, and that most—but not all—trials decrease over time.

1.2.1. Research Question 1: Which system design factors contribute most to system-level entropy?

For this research question, we explore the one-way sensitivity of each entropy response variable to changes in the levels of individual experimental design factors. This exploration includes qualitative comparisons of RV distributions when the trial data is grouped by factor-levels and statistical tests for differences between levels. These methods support a subjective evaluation of whether an RV is sensitive to changes in the level of each design factor. In Table 4, we summarize the analysis results for the current response variable for this research question.

Figure 14 presents the distributions grouped by factor-level for MI-B at the final time step, $t = 500$, using a half-violin plot. Differences between distributions among the levels for a single factor qualitatively show the effect each level has on the response. For example, the distributions for population size N are almost identical, so we infer that N is not important (i.e., does not have a significant effect on the response variable), at least over the range of levels used in the experimental design. On the other hand, strong differences between distributions are visible for the influence model, marking it as

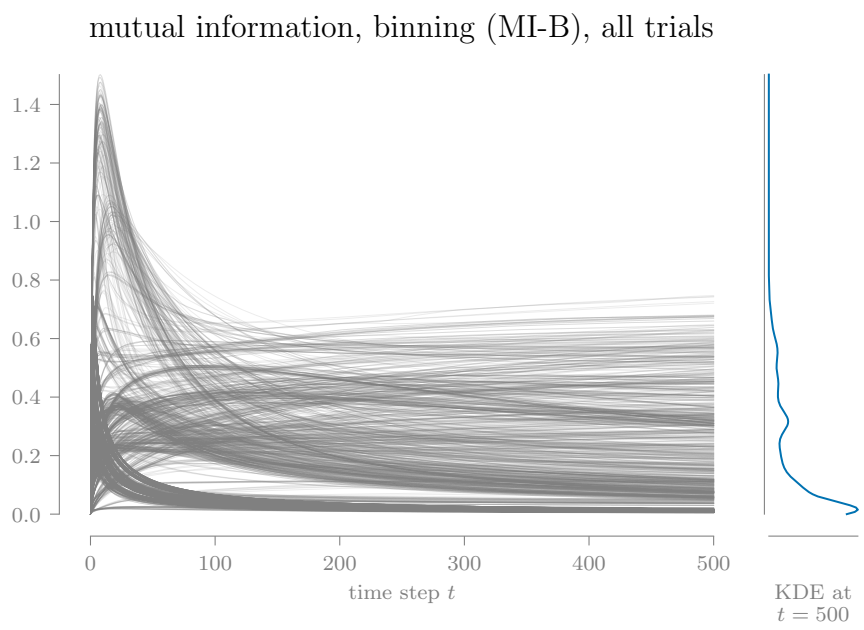

Figure 13: The time series of MI-B values for each trial are plotted on the same axes to reveal a visual cluster at the bottom of the range.

Table 4: The findings for Research Question 1 on MI-B are summarized to support the overall evaluation of each experimental design factor (final table row).

		Factor (number of levels)			
	N (3)	structure (10)	influence model (5)	error (4)	activation (3)
i.	*(Main effect plot) What differences are present between the response variable distributions for each level at the final time step?*				
	negligible	3 or 4 patterns	5 patterns	2 patterns; error vs no error	synchronous different than others
ii.	*(Grouped time series) What differences are present between the median response values over the duration of the simulation?*				
	negligible	initial grouping by density, then partial convergence	4 patterns; nonlinear+standard similar	most initially distinct, then partial convergence	minor early variation, then partial convergence
iii.	*(K-W test) Does the Kruskal-Wallace test indicate statistical differences in the response variable between each level at the final time step? (i.e., is the p-value <0.05?)*				
	no	yes	yes	yes	yes
iv.	*(M-W U test) How many pairs of levels are statistically different (p-value <0.05) according to the Mann-Whitney U test?*				
	0/3	24/45	10/10	3/6	3/3
*	*(Evaluation) Is the response variable sensitive to changes in the level for the factor?*				
	no	yes	yes	yes	yes

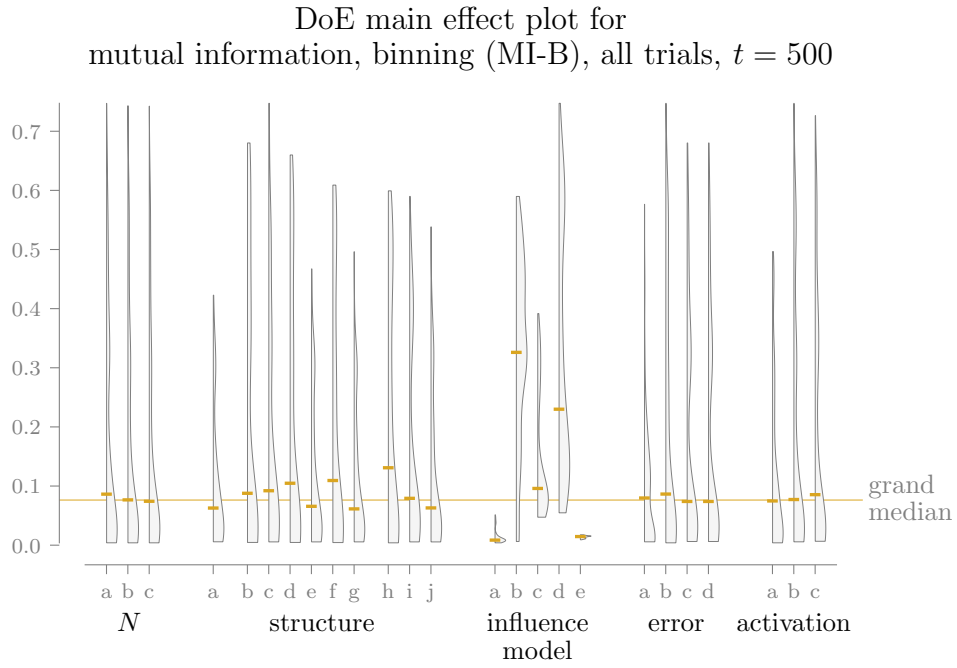

Figure 14: Each half-violin of this design of experiments (DoE) main effect plot represents the distribution of MI-B at $t = 500$ for all trials with the corresponding level on the horizontal axis, and its median is indicated with a horizontal dash; the grand median is shown for reference. This plot suggests that N is unimportant to the response variable, while influence model leads to highly varied outcomes.

important to MI-B.

Figure 14 uses data for only a single time step. To reveal the effect of time on the response variable, Figures 15-19 plot the medians of the grouped data over the full length of the simulation. The two inner quartiles (25th to 75th percentiles) are indicated by the shaded regions around each line. Overall, these figures reinforce the similarities and differences observed in the main effect plot.

Thus far, we have use qualitative approaches to show the effect of varying individual design factors. We now adopt a non-parametric approach to measuring differences between factor-levels, using the Kruskal-Wallace test and Mann-Whitney U test.

Based on the p-values from the Kruskal-Wallace test (Table 5) on MI-B at $t = 500$, when the data is split into levels for population size N, the data

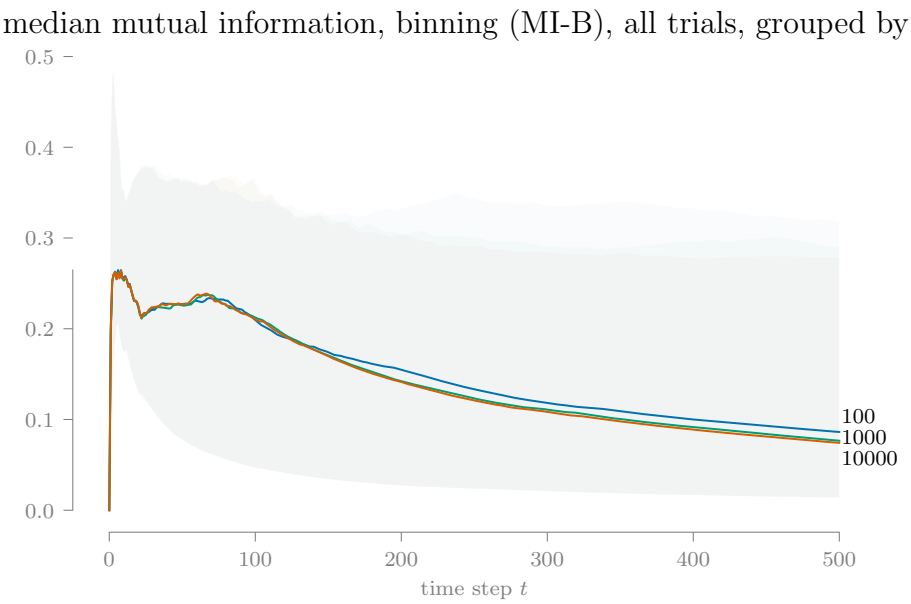

Figure 15: All trials are grouped by factor-level as in Figure 14 and the groups' median response value over time is plotted. Shaded regions around each line enclose the 25th to 75th percentiles of the data. For population size N, these regions almost entirely overlap due to the closeness of the median lines, which reinforces the low importance of N shown in the DoE main effect plot.

median mutual information, binning (MI-B), all trials, grouped by structure

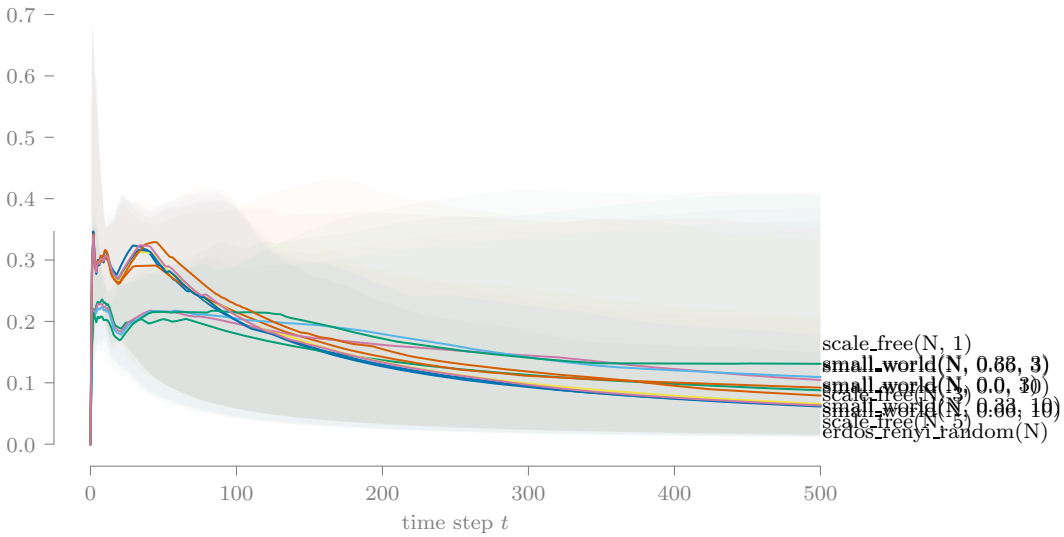

Figure 16: Following Figure 15 in design, this plot shows some differentiation between network models but many are very similar. An interesting split into two groups occurs during the initial transient between lower density (lower group) and higher density (upper group) networks.

median mutual information, binning (MI-B), all trials, grouped by influence model

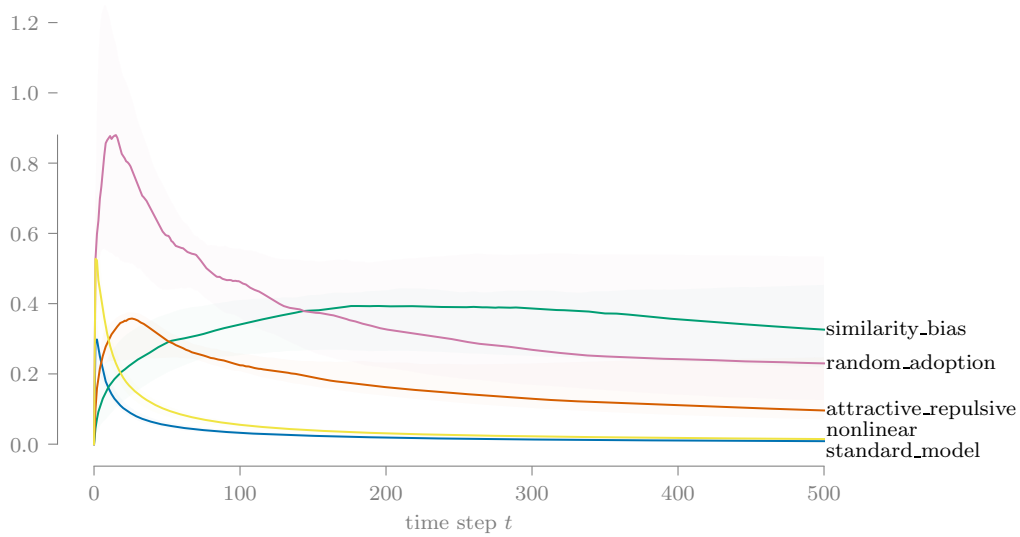

Figure 17: The standard model and nonlinear model are closely aligned in this plot, but all other levels are well-differentiated from each other. One explanation is that nonlinear and the standard model both use weighted averages of neighbor opinion, while the other models do not.

median mutual information, binning (MI-B), all trials, grouped by error

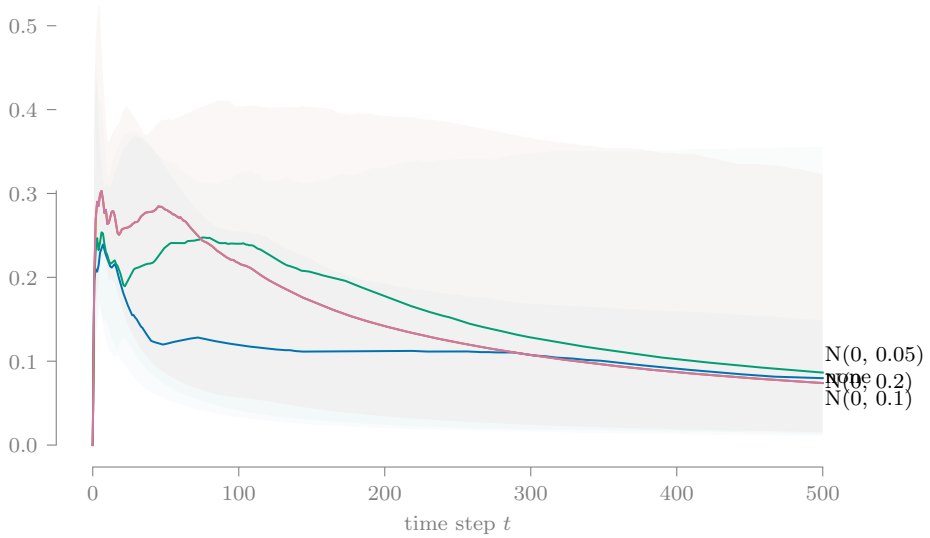

Figure 18: With the influence error distribution, we observe initial differences in the response variable, but the median lines converge after $t = 300$.

appears to come from the same population. Practically, this suggests that varying this factor—over the levels specified in our experiment—does not have a significant effect on the response variable. This agrees with what we observe in the previous figures.

Figure 20 aggregates the results of the Mann-Whitney U test applied to each pair of levels within a factor. Some interesting observations can be made about what pairwise results are/are not significant. In the network structure models, only some of the similar pairs of levels (i.e., those with grey cells in the figure) have similar densities.

Overall, for the MI-B response variable, varying population size N and agent activation regime has no significant effect; differences in network structure models and error terms have greater effects early in a run but very minor effects in the long term; and each influence model has a distinct signature.

1.2.2. Research Question 2: How is system design related to the response space of entropy time-series values?

Using the cluster analysis process described in Section 1.1, trials are assigned to clusters for both DTW and Pearson's correlation. These assign-

median mutual information, binning (MI-B), all trials, grouped by activation

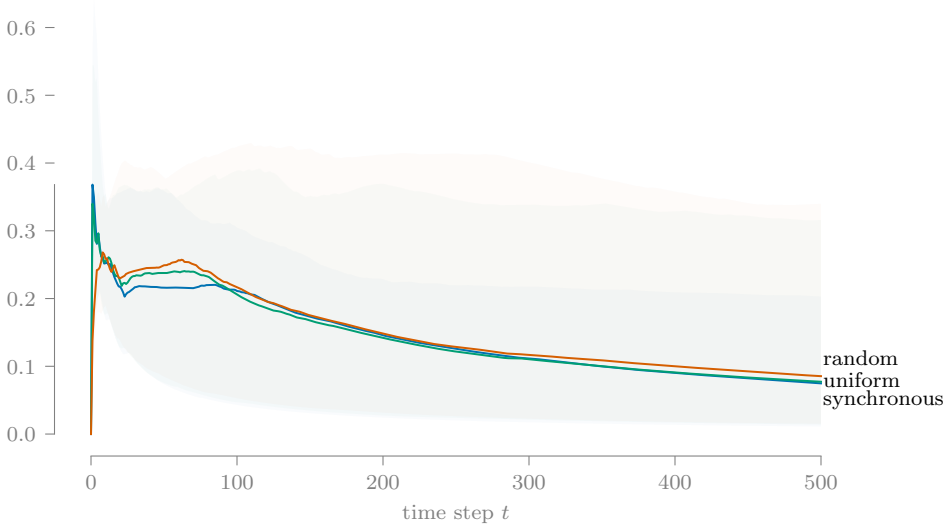

Figure 19: The median lines for the three activation regimes are nearly identical, but their variability differs, as also indicated in the main effect plot.

