Versatile Uncertainty Quantification of Contrastive Behaviors for Modeling Networked Anagram Games

Zhihao Hu[1], Xinwei Deng[1], and Chris J. Kuhlman[2(✉)]

[1] Virginia Tech, Blacksburg, VA 24061, USA
{huzhihao,xdeng}@vt.edu
[2] University of Virginia, Charlottesville, VA 22904, USA
hugo3751@gmail.com

Abstract. In a networked anagram game, each team member is given a set of letters and members collectively form as many words as possible. They can share letters through a communication network in assisting their neighbors in forming words. There is variability in behaviors of players, e.g., there can be large differences in numbers of letter requests, of replies to letter requests, and of words formed among players. Therefore, it is of great importance to understand uncertainty and variability in player behaviors. In this work, we propose versatile uncertainty quantification (VUQ) of behaviors for modeling the networked anagram game. Specifically, the proposed methods focus on building contrastive models of game player behaviors that quantify player actions in terms of worst, average, and best performance. Moreover, we construct agent-based models and perform agent-based simulations using these VUQ methods to evaluate the model building methodology and understand the impact of uncertainty. We believe that this approach is applicable to other networked games.

Keywords: Networked anagram games · Uncertainty quantification · Contrastive performance · Model explainability

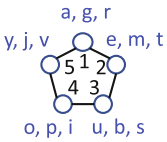
1 Introduction

1.1 Background and Motivation

Anagram games is a class of games where players are given a collection of letters and their goal is to identify the single word, or as many words as possible, that can be formed with these letters. Almost always, there is a time limit imposed on the game. Common anagram games include Scrabble and Boggle. In the literature, individual anagram games have been used to determine how players attribute their success or failure. It was found that players who performed well attributed their success to skill and those that performed poorly attributed their failure to bad luck [18]. Clearly, there can be heterogenous and contrastive behaviors among players of an anagram game.

R. M. Benito et al. (Eds.): COMPLEX NETWORKS 2021, SCI 1015, pp. 644–656, 2022.
https://doi.org/10.1007/978-3-030-93409-5_53

Our interest is networked group anagram games (NGAGs) [5,15], where players are arranged in a network configuration. They can share letters wherein players request letters from their neighbors and these neighbors decide whether or not to reply with the requested letters. The team's goal is to form as many words as possible. Figure 1 shows an illustrative game among five players, with initial letter assignments, and a sequence of player actions over a portion of the 5-min game (time is in seconds). In our experiments, each letter has infinite multiplicity: if player v_i shares a letter g with neighbor v_j, then v_i retains a copy of the letter that it shares with v_j. Note that the game configuration, and our analyses, account for agents (game players) with different degrees.



time	player	action	payload	description
4	2	request letter	g	2 requests g from 1
9	2	form word	meet	3 forms word "meet"
11	3	form word	bus	3 forms word "bus"
17	4	request letter	s	4 requests s from 3
27	3	reply letter	s	3 replies with s to 4
35	5	request letter	o	5 requests o from 4

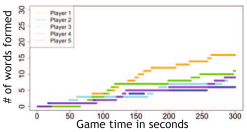


Fig. 1. (Left) Networked group anagram game (NGAG) configuration where each player (human subject) has three initial letters (in blue) and two neighbors. This configuration is a circle-5 graph, Circ_5. (Middle) Illustrative actions of players in the game. Players can use a letter any number of times, as evidenced by player 2 forming *meet* with letters m, e, and t. (Right) Experimental data on number of words formed by players in time; there is variability among player behaviors.

It is seen that the behaviors of players consist of multiple actions, i.e., requesting letters, replying to a request, forming words, or idle (thinking). Player performance is affected by these interactions. For example, the more letters a person has, the more words that she can presumably form. There are various uncertainties in terms of players' behaviors and the numbers of words formed in the team effort. Moreover, the heterogenous characteristics of players can involve contrastive behaviors, such as some players rarely requesting letters from their neighbors, while others request several. Therefore, it is of great importance to understand the uncertainty of the NGAG, and to have a flexible framework for quantifying the uncertainty of contrastive behaviors of players in the game.

In this work, our goals are to: (*i*) build explainable models of game player behaviors that quantify contrastive behaviors in terms of worst, average, and best performance based on game data; (*ii*) construct agent-based models (ABMs) and perform agent-based simulations (ABSs) using these models; and (*iii*) evaluate the model building methodology and understand the impact of uncertainty for these models.