Table 5: The Kruskal-Wallace test is ran on trial-level MI-B values at $t = 500$ to test if changing the level for a factor has a statistical effect on the response value. The asterisk indicates that population size N has no significant impact on MI-B.

	test stat	p-value
N	1.12	* 5.69e-01
structure	28.80	6.99e-04
influence model	1317.66	4.93e-284
error	47.96	2.16e-10
activation	41.50	9.70e-10

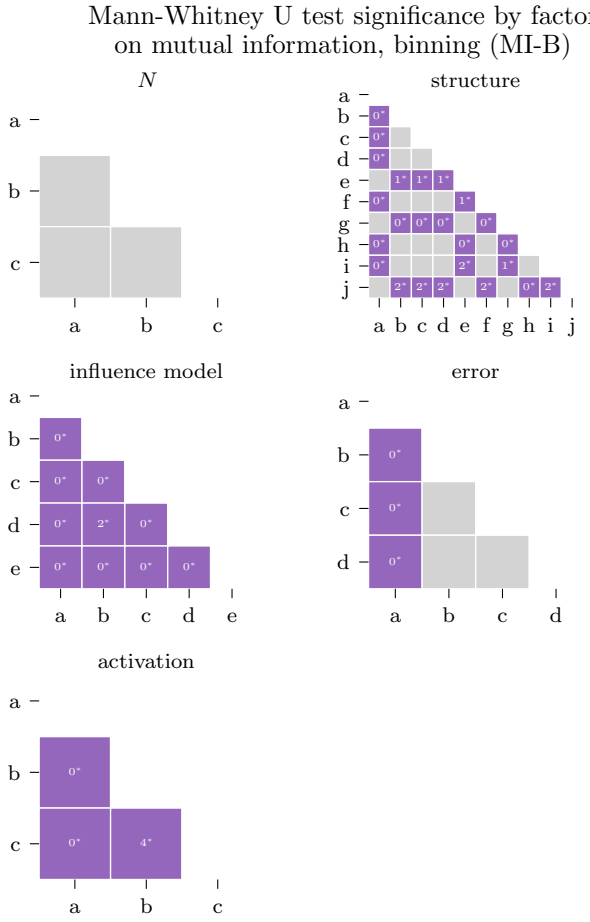

Figure 20: We use the Mann-Whitney U test to determine which levels are statistically different within each factor. The numbers in cells for the pairs with a significant test statistic (< 0.05) express the p-value as a percentage (e.g. 3* means $0.03 \leq$ p-value < 0.04). The non-significant results for N are consistent with the previous findings.

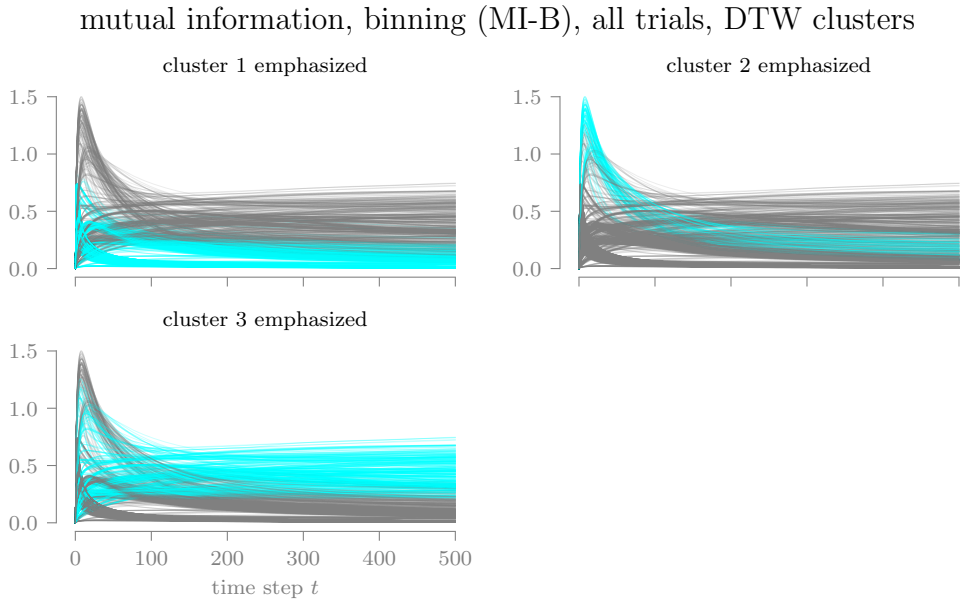

Figure 21: Using dynamic time warping (DTW) as the distance measure between pairs of response variable time series, the consensus method produces three clusters, each highlighted here using the original time series plot (Figure 13). The densely grouped nature of these clusters suggest a reasonable level of cluster quality.

ments are summarized in the following figures. DTW for MI-B produced three clusters (Figure 21), while Pearson's correlation produced twelve clusters (Figure 22), the maximum number of clusters considered by our analysis process. (Conventional guidance says that if multiple clustering methods call for the minimum/maximum number of clusters, then the selected methods or distance metric may be unsuitable for clustering the data.) With respect to the time series plots, DTW led to rather differentiated clusters, while Pearson's correlation produced highly homogeneous clusters.

In Figures 23 and 24, we conduct a "census" of the trials assigned to each cluster, with respect to the experimental design factors. For DTW, cluster 1 contains exclusively trials with the random adoption influence model (level d) and is low in trials with tree-like network structures For Pearson's correlation, cluster membership is homogeneous and indistinct.

In summary, the variation in system design studied here can produce meaningful clusters, with respect to the experimental design factors, in the response space for MI-B. This effect is achieved when using dynamic time

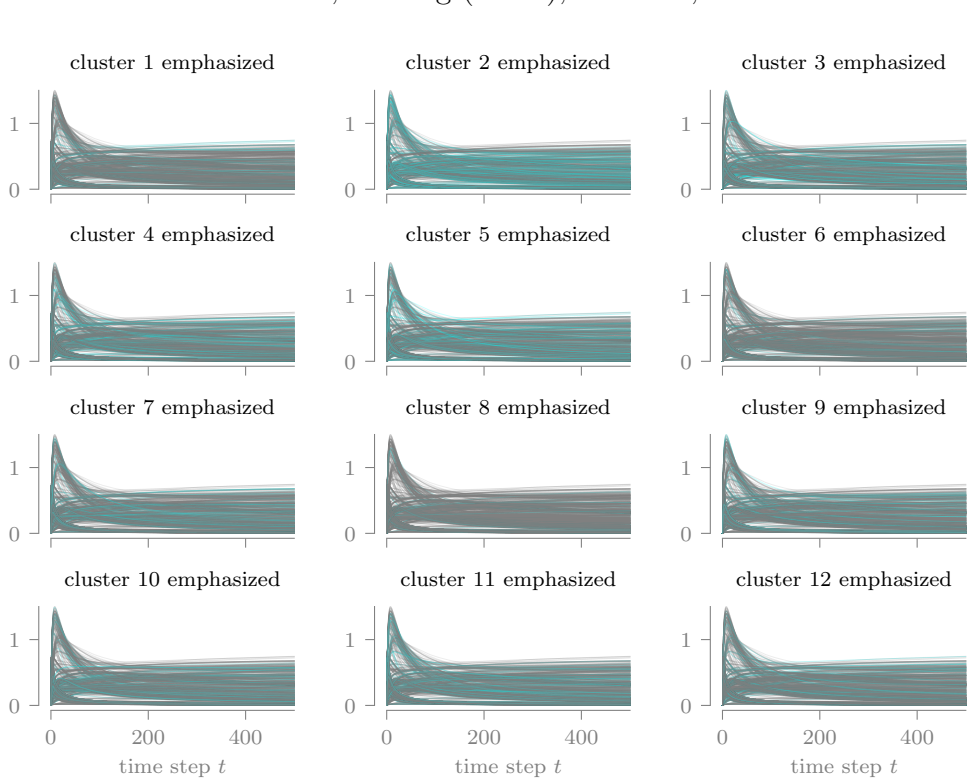

Figure 22: Using Pearson's correlation as the distance measure, the consensus method produces twelve clusters. The results show no clear pattern and may indicate less meaningful clusters.

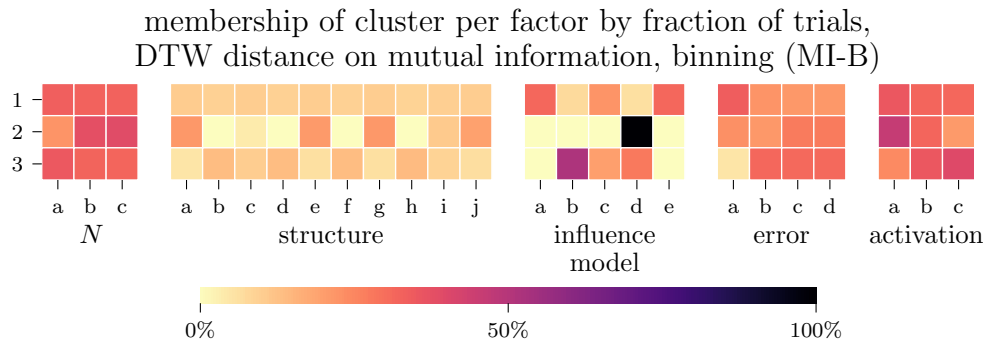

Figure 23: For each cluster produced through DTW, the trials assigned to the cluster are grouped by factor-level in order to find the percentage of a cluster associated with each factor-level. For example, all trials assigned to cluster 2 use influence model d (random adoption) and include more of the higher density network structures.

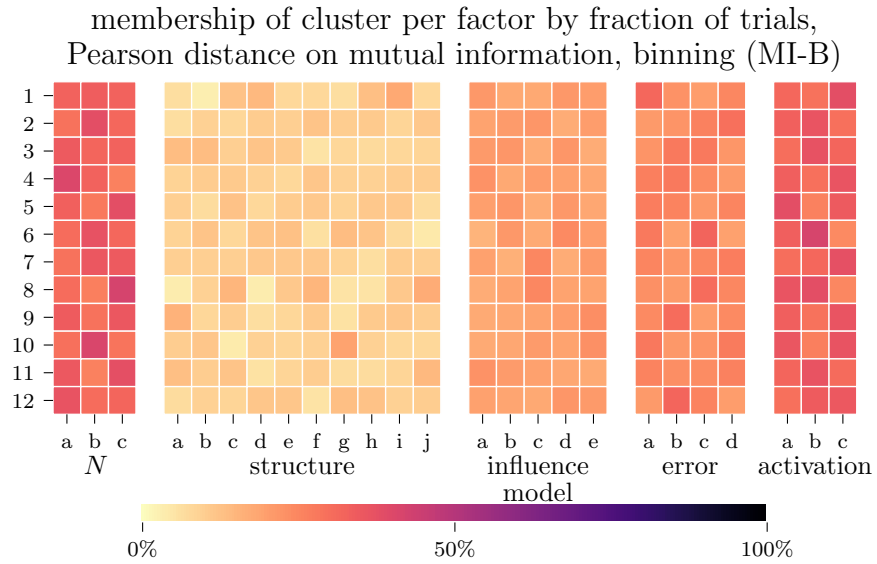

Figure 24: Clusters produced through Pearson's correlation are mostly undifferentiated, suggesting this distance measure is unsuitable for the response variable.

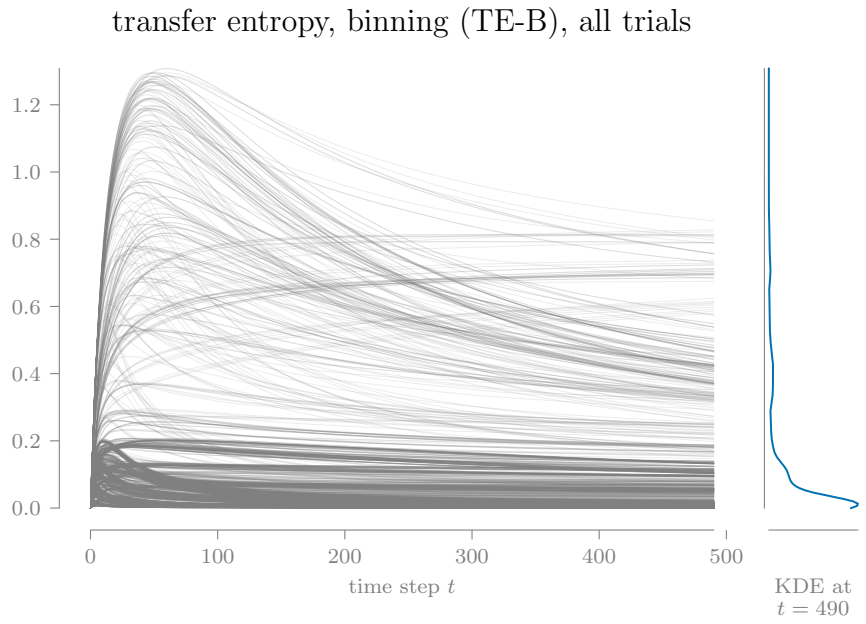

Figure 25: The time series of TE-B values for each trial are plotted on the same axes to reveal a visual cluster at the bottom of the range.

warping as the distance measure, but not when using Pearson's correlation coefficient.

1.3. Response variable 3 - transfer entropy, binning (TE-B)

Transfer entropy, binning (TE-B) assigns agent opinion to one of a set of equal-width bins and computes the transfer entropy between each agent-neighbor pair, averages across the neighbors for each agent, and then averages across each agent and each replication to produce the trial-level response. Figure 25 shows the time series of TE-B for each trial and an associated kernel density estimate (KDE) for the final time step.[2] This shows a grouping near zero transfer entropy, and that most—but not all—trials decrease over time.

[2]Because transfer entropy as calculated here uses two time steps per calculation, the final time step of the simulation has no response value. We trim the data to $t = 490$ for ease of reading but observe no unusual behavior after 490.

1.3.1. Research Question 1: Which system design factors contribute most to system-level entropy?

For this research question, we explore the one-way sensitivity of each entropy response variable to changes in the levels of individual experimental design factors. This exploration includes qualitative comparisons of RV distributions when the trial data is grouped by factor-levels and statistical tests for differences between levels. These methods support a subjective evaluation of whether an RV is sensitive to changes in the level of each design factor. In Table 6, we summarize the analysis results for the current response variable for this research question.

Table 6: The findings for Research Question 1 on TE-B are summarized to support the overall evaluation of each experimental design factor (final table row).

	Factor (number of levels)				
	N (3)	structure (10)	influence model (5)	error (4)	activation (3)
i.	*(Main effect plot) What differences are present between the response variable distributions for each level at the final time step?*				
	negligible	2 or 3 patterns	5 patterns	2 patterns; error vs no error	3 patterns
ii.	*(Grouped time series) What differences are present between the median response values over the duration of the simulation?*				
	negligible	initial grouping by density, then partial convergence	3 patterns; random adoption is outlier	3 patterns; high+medium variance identical	minor early variation, then partial convergence
iii.	*(K-W test) Does the Kruskal-Wallace test indicate statistical differences in the response variable between each level at the final time step? (i.e., is the p-value <0.05?)*				
	no	no	yes	yes	yes
iv.	*(M-W U test) How many pairs of levels are statistically different (p-value <0.05) according to the Mann-Whitney U test?*				
	0/3	3/45	10/10	5/6	2/3
*	*(Evaluation) Is the response variable sensitive to changes in the level for the factor?*				
	no	yes	yes	yes	yes

Figure 26 presents the distributions grouped by factor-level for TE-B at the final time step, $t = 490$, using a half-violin plot. Differences between distributions among the levels for a single factor qualitatively show the effect each level has on the response. For example, the distributions for population size N are almost identical, so we infer that N is not important (i.e., does not have a significant effect on the response variable), at least over the range

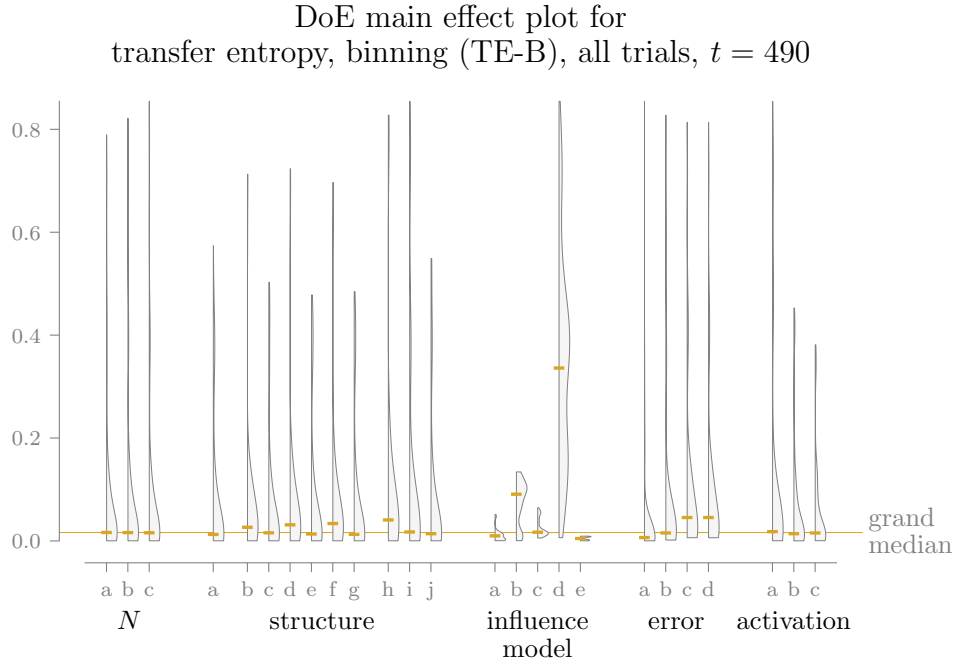

Figure 26: Each half-violin of this design of experiments (DoE) main effect plot represents the distribution of TE-B at $t = 500$ for all trials with the corresponding level on the horizontal axis, and its median is indicated with a horizontal dash; the grand median is shown for reference. This plot suggests that N is unimportant to the response variable, while influence model leads to highly varied outcomes.

of levels used in the experimental design. On the other hand, strong differences between distributions are visible for the influence model, marking it as important to TE-B.

Figure 26 uses data for only a single time step. To reveal the effect of time on the response variable, Figures 27-31 plot the medians of the grouped data over the full length of the simulation. The two inner quartiles (25th to 75th percentiles) are indicated by the shaded regions around each line. Overall, these figures reinforce the similarities and differences observed in the main effect plot.

Thus far, we have use qualitative approaches to show the effect of varying individual design factors. We now adopt a non-parametric approach to measuring differences between factor-levels, using the Kruskal-Wallace test and Mann-Whitney U test.

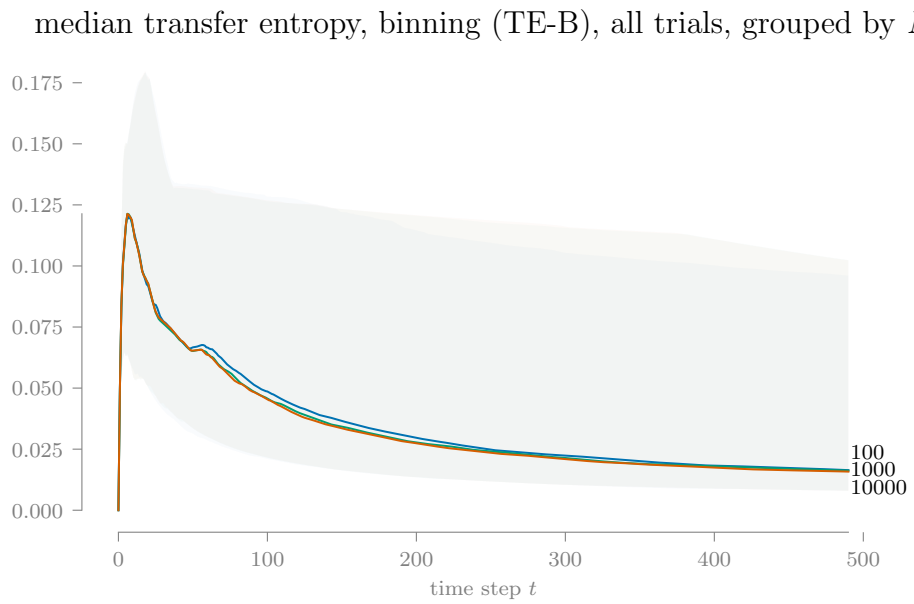

Figure 27: All trials are grouped by factor-level as in Figure 26 and the groups' median response value over time is plotted. Shaded regions around each line enclose the 25th to 75th percentiles of the data. For population size N, these regions almost entirely overlap due to the closeness of the median lines, which reinforces the low importance of N shown in the DoE main effect plot.

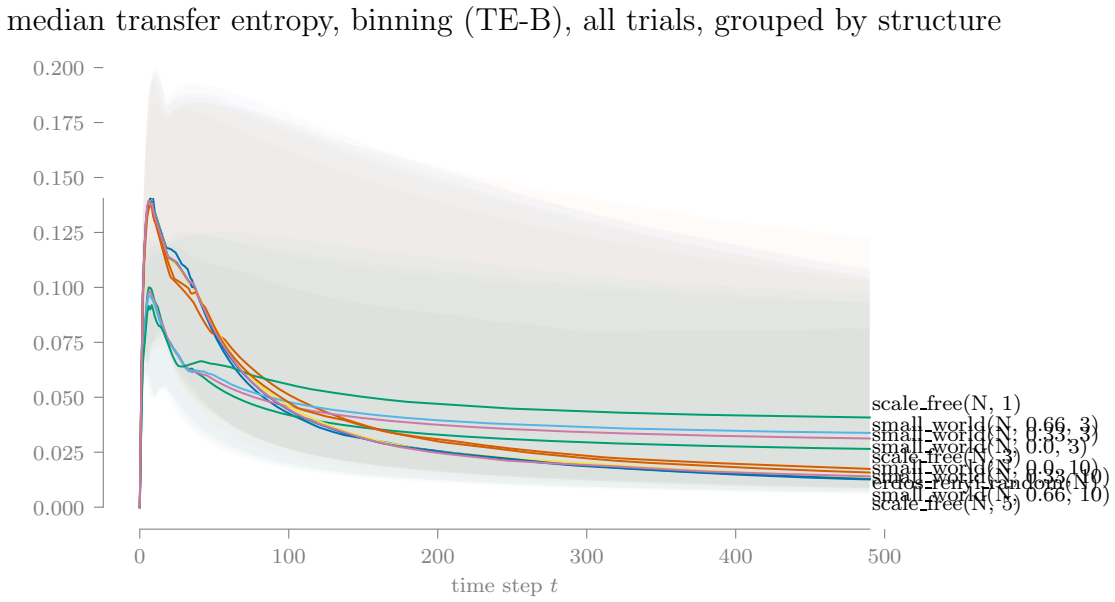

Figure 28: Following Figure 27 in design, this plot shows some differentiation between network models but many are very similar. A distinct split into two groups occurs during the initial transient between lower density (lower group) and higher density (upper group) network structures.

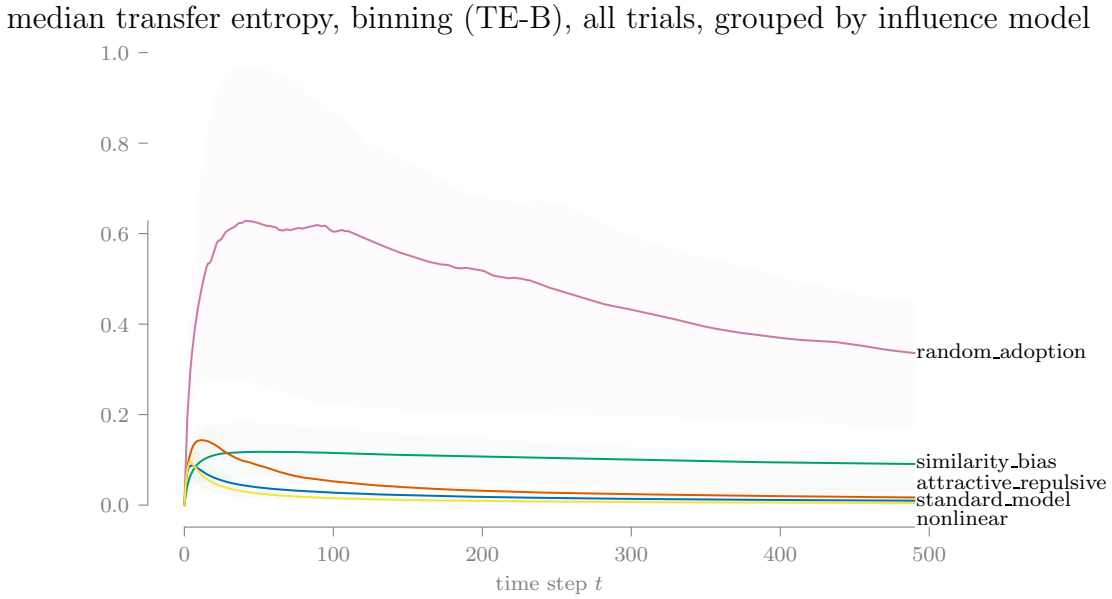

Figure 29: The standard model, attractive-repulsive model, and nonlinear model are closely aligned in this plot; similarity bias has a different characteristic shape, and random adoption is an extreme outlier.

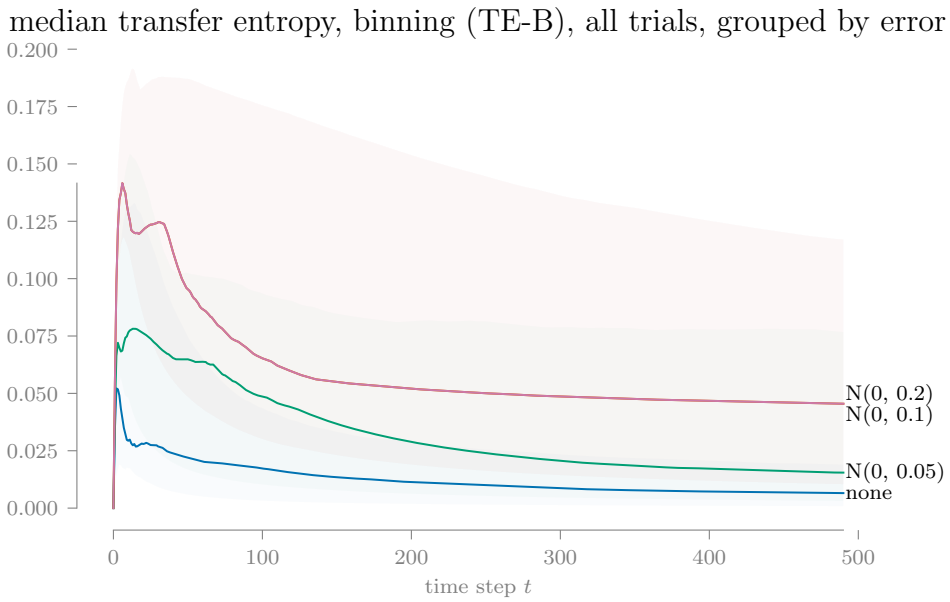

Figure 30: With the influence error distribution, we observe clear differences in the response variable, but the differences are very small relative to the total range of transfer entropy observed in the trials.

median transfer entropy, binning (TE-B), all trials, grouped by activation

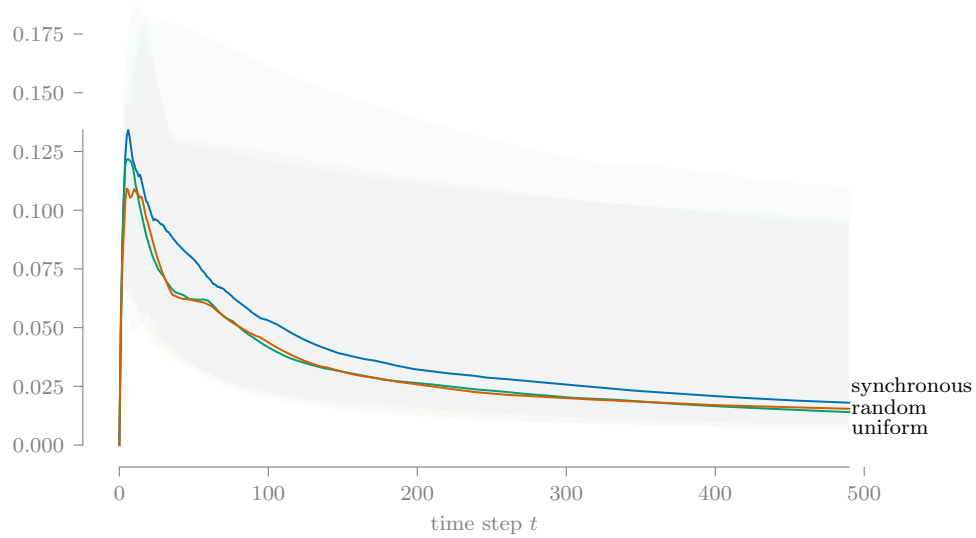

Figure 31: The median lines for the three activation regimes are very similar, but their variability differs, as also indicated in the main effect plot.

Based on the p-values from the Kruskal-Wallace test (Table 7) on TE-B at $t = 490$, when the data is split into levels for population size N and network structure model, the data appears to come from the same population. Practically, this suggests that varying these factors—over the levels specified in our experiment—does not have a significant effect on the response variable. This agrees with what we observe in the previous figures.

Figure 32 aggregates the results of the Mann-Whitney U test applied to each pair of levels within a factor.

Overall, for the TE-B response variable, varying population size N and agent activation regime have no real effect; the different densities of the network structure models have an effect on the response; and influence models and error distributions have distinct signatures.

1.3.2. Research Question 2: How is system design related to the response space of entropy time-series values?

Using the cluster analysis process described in Section 1.1, trials are assigned to clusters for both DTW and Pearson's correlation. These assignments are summarized in the following figures. Both DTW and Pearson's

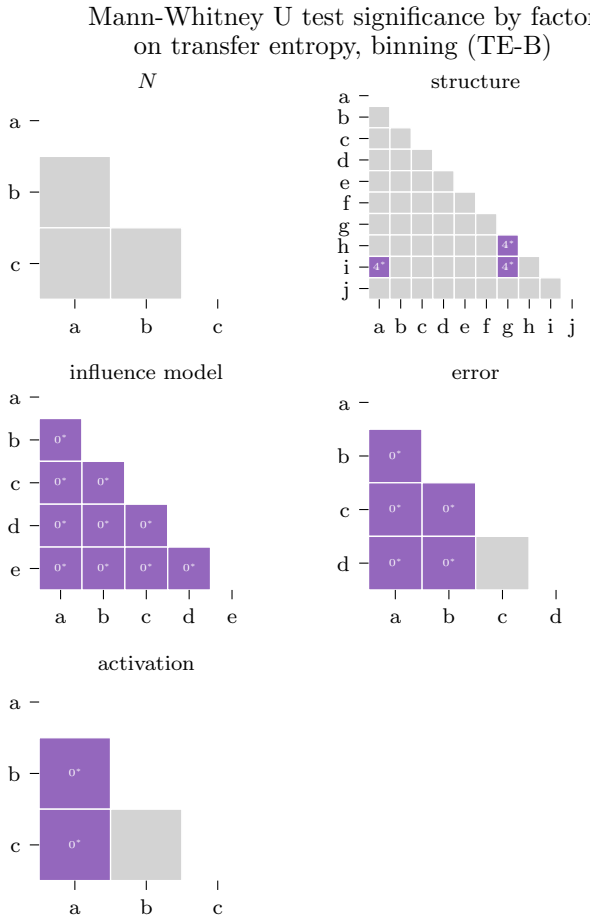

Figure 32: We use the Mann-Whitney U test to determine which levels are statistically different within each factor. The numbers in cells for the pairs with a significant test statistic (< 0.05) express the p-value as a percentage (e.g. 3* means $0.03 \leq$ p-value < 0.04). The non-significant results for N are consistent with the previous findings. The low number of differences in the network structure models is unexpected, since the main effect plot showed marked differences in distribution tail length among some levels.

Table 7: The Kruskal-Wallace test is ran on trial-level TE-B values at $t = 490$ to test if changing the level for a factor has a statistical effect on the response value. The asterisk indicates that population size N has no significant impact on TE-B.

	test stat	p-value
N	0.04	* 9.76e-01
structure	7.11	* 6.24e-01
influence model	1284.07	9.41e-277
error	267.21	1.23e-57
activation	12.38	2.04e-03

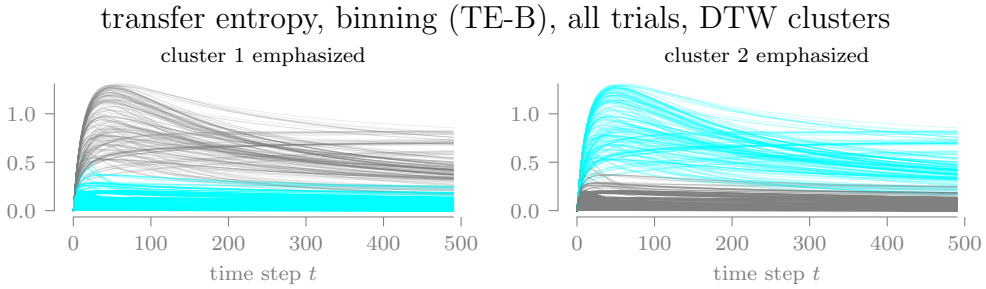

Figure 33: Using dynamic time warping (DTW) as the distance measure between pairs of response variable time series, the consensus method produces two clusters, each highlighted here using the original time series plot (Figure 25). The densely grouped nature of these clusters suggest a reasonable level of cluster quality.

correlation produced two clusters for TE-B (Figures 34 and Figure 33). (Conventional guidance says that if multiple clustering methods call for the minimum/maximum number of clusters, then the selected methods or distance metric may be unsuitable for clustering the data.) With respect to the time series plots, DTW led to rather differentiated clusters, while Pearson's correlation did not.

In Figures 35 and 36, we conduct a "census" of the trials assigned to each cluster, with respect to the experimental design factors. For DTW, one cluster contains exclusively trials with influence model d (random adoption), but is also low in trials with tree-like network structures. For Pearson's correlation, cluster membership is almost entirely homogeneous.