Our exemplar is the NGAG, but the proposed approach can be used in other networked games (e.g., [4,6,11,12]) and with observational data (e.g., [21]) where human behavior data are collected. Such behaviors are notorious for having significant uncertainty across players [8,19,20]. Consequently, uncertainty quantification (UQ) methods are essential in complicated games like ours, where: (*i*) players have several *types* of actions that they can take (e.g., form word, request letter), (*ii*) players can take these actions many times throughout a game, and (*iii*) we seek to combine behaviors of a collection of game players in order to build one model of behavior (because there is insufficient data to generate a model from one player's game data).

One use of such models is analogous to the earlier work cited above. A human subject could be embedded in a networked anagram game where other players are bots. Skill levels of the bots are controlled (e.g., as all high performers or all worst performers, with the models developed in this study). The goal is to understand how players attribute their success or failure, in a group setting, for different pre-determined bot play performance values and network configurations. These types of questions—in group settings—are in the realm of social psychology [3]. This is analogous to individual anagram games wherein experimenters found that they could control solution times by varying letter order [13].

Novelty of Our Work. The novelty of this work unfolds as follows. First, different from the previous work in [10], our current method is not restricted to clustering of players to differentiate the heterogeneous behaviors of players. The key idea of the proposed method is leveraging the uncertainty of model parameters to quantify the uncertainty of players' behaviors. Specifically, we propose a novel approach of mapping model parameters to the probabilities of players' actions, to better represent the uncertainty of behaviors in the game.

Second, we propose a versatile uncertainty quantification (VUQ) framework to enable the quantification of contrastive behaviors in terms of worst, average, and best performance to better understand player behaviors based on game data. Different from the previous work in [10], the proposed framework takes advantage of the $(1-\alpha)$ confidence set of model parameters to enable the quantification of contrastive behaviors with appealing visualization. Third, we integrate the VUQ framework into an ABS framework.

1.2 Our Contributions and Their Implications

Our first contribution is a VUQ approach to UQ. The proposed VUQ method can be used to understand characteristics of a game, including contrastive behaviors, bot effects in the network, and demographic differences of players. We use a multinomial logistic regression model to characterize the probabilities π_{ij} of a player taking a particular action a_j at time $(t+1)$ based on a state vector and the player's action a_i at t. Uncertainty is embedded in a parameter matrix $\boldsymbol{B}^{(i)}$, as described in Equation (1). We employ a sampling technique that uses contours of $(1-\alpha)\times 100\%$ confidence regions to construct $\boldsymbol{B}^{(i)}$ in terms of π_{ij}. We thus quantify uncertainty with $\boldsymbol{B}^{(i)}$ via π_{ij}. Preliminaries (i.e., previous work that

serves as the point of departure for our methods here) are provided in Sect. 3 and our methods are presented in Sect. 4.

The second contribution is the models that result from the VUQ methods and their use in an agent-based modeling and simulation (ABMS) platform. From the data for a particular collection of players, we determine worst, average, and best player models. These models are integrated into an ABMS platform so that NGAGs can be simulated for any number of players with different levels of performance, any specified communication network, and different numbers of initial letter assignments per player. Our ABMS work and illustrative results are in Sect. 6. We demonstrate, for example, that the number of words formed by players in a game increases by 25% in going from worst to best behavior. This is an example of contrastive behavior: a contrast (difference) in results produced by differences in models.

Our third contribution is to illustrate important implications of the preceding two contributions. The proposed VUQ significantly enhances model transparency and model explainablity [14], as well as elaborates the impact of data quality (sufficiency versus scarcity). By mapping model parameter uncertainty to the uncertainty of players' behaviors in terms of possible game actions, our method provides a useful technique to make UQ more transparent in term of the players' behaviors. The use of a $(1-\alpha) \times 100\%$ confidence set provides a clear and simple tool to visualize the effects of the uncertainty by sampling multiple model parameters from the confidence set. This is similar in spirit to other sampling techniques used to generate graphs [7]. Moreover, the comprehensive quantification of uncertainty of contrastive behaviors surprisingly uncovers the impact of data scarcity and data sufficiency in the modeling and uncertainty quantification of data. We provide concrete examples in Sect. 5.

2 Related Work

Modeling of Group Anagram Games. An ABM was constructed from NGAG experiments in [5,15]. The model computed player behaviors in time. The model also accounted for the number of neighbors that a player (agent) had in the anagram interaction network. In [10], behavior models were made more parsimonious by clustering players based on experimental game data and their degrees k in the game network. Each model was based on the average behavior within a cluster. This current work differs from the above works in that we are analyzing each cluster to produce models for worst, average, and best performance behaviors per cluster. Hence, in this work, an agent's assigned model is based on its degree in the network, cluster number, and performance specification.