In summary, the variation in system design studied here does produce a

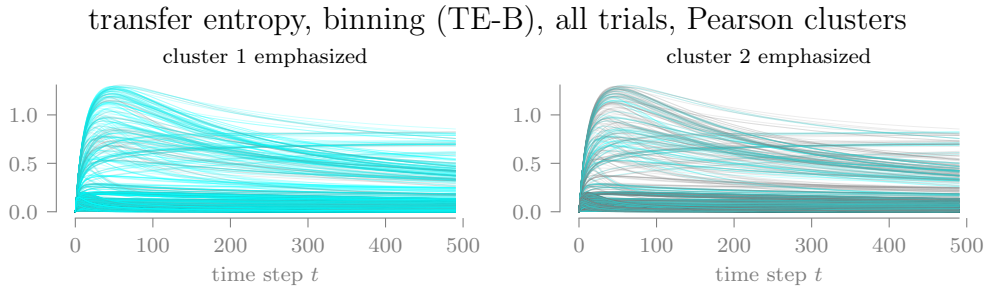

Figure 34: Using Pearson's correlation as the distance measure, the consensus method produces two clusters. The results show no clear pattern and may indicate less meaningful clusters. Difference in perceived brightness is due to the z-ordering of the lines and different cluster sizes.

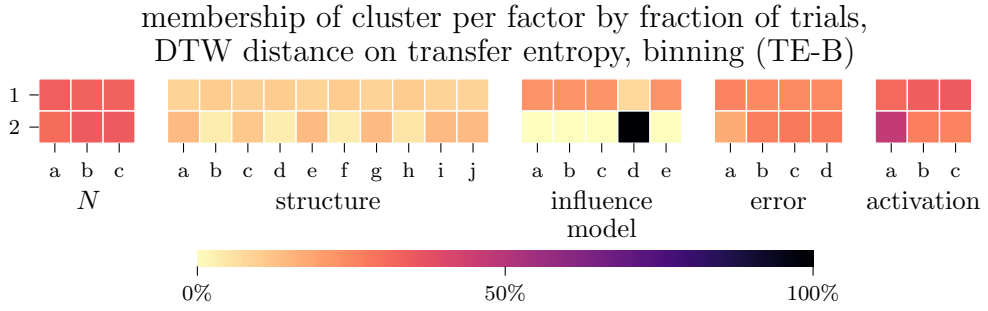

Figure 35: For each cluster produced through DTW, the trials assigned to the cluster are grouped by factor-level in order to find the percentage of a cluster associated with each factor-level. For example, all trials assigned to cluster 2 use influence model d (random adoption) and fewer trials with lower density networks (b, d, f, and h).

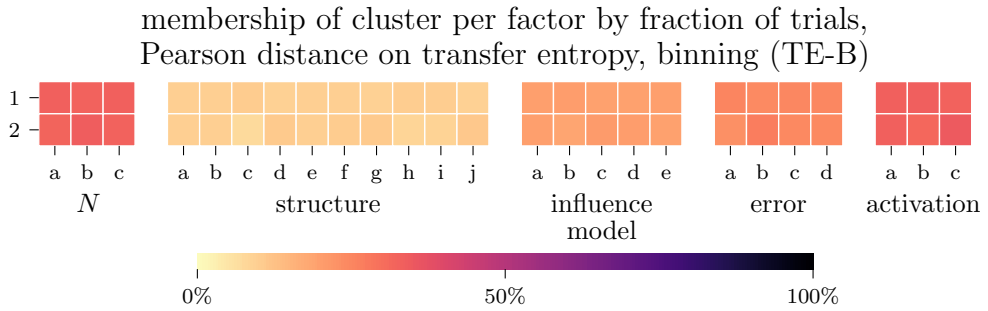

Figure 36: Clusters produced through Pearson's correlation are completely undifferentiated, suggesting this distance measure is unsuitable for the response variable.

somewhat meaningful cluster, with respect to the experimental design factors, in the response space for TE-B. This effect is achieved when using dynamic time warping as the distance measure, but not when using Pearson's correlation coefficient. However, in both cases, only two clusters were created, which limits the usefulness of cluster analysis for this response variable.

1.4. Response variable 4 - relative entropy, symbolic approach (RE-S)

Relative entropy, symbolic approach (RE-S) transforms each agent's sequence of opinion values into a pattern of relative orderings and computes the relative entropy of the resulting distribution $p(x)$ with respect to the uniform distribution $q(x)$, averaging across each agent and each replication to produce the trial-level response. Figure 37 shows the time series of RE-S for each trial and an associated kernel density estimate (KDE) for the final time step.[3] The upper extreme appears to correspond to trials where individual agent opinion converged; we use six symbols (patterns) for RE-B, so the upper limit for relative entropy with respect to the uniform distribution occurs when its opinion takes on only a single value:

$$D_X(p \parallel q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} = 1 \log_2 \frac{1}{1/6} \approx 2.585. \tag{2}$$

1.4.1. Research Question 1: Which system design factors contribute most to system-level entropy?

For this research question, we explore the one-way sensitivity of each entropy response variable to changes in the levels of individual experimental design factors. This exploration includes qualitative comparisons of RV distributions when the trial data is grouped by factor-levels and statistical tests for differences between levels. These methods support a subjective evaluation of whether an RV is sensitive to changes in the level of each design factor. In Table 8, we summarize the analysis results for the current response variable for this research question.

Figure 38 presents the distributions grouped by factor-level for RE-S at the final time step, $t = 490$, using a half-violin plot. Differences between

[3]Because the symbolic method uses multiple time steps per calculation, the final time steps of the simulation have no response value. We trim the data to $t = 490$ for ease of reading.

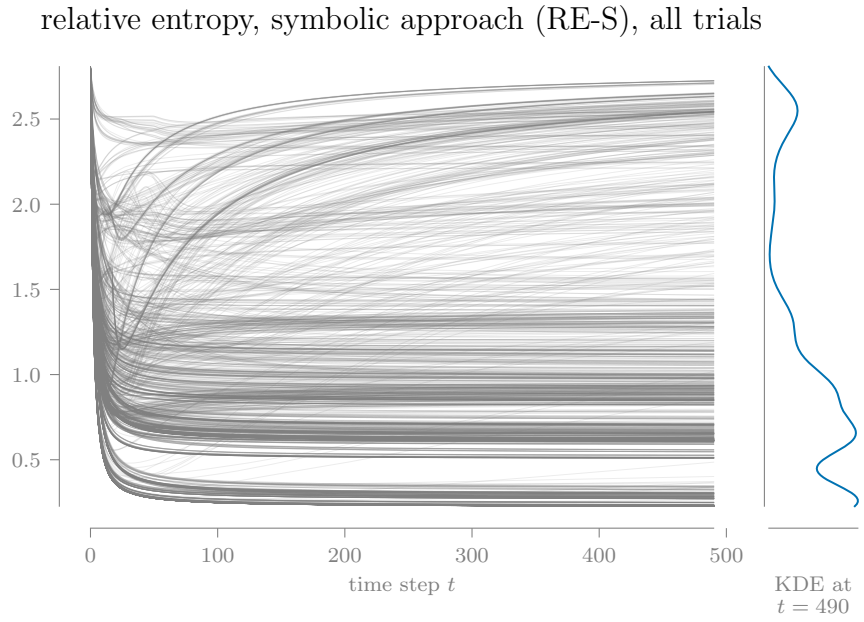

Figure 37: The time series of RE-S values for each trial are plotted on the same axes to reveal several visual clusters near the bottom of the range and a small grouping near the maximum value.

Table 8: The findings for Research Question 1 on RE-S are summarized to support the overall evaluation of each experimental design factor (final table row).

	Factor (number of levels)				
	N (3)	structure (10)	influence model (5)	error (4)	activation (3)
i.	*(Main effect plot) What differences are present between the response variable distributions for each level at the final time step?*				
	negligible	2 patterns	2 or 3 patterns; attractive-repulsive significantly different	2 patterns; error vs no error	random adoption different than others
ii.	*(Grouped time series) What differences are present between the median response values over the duration of the simulation?*				
	negligible	overall similar shapes; small divide affected by density	2 patterns; attractive-repulsive distinct	no error significantly different from rest	2 patterns
iii.	*(K-W test) Does the Kruskal-Wallace test indicate statistical differences in the response variable between each level at the final time step? (i.e., is the p-value <0.05?)*				
	no	yes	yes	yes	yes
iv.	*(M-W U test) How many pairs of levels are statistically different (p-value <0.05) according to the Mann-Whitney U test?*				
	0/3	37/45	7/10	3/6	2/3
*	*(Evaluation) Is the response variable sensitive to changes in the level for the factor?*				
	no	yes	yes	yes	no

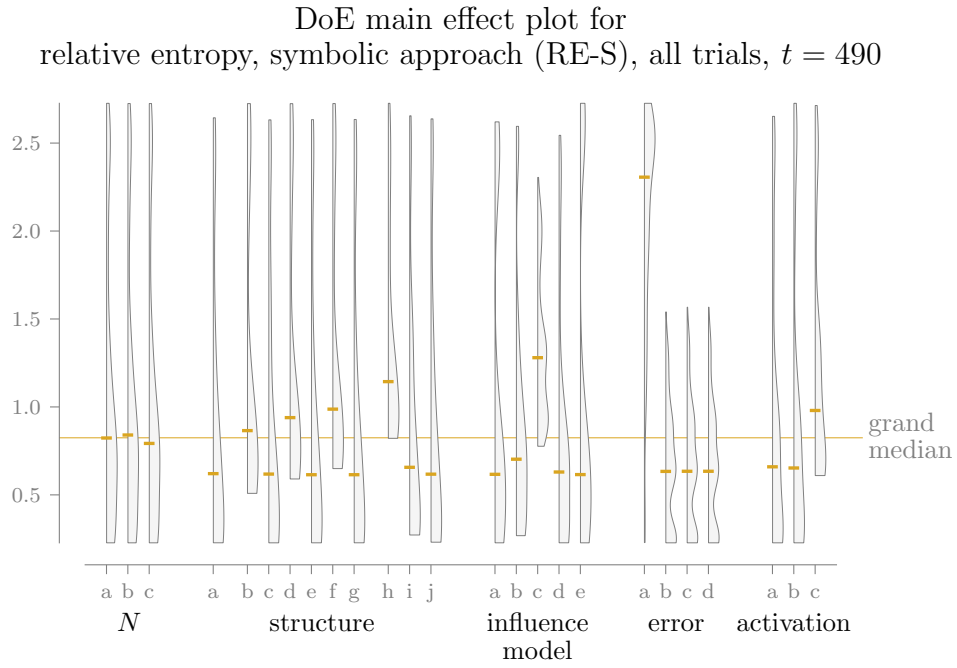

Figure 38: Each half-violin of this design of experiments (DoE) main effect plot represents the distribution of RE-S at $t = 490$ for all trials with the corresponding level on the horizontal axis, and its median is indicated with a horizontal dash; the grand median is shown for reference. This plot suggests that N is unimportant to the response variable, while the other factors lead to varied outcomes, to greater or lesser extents.

distributions among the levels for a single factor qualitatively show the effect each level has on the response. For example, the distributions for population size N are almost identical, so we infer that N is not important (i.e., does not have a significant effect on the response variable), at least over the range of levels used in the experimental design. On the other hand, strong differences between distributions are visible for the influence model and error term, marking them as important to RE-S. The results for the network structure shows four levels above the grand median line, which happen to be the four network structures with lower density (Appendix B).

Figure 38 uses data for only a single time step. To reveal the effect of time on the response variable, Figures 39-43 plot the medians of the grouped data over the full length of the simulation. The two inner quartiles (25th to 75th percentiles) are indicated by the shaded regions around each line. Overall,

median relative entropy, symbolic approach (RE-S), all trials, grouped by N

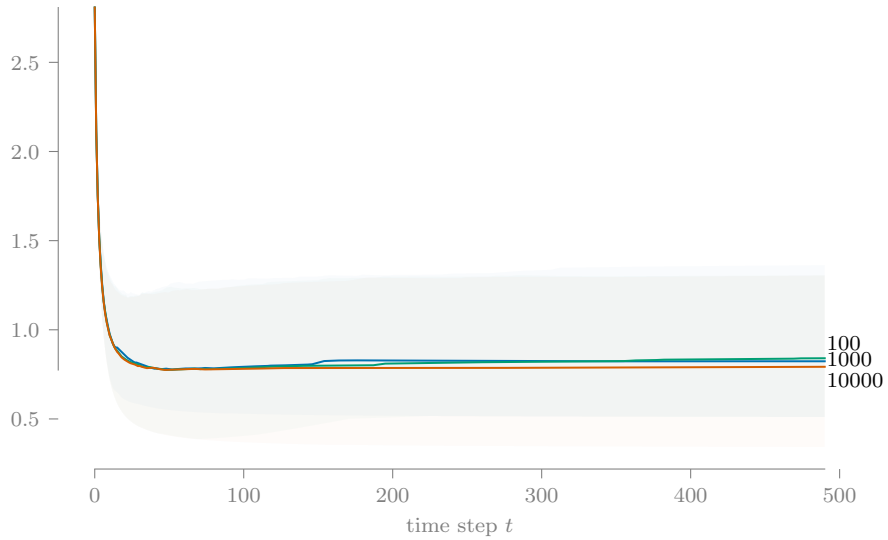

Figure 39: All trials are grouped by factor-level as in Figure 38 and the groups' median response value over time is plotted. Shaded regions around each line enclose the 25th to 75th percentiles of the data. For population size N, these regions almost entirely overlap due to the closeness of the median lines, which reinforces the low importance of N shown in the DoE main effect plot.

these figures reinforce the similarities and differences observed in the main effect plot.

Thus far, we have use qualitative approaches to show the effect of varying individual design factors. We now adopt a non-parametric approach to measuring differences between factor-levels, using the Kruskal-Wallace test and Mann-Whitney U test.

Based on the p-values from the Kruskal-Wallace test (Table 9) on RE-S at $t = 490$, when the data is split into levels for population size N, the data appears to come from the same population. Practically, this suggests that varying this factor—over the levels specified in our experiment—does not have a significant effect on the response variable. This agrees with what we observe in the previous figures.

Figure 44 aggregates the results of the Mann-Whitney U test applied to each pair of levels within a factor. Level pairs for network structure that are similar (grey cells) are almost all for pairs of higher density graphs.

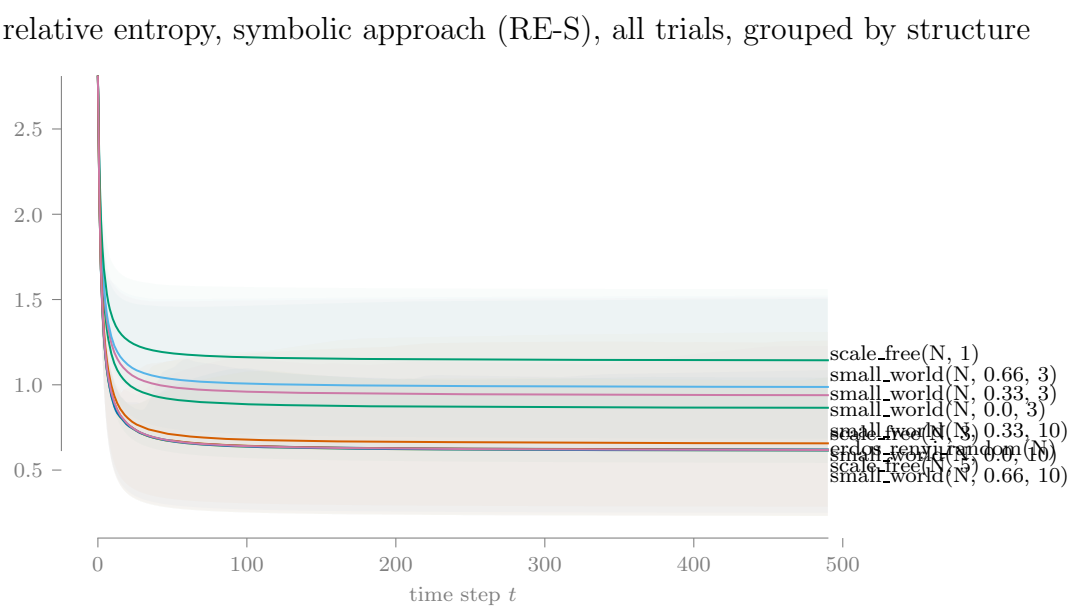

Figure 40: Following Figure 39 in design, this plot shows some differences in magnitude between network models but all have the same characteristic shape. Lower density structures have higher median values for RE-S.

Table 9: The Kruskal-Wallace test is ran on trial-level RE-S values at $t = 490$ to test if changing the level for a factor has a statistical effect on the response value. The asterisk indicates that population size N has no significant impact on RE-S.

	test stat	p-value
N	1.14	* 5.65e-01
structure	234.41	1.92e-45
influence model	231.76	5.51e-49
error	859.47	5.46e-186
activation	159.54	2.26e-35

median relative entropy, symbolic approach (RE-S), all trials, grouped by influence model

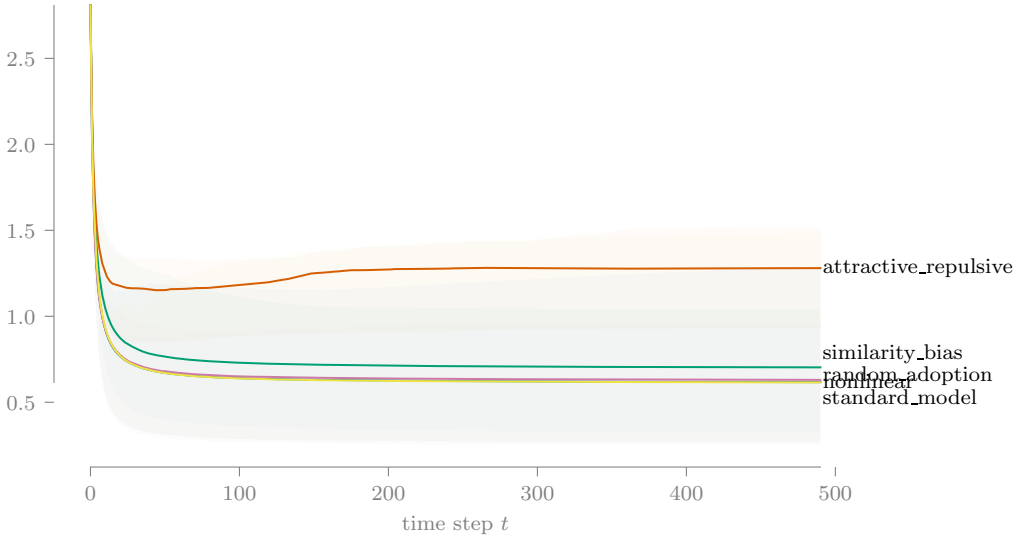

Figure 41: The attractive-repulsive model is clearly differentiated from the other influence models in both magnitude and shape of the median response line.

Overall, for the RE-S response variable, varying population size N has no significant effect; lower density and higher density network structure models lead to different outcomes; trials with no error term have much higher RE-S than any trial with an error term; the random activation regime yields somewhat higher RE-S; and results for each influence model, except attractive-repulsive, are quite similar.

1.4.2. Research Question 2: How is system design related to the response space of entropy time-series values?

Using the cluster analysis process described in Section 1.1, trials are assigned to clusters for both DTW and Pearson's correlation. These assignments are summarized in the following figures. Both DTW and Pearson's correlation produced two clusters for RE-S (Figures 46 and Figure 45). (Conventional guidance says that if multiple clustering methods call for the minimum/maximum number of clusters, then the selected methods or distance metric may be unsuitable for clustering the data.) With respect to the time series plots, DTW led to rather differentiated clusters, while Pearson's correlation did not.

median relative entropy, symbolic approach (RE-S), all trials, grouped by error

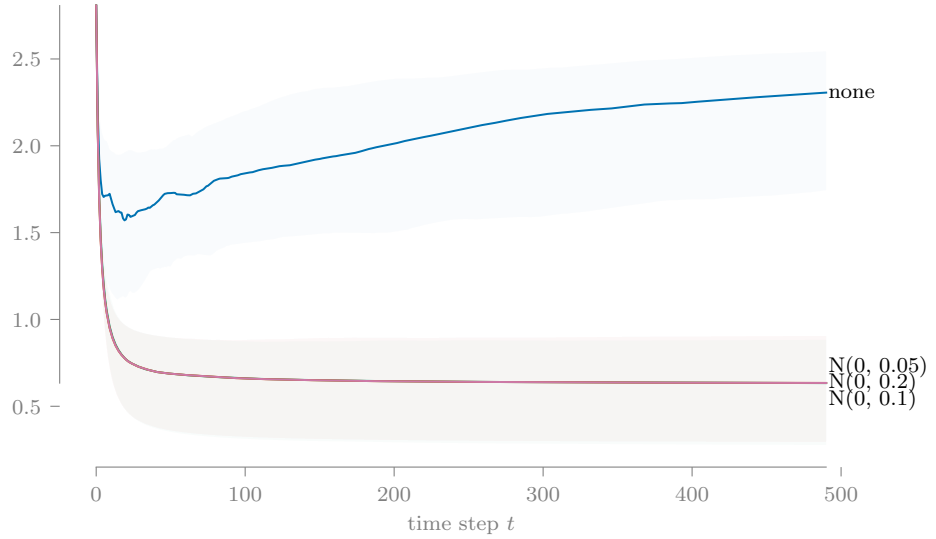

Figure 42: With the influence error distribution, we observe clear differences between no error term and the normally distributed error terms, while the median lines for the three normally distributed terms are indistinguishable.

median relative entropy, symbolic approach (RE-S), all trials, grouped by activation

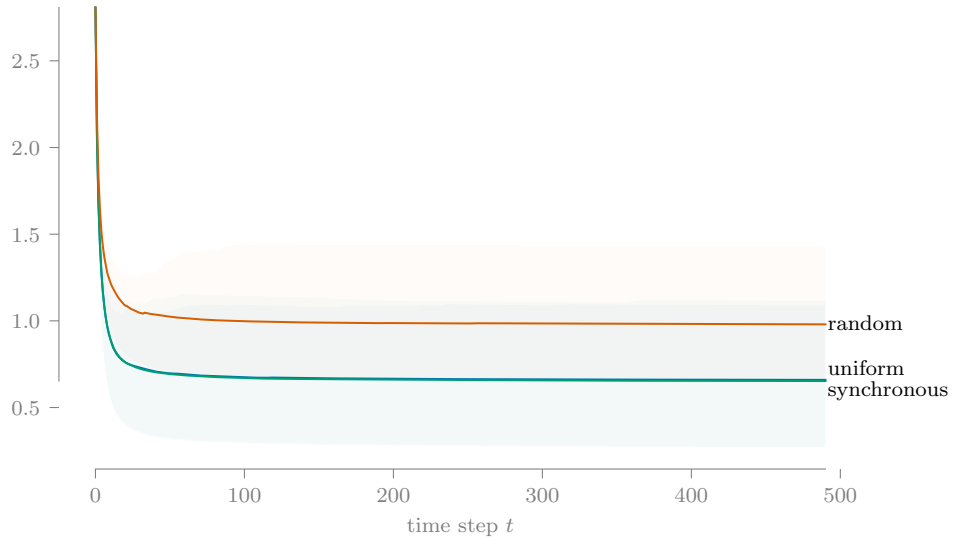

Figure 43: The median lines for the three activation regimes are very similar in shape, but the random regime converges to a slightly higher entropy value than the others.

In Figures 47 and 48, we conduct a "census" of the trials assigned to each cluster, with respect to the experimental design factors. For DTW, one cluster contains exclusively trials with error term a (no error). For Pearson's correlation, cluster membership is almost entirely homogeneous.

In summary, the variation in system design studied here does produce a somewhat meaningful cluster, with respect to the experimental design factors, in the response space for RE-S. This effect is achieved when using dynamic time warping as the distance measure, but not when using Pearson's correlation coefficient. However, in both cases, only two clusters were created, which limits the usefulness of cluster analysis for this response variable.

1.5. Response variable 5 - mutual information, symbolic approach (MI-S)

Mutual information, symbolic approach (MI-S) transforms each agent's sequence of opinion values into a pattern of relative orderings and computes the mutual information between each agent-neighbor pair, averages across the neighbors for each agent, and then averages across each agent and each replication to produce the trial-level response. Figure 49 shows the time

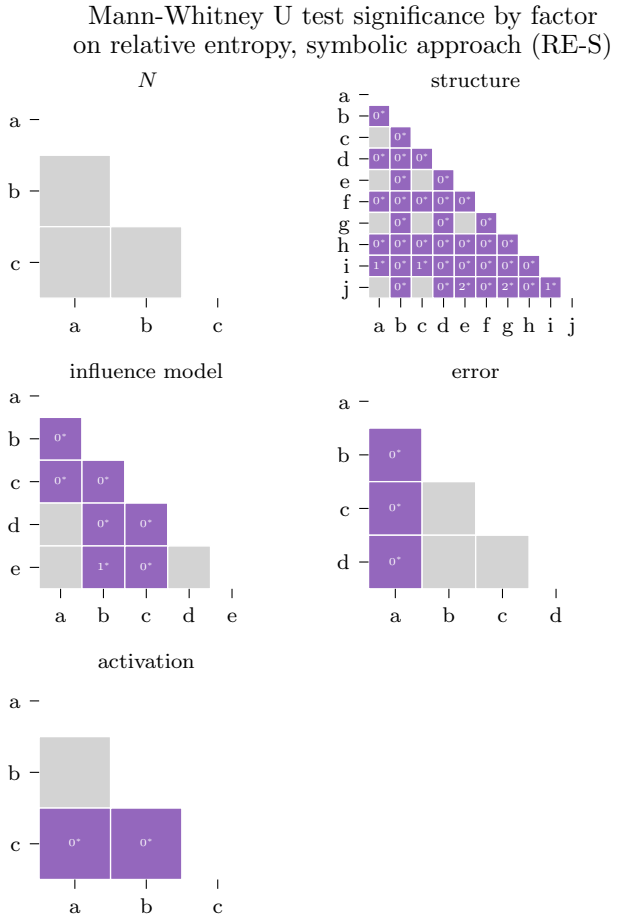

Figure 44: We use the Mann-Whitney U test to determine which levels are statistically different within each factor. The numbers in cells for the pairs with a significant test statistic (< 0.05) express the p-value as a percentage (e.g. 3* means $0.03 \leq$ p-value < 0.04). The non-significant results for N are consistent with the previous findings.

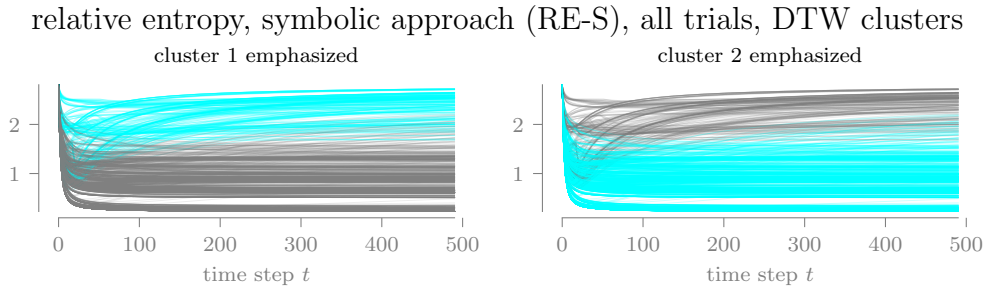

Figure 45: Using dynamic time warping (DTW) as the distance measure between pairs of response variable time series, the consensus method produces two clusters, each highlighted here using the original time series plot (Figure 37). The densely grouped nature of these clusters suggest a reasonable level of cluster quality.

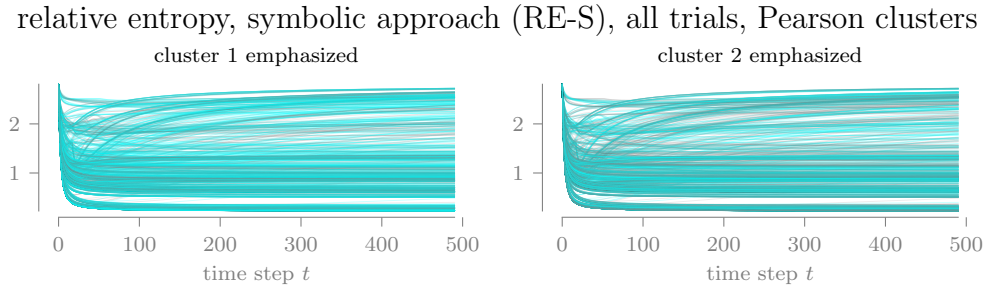

Figure 46: Using Pearson's correlation as the distance measure, the consensus method produces two clusters. The results show no clear pattern and may indicate less meaningful clusters.

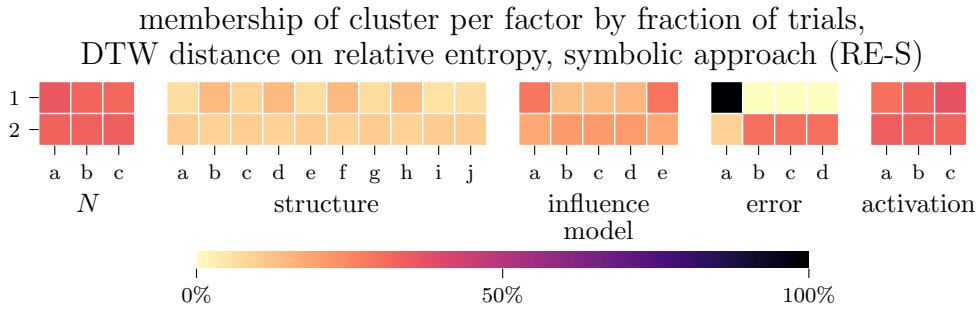

Figure 47: For each cluster produced through DTW, the trials assigned to the cluster are grouped by factor-level in order to find the percentage of a cluster associated with each factor-level. For example, all trials assigned to cluster 2 use error term a (no error) and have a slight preference for lower density networks (b, d, f, and h).

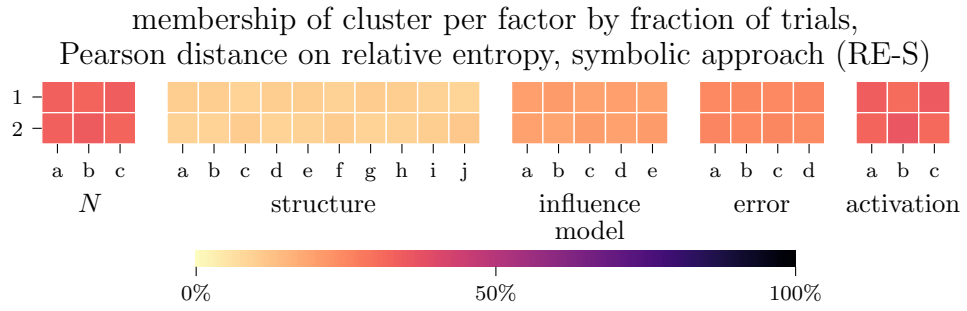

Figure 48: Clusters produced through Pearson's correlation are completely undifferentiated, suggesting this distance measure is unsuitable for the response variable.

series of MI-S for each trial and an associated kernel density estimate (KDE) for the final time step.[4]

1.5.1. Research Question 1: Which system design factors contribute most to system-level entropy?

For this research question, we explore the one-way sensitivity of each entropy response variable to changes in the levels of individual experimental design factors. This exploration includes qualitative comparisons of RV distributions when the trial data is grouped by factor-levels and statistical tests for differences between levels. These methods support a subjective evaluation of whether an RV is sensitive to changes in the level of each design factor. In Table 10, we summarize the analysis results for the current response variable for this research question.

Figure 50 presents the distributions grouped by factor-level for MI-S at the final time step, $t = 490$, using a half-violin plot. Differences between distributions among the levels for a single factor qualitatively show the effect each level has on the response. For example, the distributions for population size N are almost identical (except in the length of the upper tails), so we infer that N is not very important (i.e., does not have a significant effect on the response variable), at least over the range of levels used in the experimental design. On the other hand, strong differences between distributions

[4]Because the symbolic method uses multiple time steps per calculation, the final time steps of the simulation have no response value. We trim the data to $t = 490$ for ease of reading.

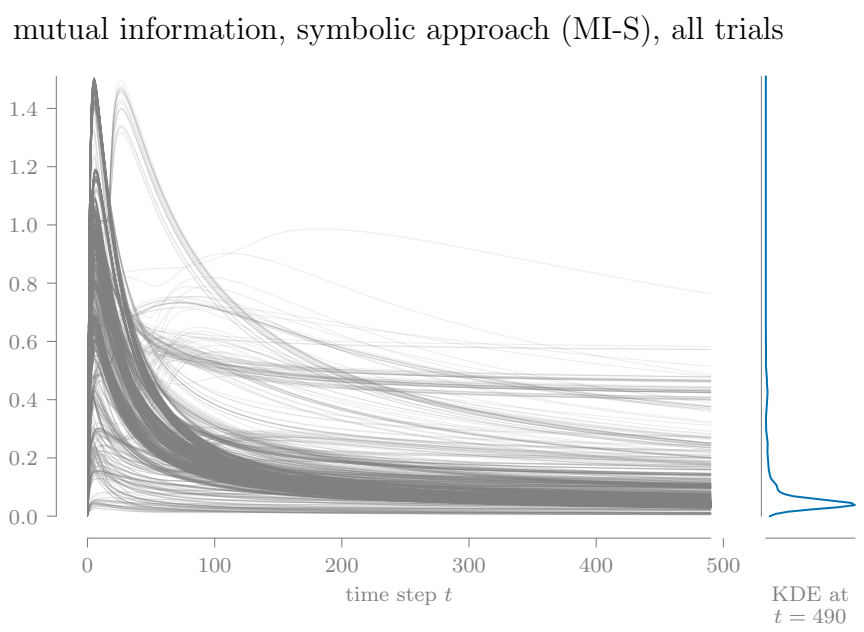

Figure 49: The time series of MI-S values for each trial are plotted on the same axes to reveal a dense visual cluster near the bottom of the range.

Table 10: The findings for Research Question 1 on MI-S are summarized to support the overall evaluation of each experimental design factor (final table row).