Uncertainty Quantification Methods and Analyses. Experimental uncertainty and parameter uncertainty are two common sources of uncertainty. Alam et al. [1] use design of experiments (DoE) to quantify experimental uncertainty and analyze sensitivity. Regression models and Bayesian approaches are commonly used to quantify parameter uncertainty. For example, Arendt et al. [2]

quantify uncertainty using gaussian process and the variance of posterior distribution. Simulation-based modeling and analyses are also used for uncertainty quantification [16,17]. In this work, we aim to quantify uncertainty in terms of both model parameters and players' behaviors.

3 Previous Models

Our models use a network configuration to capture communication among game players. A player's neighbors are the players at distance 1 in the NGAG (see Fig. 1). In our previous work [10], a clustering-based UQ method is used for building ABMs of human behavior in the NGAG. The UQ approach in [10] focuses on the following aspects. First, players are partitioned based on their activity in a game by creating two variables as $x_{engagement}$ and x_{word}. Here $x_{engagement}$ is the sum of the number of requests and number of replies of a player, and x_{word} is the number of words a player forms in a game. We conducted hypothesis testing, and the results showed that we can categorize players into two groups: those players with two neighbors (group $g = 1$) and those players with three or more neighbors (group $g = 2$). Second, players are further partitioned within each group. We used the k-means clustering method [9] to form four clusters based on the Bayesian information criterion. Specifically, we cluster player behaviors in terms of $x_{engagement}$ and x_{word}, so that players in a single cluster have similar numbers of actions in a NGAG. Third, player behaviors in a game are modeled and four variables are introduced in Equation (1) below: size $Z_B(t)$ of the buffer of letter requests that player v has yet to reply to at time t; number $Z_L(t)$ of letters that v has available to use (i.e., in hand) at t to form words; number $Z_W(t)$ of valid words that v has formed up to t; and number $Z_C(t)$ of consecutive time steps that v has taken the same action. Let $\mathbf{z} = (1, Z_B(t), Z_L(t), Z_W(t), Z_C(t))_{5\times 1}$. Let action a_1 represent thinking or idle, a_2 represent letter reply, a_3 represent requesting a letter, and a_4 represent forming a word. The multinomial logistic regression is used to model π_{ij}—the probability of a player taking action a_j at time $t+1$, given that the player took action a_i at time t—as

$$\pi_{ij} = \frac{\exp(\mathbf{z}'\boldsymbol{\beta}_j^{(i)})}{\sum_{l=1}^{4}\exp(\mathbf{z}'\boldsymbol{\beta}_l^{(i)})},\ j = 1, 2, 3, 4 \tag{1}$$

where $\boldsymbol{\beta}_j^{(i)} = (\beta_{j1}^{(i)}, \ldots, \beta_{j,5}^{(i)})'$ are the corresponding regression coefficients. For a given action a_i at time t, the parameters can be expressed as a matrix $\boldsymbol{B}^{(i)} = (\beta_{j,h}^{(i)})_{4\times 5}$ for i = 1, …, 4. Note that the estimation of $\boldsymbol{B}^{(i)}$ generates a corresponding transition probability matrix which quantifies the activity levels of players in the game. In previous work and this current work, we use all NGAG experimental data for parameter estimation.

As a result, one can infer the activity level of a player in a cluster based on its engagement and words [10]. However, there are two limitations of such an approach. First, if we cluster players using a large number of variables, then it

would be impossible to tell the activity level of a cluster based on a high dimension plot. Thus, the approach requires a more flexible method to infer which model parameters correspond to contrastive behaviors (i.e., worst, average, and best). Second, there are different levels of variability within clusters of players in going from worst to best behavior. The variabilities of some clusters may be small while the variabilities of other clusters may be large. Hence, it is important to quantify the within-cluster uncertainties and integrate them with the ABMs.

4 The Proposed VUQ Method

In a group anagram game, it is important to identify which players are more active and which players are less active, i.e., contrastive behaviors of players. As shown in our previous work [10], players have different behaviors in different clusters. Also, it is essential to quantify uncertainties of players within clusters.