	Factor (number of levels)				
	N (3)	structure (10)	influence model (5)	error (4)	activation (3)
i.	*(Main effect plot) What differences are present between the response variable distributions for each level at the final time step?*				
	$N = 10k$ has much longer tail	2 or 3 patterns	5 patterns	2 patterns; error vs no error	3 patterns
ii.	*(Grouped time series) What differences are present between the median response values over the duration of the simulation?*				
	negligible	initial grouping by density, then partial convergence	initial separation, then partial convergence	no error somewhat different from rest	minor early variation, then partial convergence
iii.	*(K-W test) Does the Kruskal-Wallace test indicate statistical differences in the response variable between each level at the final time step? (i.e., is the p-value <0.05?)*				
	no	yes	yes	yes	yes
iv.	*(M-W U test) How many pairs of levels are statistically different (p-value <0.05) according to the Mann-Whitney U test?*				
	0/3	12/45	9/10	5/6	3/3
*	*(Evaluation) Is the response variable sensitive to changes in the level for the factor?*				
	no	yes	yes	yes	yes

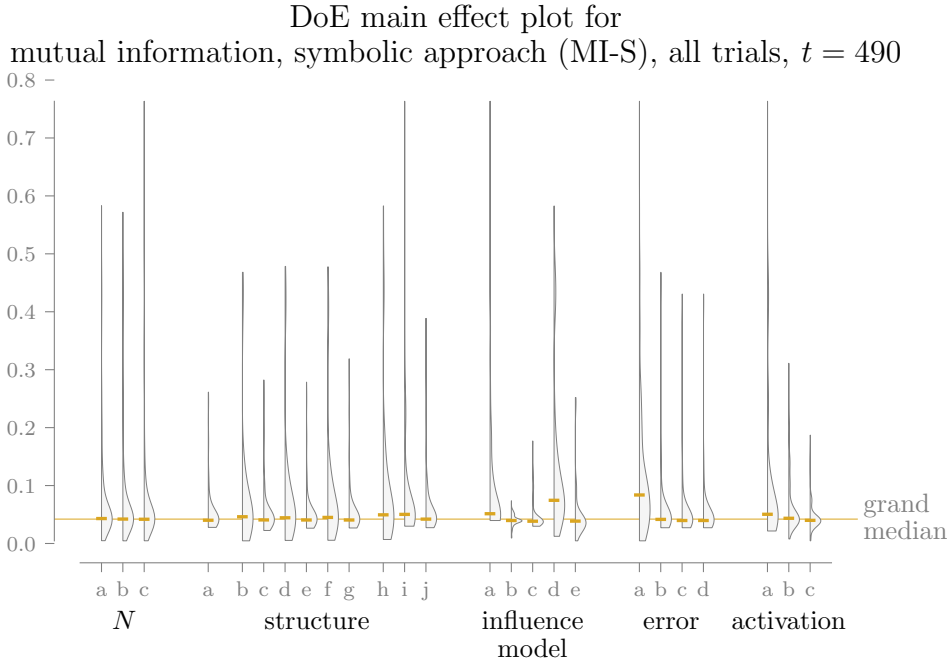

Figure 50: Each half-violin of this design of experiments (DoE) main effect plot represents the distribution of MI-S at $t = 490$ for all trials with the corresponding level on the horizontal axis, and its median is indicated with a horizontal dash; the grand median is shown for reference. This plot suggests that N is fairly unimportant to the response variable, while changes in influence model lead to more varied outcomes.

are visible for the influence model, marking it as important to MI-S.

Figure 50 uses data for only a single time step. To reveal the effect of time on the response variable, Figures 51-55 plot the medians of the grouped data over the full length of the simulation. The two inner quartiles (25th to 75th percentiles) are indicated by the shaded regions around each line. Overall, these figures reinforce the similarities and differences observed in the main effect plot.

Thus far, we have use qualitative approaches to show the effect of varying individual design factors. We now adopt a non-parametric approach to measuring differences between factor-levels, using the Kruskal-Wallace test and Mann-Whitney U test.

Based on the p-values from the Kruskal-Wallace test (Table 11) on MI-S at $t = 490$, when the data is split into levels for population size N, the data

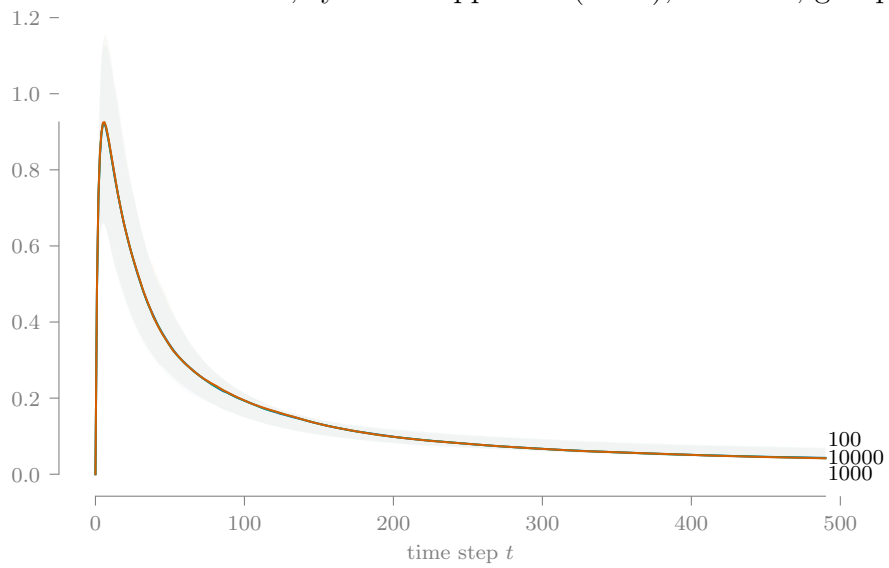

Figure 51: All trials are grouped by factor-level as in Figure 50 and the groups' median response value over time is plotted. Shaded regions around each line enclose the 25th to 75th percentiles of the data. For population size N, these regions almost entirely overlap due to the closeness of the median lines, which reinforces the low importance of N shown in the DoE main effect plot.

median mutual information, symbolic approach (MI-S), all trials, grouped by structure

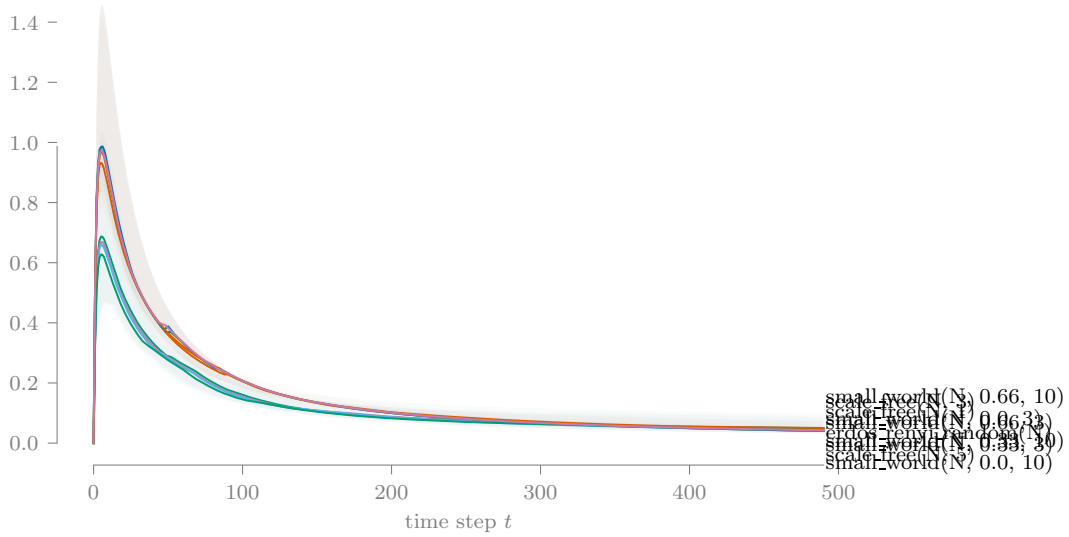

Figure 52: Following Figure 51 in design, an interesting split into two groups occurs during the initial transient between lower density (lower group) and upper density (upper group) networks, but all converge by the end.

median mutual information, symbolic approach (MI-S), all trials, grouped by influence model

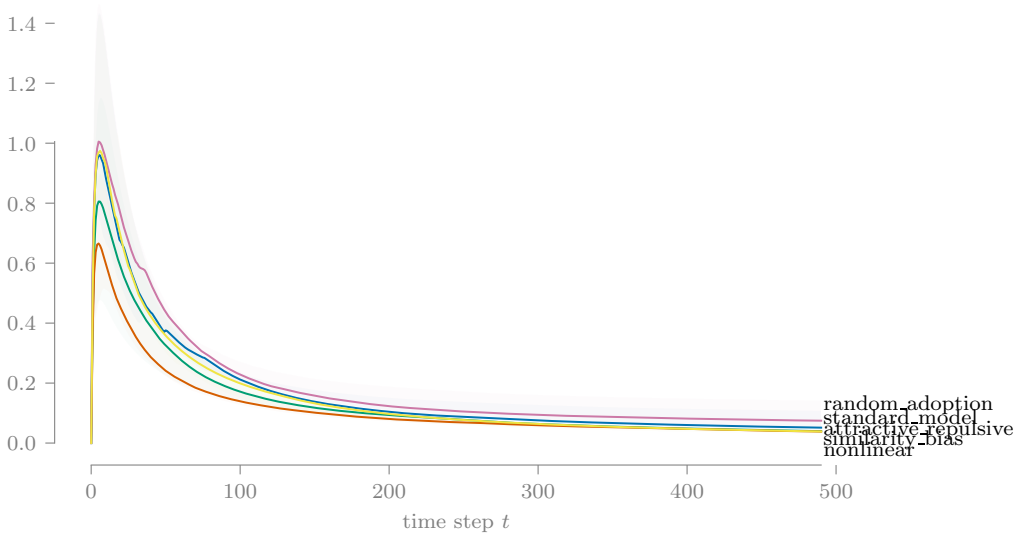

Figure 53: The various influence models show some initial differences, but all nearly converge by the end of the run.

appears to come from the same population. Practically, this suggests that varying this factor—over the levels specified in our experiment—does not have a significant effect on the response variable. This agrees with what we observe in the previous figures.

Figure 56 aggregates the results of the Mann-Whitney U test applied to each pair of levels within a factor. Only network structure levels i and j (the two higher-density scale-free networks) show any differentiation. Similarity bias (level b) and attractive-repulsive (c) influence models test as similar here, despite their very different functional forms; however, they do both use the magnitude of opinion difference to vary the strength of the interaction.

Overall, for the MI-S response variable, varying population size N and agent activation regime has no significant effect; density differences in network structure models have greater effects early in a run but very minor effects in the long term; and each influence model has a different response distribution but nearly identical median responses in the long term.

median mutual information, symbolic approach (MI-S), all trials, grouped by error

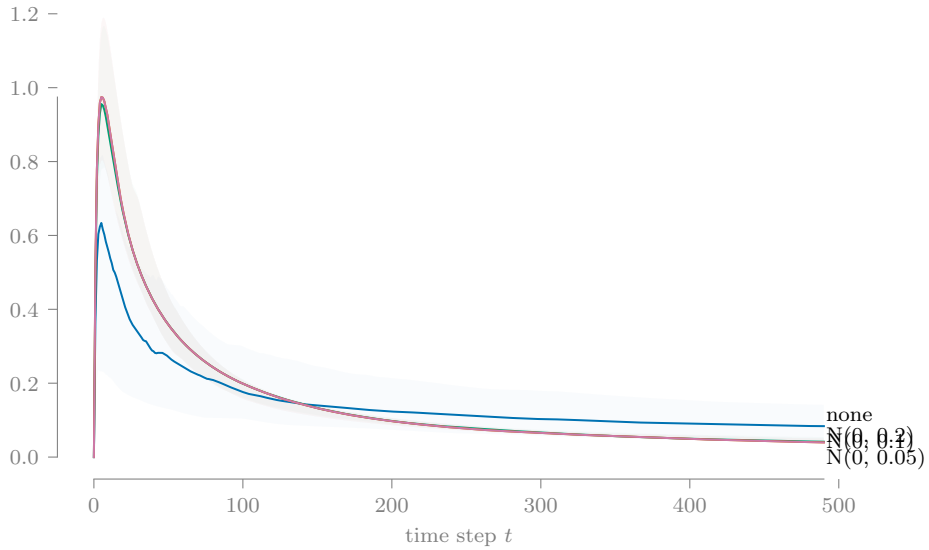

Figure 54: With the influence error distribution, we observe clear differences between no error term and the normally distributed error terms, while the median lines for the three normally distributed terms are indistinguishable.

Table 11: The Kruskal-Wallace test is ran on trial-level MI-S values at $t = 490$ to test if changing the level for a factor has a statistical effect on the response value. The asterisk indicates that population size N has no significant impact on MI-S.

	test stat	p-value
N	0.39	* 8.20e-01
structure	21.18	1.18e-02
influence model	720.43	1.30e-154
error	154.74	2.49e-33
activation	112.10	4.52e-25

median mutual information, symbolic approach (MI-S), all trials, grouped by activation

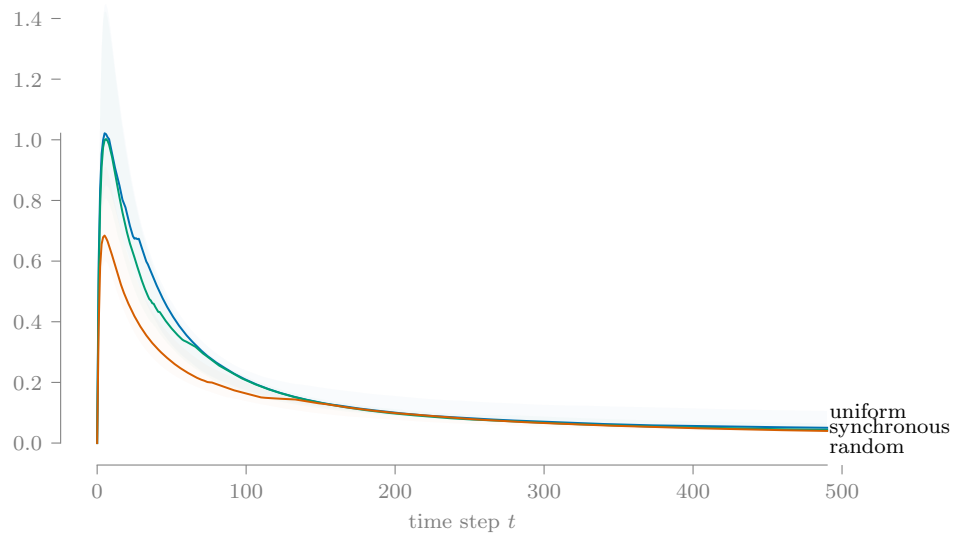

Figure 55: The median lines for the three activation regimes experience initial differences but rapidly converge.

1.5.2. Research Question 2: How is system design related to the response space of entropy time-series values?

Using the cluster analysis process described in Section 1.1, trials are assigned to clusters for both DTW and Pearson's correlation. These assignments are summarized in the following figures. DTW for MI-S produced five clusters (Figure 57), while Pearson's correlation produced two clusters (Figure 58). With respect to the time series plots, DTW led to somewhat differentiated clusters, while Pearson's correlation did not.

In Figures 59 and 60, we conduct a "census" of the trials assigned to each cluster, with respect to the experimental design factors. For DTW, cluster 2 contains exclusively trials with the synchronous activation regime (level a), while cluster 3 contains only trials with no error term (level a); cluster 5 omits random activation and lower density network structures. For Pearson's correlation, cluster membership is almost entirely homogeneous.

In summary, the variation in system design studied here does produce a somewhat meaningful cluster, with respect to the experimental design factors, in the response space for MI-S. This effect is achieved when using dy-

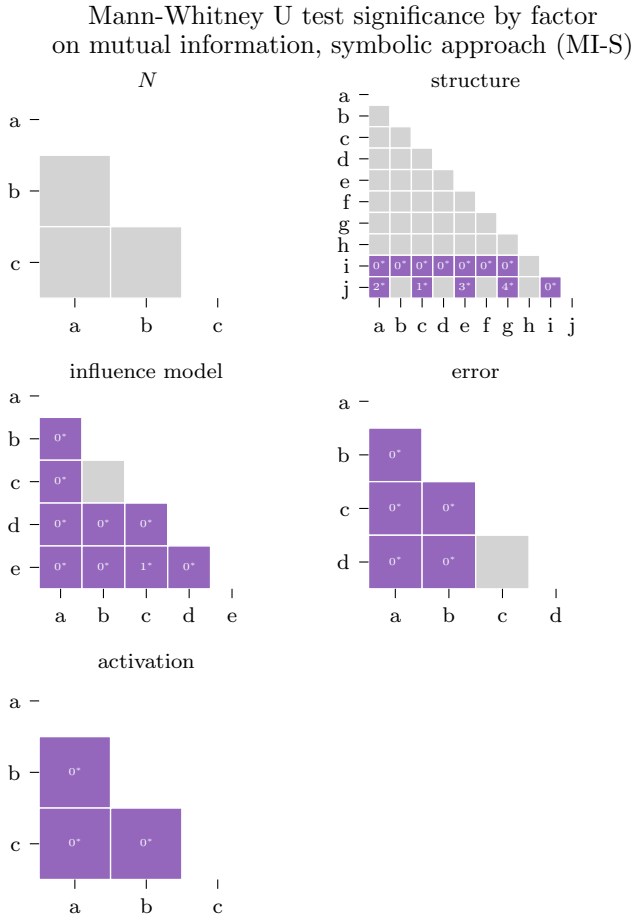

Figure 56: We use the Mann-Whitney U test to determine which levels are statistically different within each factor. The numbers in cells for the pairs with a significant test statistic (< 0.05) express the p-value as a percentage (e.g. 3* means $0.03 \leq$ p-value < 0.04). The non-significant results for N are consistent with the previous findings.

mutual information, symbolic approach (MI-S), all trials, DTW clusters

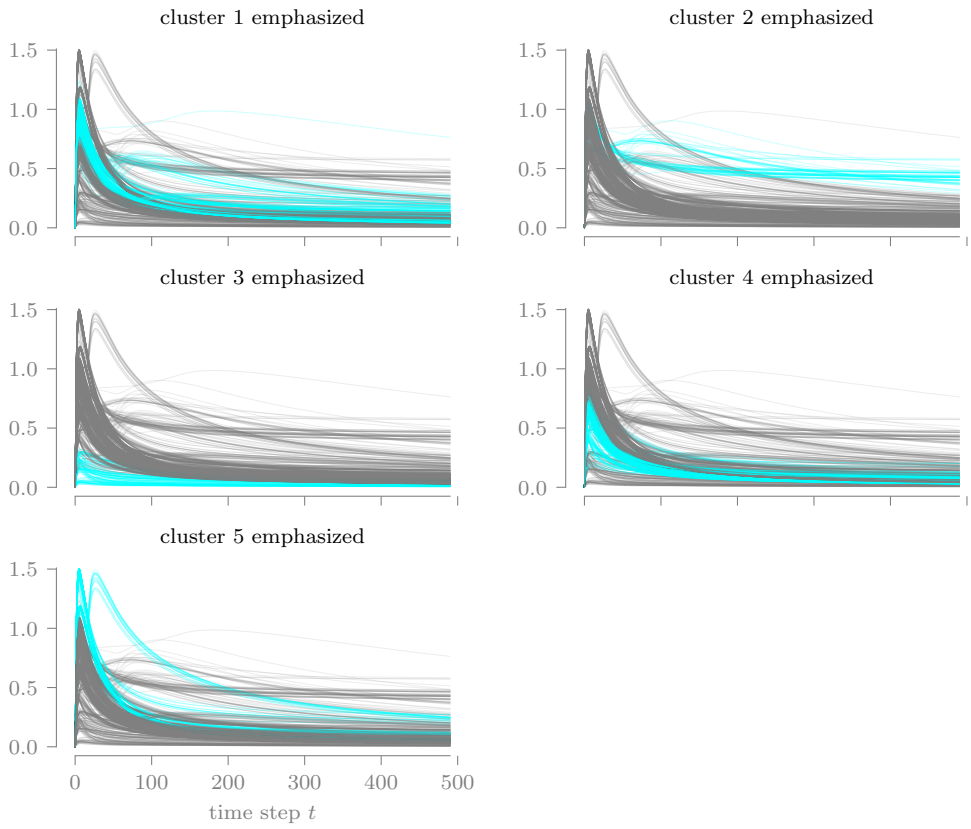

Figure 57: Using dynamic time warping (DTW) as the distance measure between pairs of response variable time series, the consensus method produces two clusters, each highlighted here using the original time series plot (Figure 49). The densely grouped nature of these clusters suggest a moderate level of cluster quality.

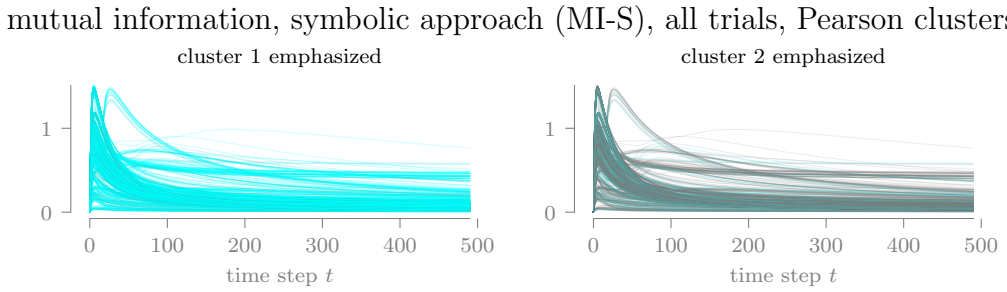

Figure 58: Using Pearson's correlation as the distance measure, the consensus method produces two clusters. The results show no clear pattern and may indicate less meaningful clusters.

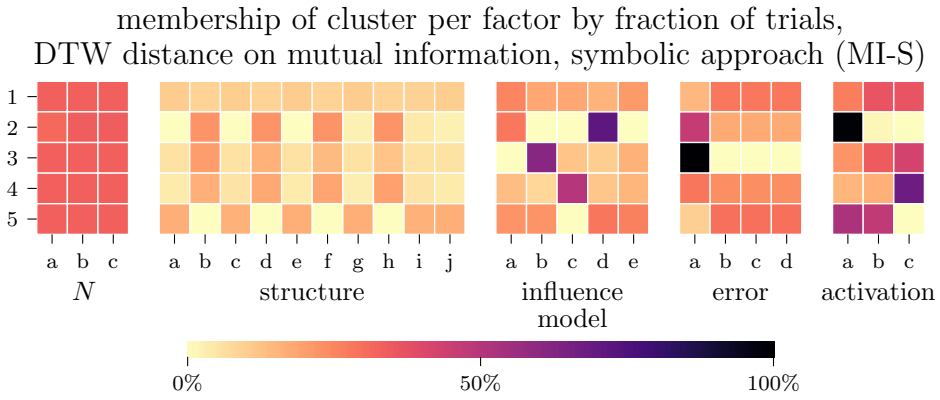

Figure 59: For each cluster produced through DTW, the trials assigned to the cluster are grouped by factor-level in order to find the percentage of a cluster associated with each factor-level. For example, all trials assigned to cluster 3 use error term a (no error).

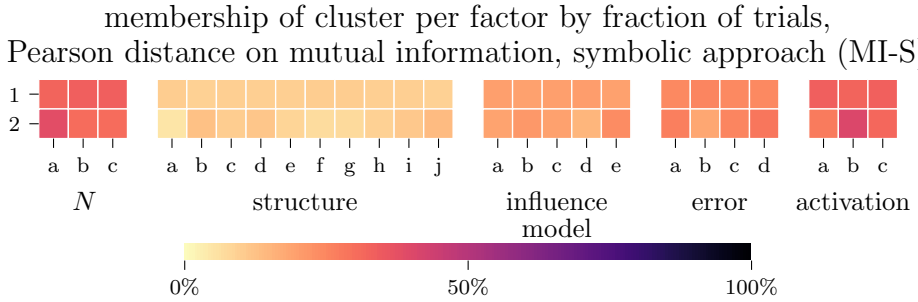

Figure 60: Clusters produced through Pearson's correlation are completely undifferentiated, suggesting this distance measure is unsuitable for the response variable.

namic time warping as the distance measure, but not when using Pearson's correlation coefficient.

1.6. Response variable 6 - transfer entropy, symbolic approach (TE-S)

Transfer entropy, symbolic approach (TE-S) transforms each agent's sequence of opinion values into a pattern of relative orderings and computes the transfer entropy between each agent-neighbor pair, averages across the neighbors for each agent, and then averages across each agent and each replication to produce the trial-level response. Figure 61 shows the time series of TE-S for each trial and an associated kernel density estimate (KDE) for the final time step.[5]

1.6.1. Research Question 1: Which system design factors contribute most to system-level entropy?

For this research question, we explore the one-way sensitivity of each entropy response variable to changes in the levels of individual experimental design factors. This exploration includes qualitative comparisons of RV distributions when the trial data is grouped by factor-levels and statistical tests for differences between levels. These methods support a subjective evaluation of whether an RV is sensitive to changes in the level of each design factor. In Table 12, we summarize the analysis results for the current response variable for this research question.

[5]Because the symbolic method uses multiple time steps per calculation, the final time steps of the simulation have no response value. We trim the data to $t = 490$ for ease of reading.

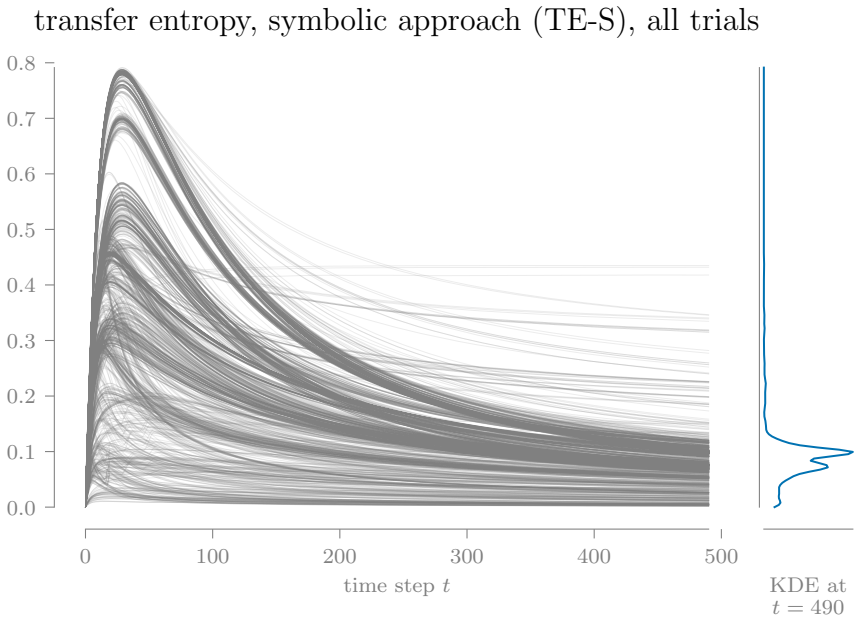

Figure 61: The time series of TE-S values for each trial are plotted on the same axes to reveal a dense visual cluster near the bottom of the range.

Table 12: The findings for Research Question 1 on TE-S are summarized to support the overall evaluation of each experimental design factor (final table row).

	Factor (number of levels)				
	N (3)	structure (10)	influence model (5)	error (4)	activation (3)
i.	*(Main effect plot) What differences are present between the response variable distributions for each level at the final time step?*				
	negligible	4 or 5 patterns	3 patterns	2 patterns; error vs no error	significant differences between all levels
ii.	*(Grouped time series) What differences are present between the median response values over the duration of the simulation?*				
	negligible	overall similar shapes; small divide affected by density	initial separation between levels, then partial convergence	no error significantly different from rest	random adoption initially different, then partial convergence
iii.	*(K-W test) Does the Kruskal-Wallace test indicate statistical differences in the response variable between each level at the final time step? (i.e., is the p-value <0.05?)*				
	no	yes	yes	yes	yes
iv.	*(M-W U test) How many pairs of levels are statistically different (p-value <0.05) according to the Mann-Whitney U test?*				
	0/3	25/45	10/10	3/6	3/3
*	*(Evaluation) Is the response variable sensitive to changes in the level for the factor?*				
	no	yes	yes	yes	yes

Figure 62 presents the distributions grouped by factor-level for TE-S at the final time step, $t = 490$, using a half-violin plot. Differences between distributions among the levels for a single factor qualitatively show the effect each level has on the response. For example, the distributions for population size N are almost identical (except in the length of the upper tails), so we infer that N is not very important (i.e., does not have a significant effect on the response variable), at least over the range of levels used in the experimental design. On the other hand, strong differences between distributions are visible for the influence model, marking it as important to TE-S. While the shape of the distributions for the lower density networks (b, d, f, and h) are inconsistent, they do happen to be the four levels with medians below the grand median.

Figure 62 uses data for only a single time step. To reveal the effect of time on the response variable, Figures 63-67 plot the medians of the grouped data over the full length of the simulation. The two inner quartiles (25th to 75th percentiles) are indicated by the shaded regions around each line. Overall, these figures reinforce the similarities and differences observed in the main

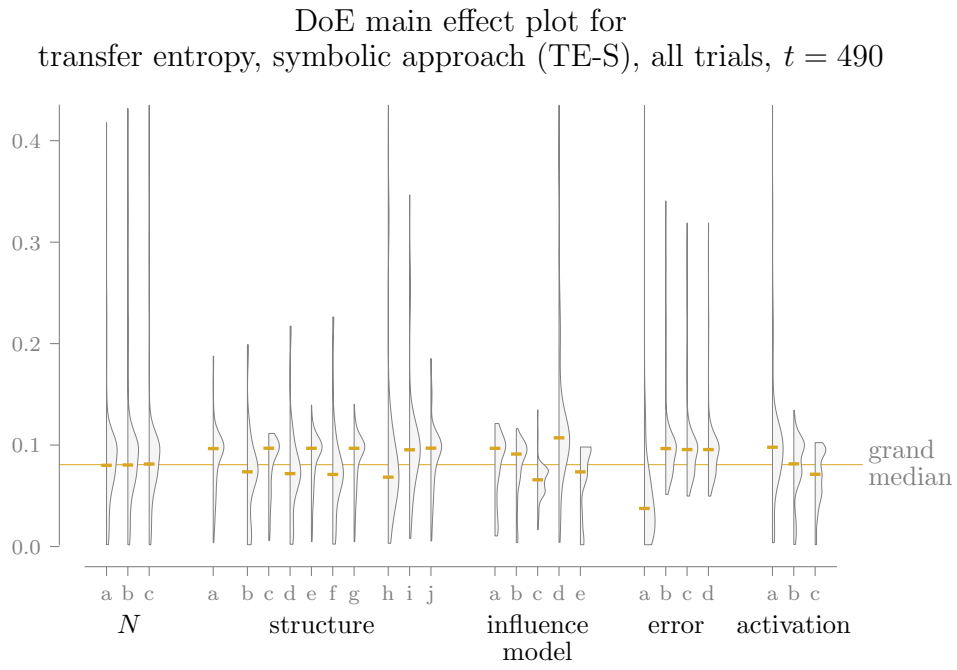

Figure 62: Each half-violin of this design of experiments (DoE) main effect plot represents the distribution of TE-S at $t = 490$ for all trials with the corresponding level on the horizontal axis, and its median is indicated with a horizontal dash; the grand median is shown for reference. This plot suggests that N is fairly unimportant to the response variable, while changes in influence model lead to more varied outcomes.

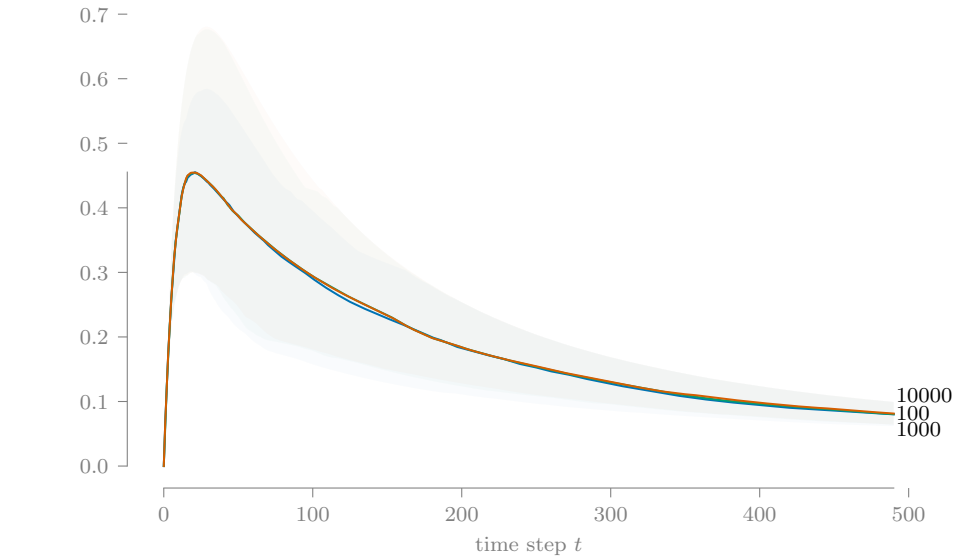

Figure 63: All trials are grouped by factor-level as in Figure 62 and the groups' median response value over time is plotted. Shaded regions around each line enclose the 25th to 75th percentiles of the data. For population size N, these regions almost entirely overlap due to the closeness of the median lines, which reinforces the low importance of N shown in the DoE main effect plot.

effect plot.

Thus far, we have use qualitative approaches to show the effect of varying individual design factors. We now adopt a non-parametric approach to measuring differences between factor-levels, using the Kruskal-Wallace test and Mann-Whitney U test.

Based on the p-values from the Kruskal-Wallace test (Table 13) on TE-S at $t = 490$, when the data is split into levels for population size N, the data appears to come from the same population. Practically, this suggests that varying this factor—over the levels specified in our experiment—does not have a significant effect on the response variable. This agrees with what we observe in the previous figures.

Figure 68 aggregates the results of the Mann-Whitney U test applied to each pair of levels within a factor. Structure pairs are primarily different if they are have different relative network densities (lower vice higher density, with respect to all network structures in the experiment).

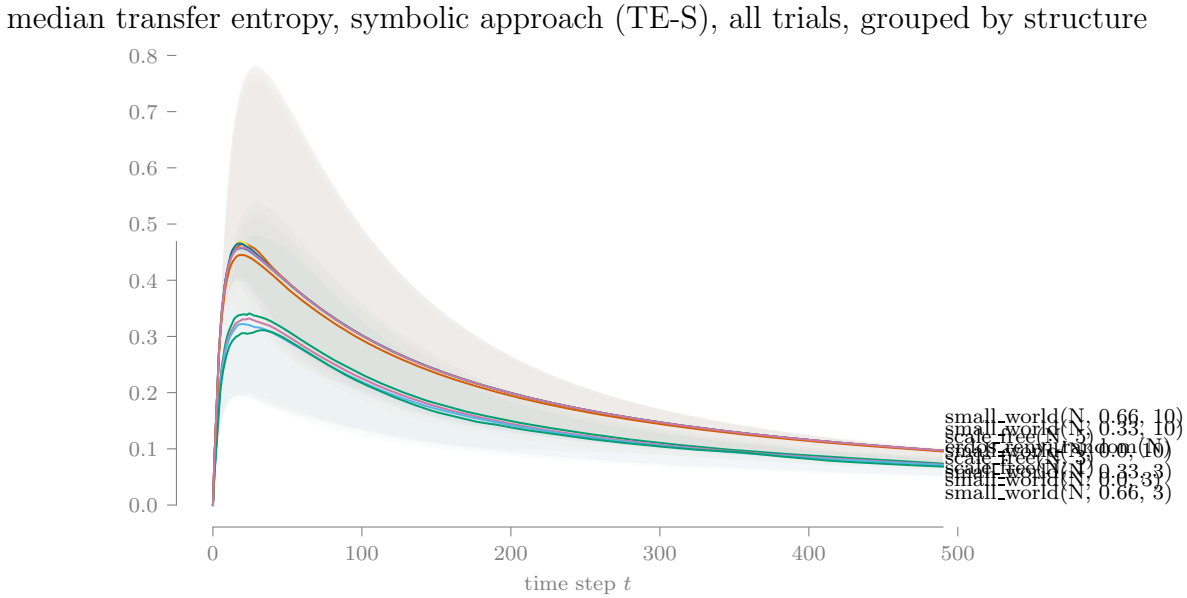

Figure 64: Following Figure 63 in design, this plot shows two groupings of network models: lower density (lower group) and higher density (upper group).

Table 13: The Kruskal-Wallace test is ran on trial-level TE-S values at $t = 490$ to test if changing the level for a factor has a statistical effect on the response value. The asterisk indicates that population size N has no significant impact on TE-S.

	test stat	p-value
N	0.23	* 8.91e-01
structure	133.57	2.17e-24
influence model	458.17	7.41e-98
error	558.63	9.35e-121
activation	206.61	1.36e-45

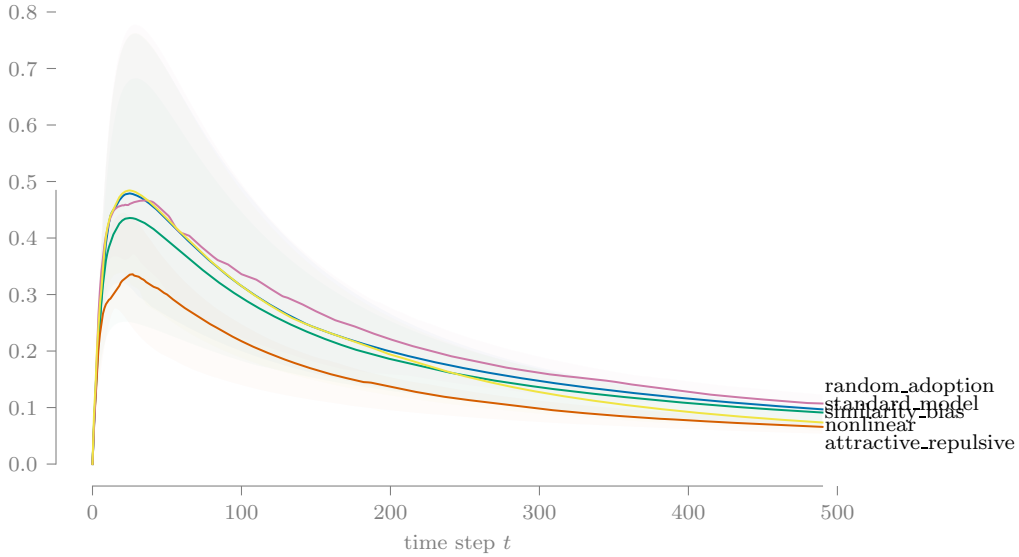

Figure 65: The various influence models show some initial differences, but begin to converge by the end of the run.

Overall, for the TE-S response variable, varying population size N has no significant effect, while varying all other factors produce noticeable differences in response distributions.

1.6.2. Research Question 2: How is system design related to the response space of entropy time-series values?

Using the cluster analysis process described in Section 1.1, trials are assigned to clusters for both DTW and Pearson's correlation. These assignments are summarized in the following figures. DTW for TE-S produced four clusters (Figure 69), while Pearson's correlation produced three clusters (Figure 70). With respect to the time series plots, DTW led to somewhat differentiated clusters, while Pearson's correlation did not.

In Figures 71 and 72, we conduct a "census" of the trials assigned to each cluster, with respect to the experimental design factors. For DTW, cluster 2 contains exclusively trials with no error term (level a) and is somewhat higher in tree-like network structures; cluster 1 is strong in the random activation regime (level c), while cluster 4 is strong in the other regimes and happens to omit lower density networks. For Pearson's correlation, cluster membership

median transfer entropy, symbolic approach (TE-S), all trials, grouped by error

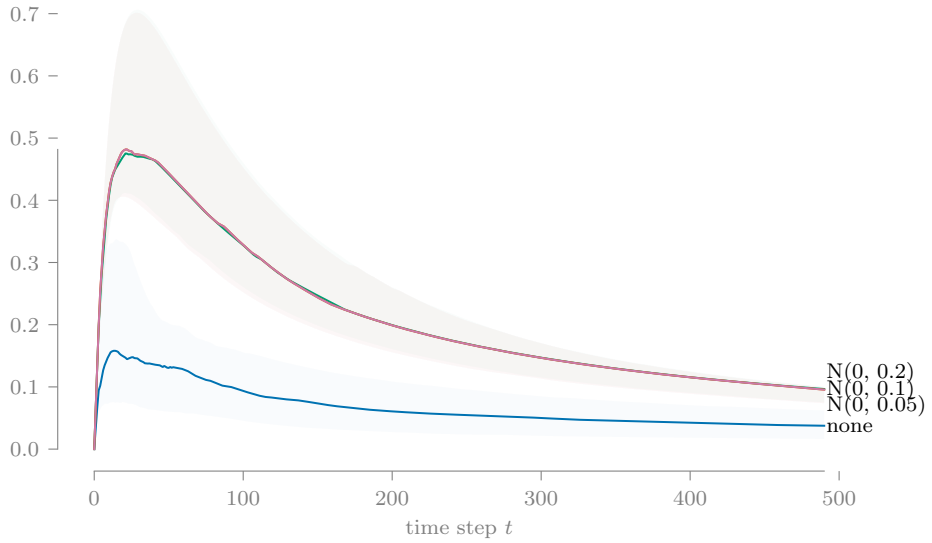

Figure 66: With the influence error distribution, we observe clear differences between no error term and the normally distributed error terms, while the median lines for the three normally distributed terms are indistinguishable.

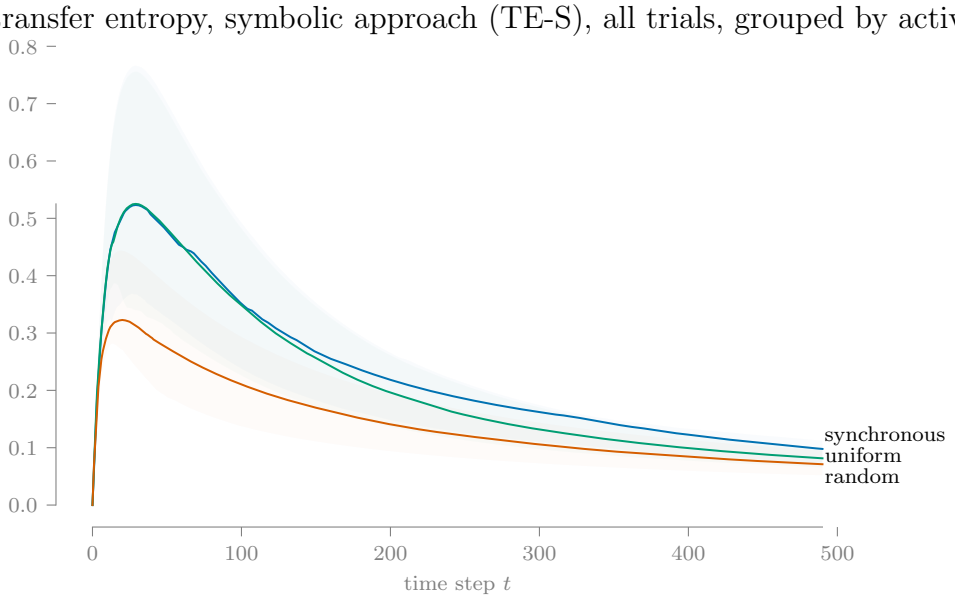

Figure 67: The median lines for the three activation regimes experience initial differences but begin to converge at the end of the run.

is almost entirely homogeneous.

In summary, the variation in system design studied here can produce meaningful clusters, with respect to the experimental design factors, in the response space for TE-S. This effect is achieved when using dynamic time warping as the distance measure, but not when using Pearson's correlation coefficient.

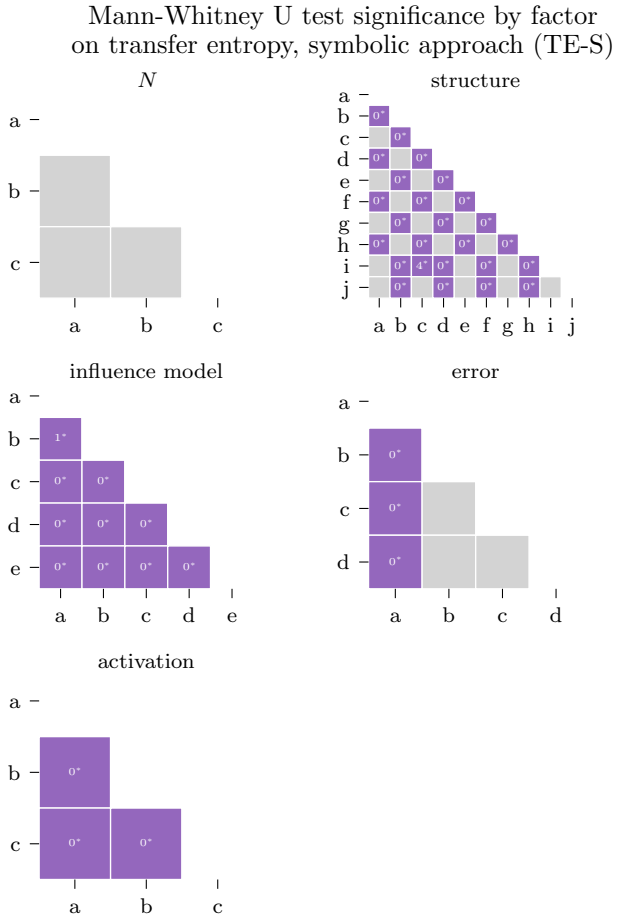

Figure 68: We use the Mann-Whitney U test to determine which levels are statistically different within each factor. The numbers in cells for the pairs with a significant test statistic (< 0.05) express the p-value as a percentage (e.g. 3* means $0.03 \leq$ p-value < 0.04). The non-significant results for N are consistent with the previous findings.

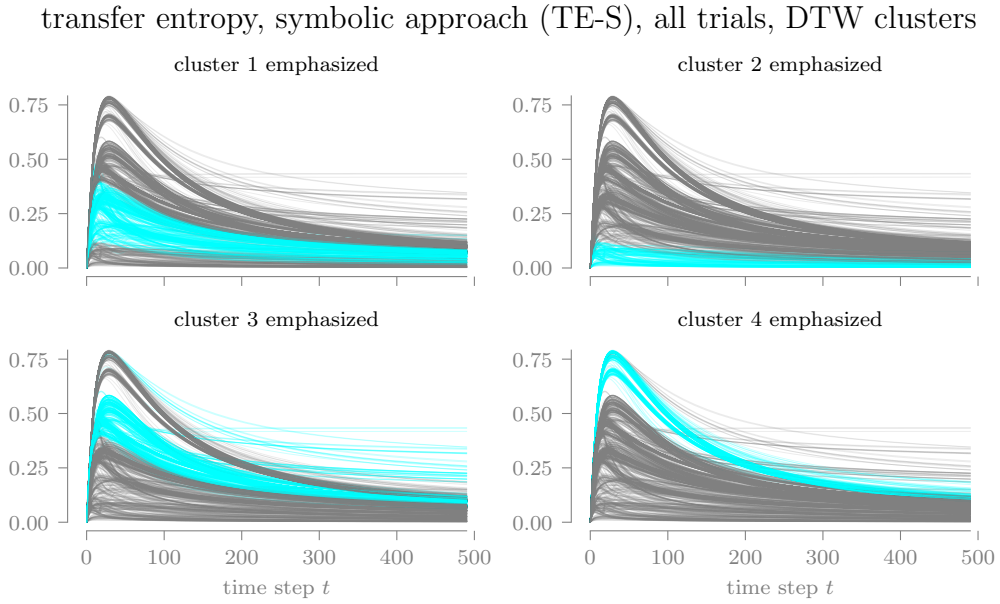

Figure 69: Using dynamic time warping (DTW) as the distance measure between pairs of response variable time series, the consensus method produces four clusters, each highlighted here using the original time series plot (Figure 61). The densely grouped nature of these clusters suggest a moderate level of cluster quality.

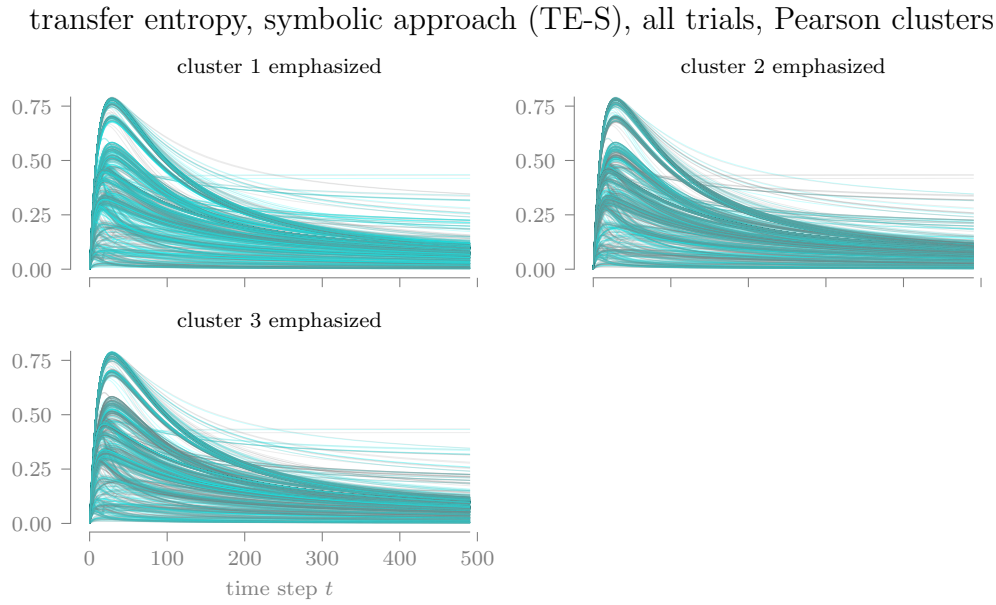

Figure 70: Using Pearson's correlation as the distance measure, the consensus method produces three clusters. The results show no clear pattern and may indicate less meaningful clusters.

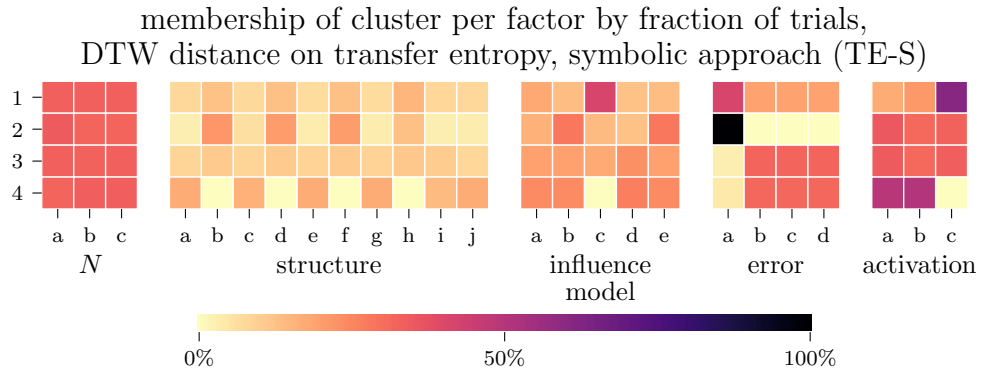

Figure 71: For each cluster produced through DTW, the trials assigned to the cluster are grouped by factor-level in order to find the percentage of a cluster associated with each factor-level. For example, all trials assigned to cluster 2 use error term a (no error).

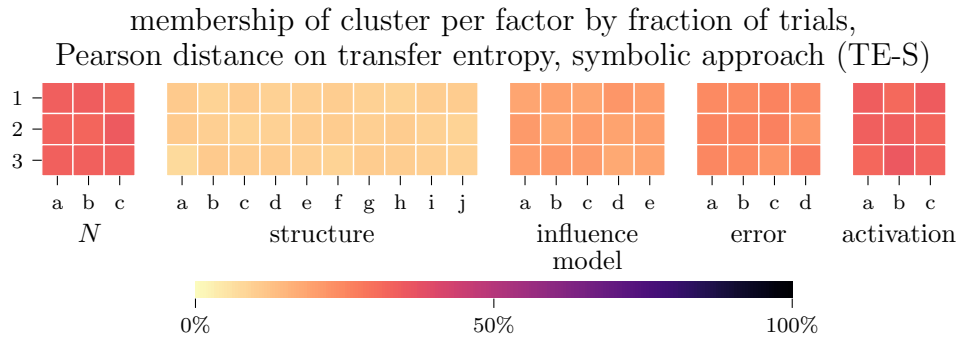

Figure 72: Clusters produced through Pearson's correlation are completely undifferentiated, suggesting this distance measure is unsuitable for the response variable.