To quantify the uncertainty within a cluster, one possible method is to start from the parameter $\boldsymbol{B}^{(i)}$ matrix. Since $\boldsymbol{B}^{(i)}$ is estimated from the multinomial logistic regression, it has an asymptotic normal distribution. Thus, we can draw random samples from the asymptotic normal distribution, and these random samples can represent the variability of that cluster. Moreover, because we are interested in player models with contrastive behaviors, we draw random samples on the contour of $(1-\alpha)\times 100\%$ confidence region. However, the sampled $\boldsymbol{B}^{(i)}$ matrices do not have a clear interpretation to quantify the corresponding activity levels (e.g., worst, average, best). While it is difficult to identify the activity level of an agent from the $\boldsymbol{B}^{(i)}$ matrix, it is easy to identify the activity level from the probability vector $\boldsymbol{\pi} = (\pi_{i1}, \ldots, \pi_{i4})$. It is known than an agent is more active if the to-idle probability (π_{i1}) is small and less active if the to-idle probability is large. To obtain the probability vector, we need the $\boldsymbol{z}$ vector. Thus, we use NGAG data (i.e., training data) to produce representative $\boldsymbol{z}$ vectors. By using these $\boldsymbol{z}$ vectors, we can compute a set of probability vectors and calculate the mean of to-idle probabilities. Then the mean to-idle probability is used to quantify the activity level of an agent via the $\boldsymbol{B}^{(i)}$ matrix. The following steps summarize the proposed method of uncertainty quantification within a cluster.

We first transform the estimated $\hat{\boldsymbol{B}}^{(i)}$ matrix to vector $\hat{\boldsymbol{\beta}}^{(i)}$, then use the asymptotic normal distribution of parameter estimators $\hat{\boldsymbol{\beta}}^{(i)}$ based on the asymptotic property of maximum likelihood estimators. The superscript i denotes the initial state. Let $\hat{\boldsymbol{B}} = \hat{\boldsymbol{\beta}}^{(i)}$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(i)} = (\hat{\boldsymbol{\beta}}_2^{(i)T}, \hat{\boldsymbol{\beta}}_3^{(i)T}, \hat{\boldsymbol{\beta}}_4^{(i)T})^T$, then $\hat{\boldsymbol{\beta}}$ follows a multivariate normal distribution, $\hat{\boldsymbol{\beta}} \sim \mathrm{MN}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$.

1. **Step 1:** Draw R random samples ($\boldsymbol{\beta}_r$ or $\boldsymbol{B}_r$, where $r = 1, \ldots, R$) on the $(1-\alpha)\times 100\%$ confidence contour of the estimated $\hat{\boldsymbol{\beta}}$ matrix. The $(1-\alpha)\times 100\%$ confidence region $S_{\hat{\boldsymbol{\beta}}}$ is defined as

$$Pr(\boldsymbol{\beta} \in S_{\hat{\boldsymbol{\beta}}}) = (1-\alpha)\times 100\%, \tag{2}$$

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \chi_d^2(1-\alpha), \tag{3}$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, and $\chi^2_d(1-\alpha)$ is the $(1-\alpha)$ quantile of Chi-squared distribution with d degrees of freedom.

2. **Step 2:** For each $\boldsymbol{\beta}_r$ drawn in step 1, apply the training data to Eq. 1 to produce n probablity vectors, $\hat{\boldsymbol{\pi}}^{r,l}$, where $l = 1, \ldots, n$ and n is the size of the training data. Then calculate the mean probability, $\bar{\boldsymbol{\pi}}^r = \frac{1}{n}\sum_{l=1}^{n} \hat{\boldsymbol{\pi}}^{r,l} = (\bar{\pi}_1^r, \bar{\pi}_2^r, \bar{\pi}_3^r, \bar{\pi}_4^r)^T$. The mean of to-idle probability is denoted as $\bar{\pi}_1^r = \frac{1}{n}\sum_{l=1}^{n} \hat{\pi}_1^{r,l}$. Then we get a set of mean to-idle probabilities, $\bar{\pi}_1^r,\ r = 1, \ldots, R$.
3. **Step 3:** The mean to-idle probabilities represent the variability within the cluster. The $\boldsymbol{\beta}_r$ vector or $\boldsymbol{B}_r$ matrix with low mean to-idle probability is more active, and the $\boldsymbol{\beta}_r$ or $\boldsymbol{B}_r$ matrix with high mean to-idle probability is less active. The $\boldsymbol{B}_r$ matrix with the maximum $\bar{\pi}_1^r$ is selected as the worst matrix, and the $\boldsymbol{B}_r$ matrix with the minimum $\bar{\pi}_1^r$ is selected as the best matrix.

One advantage of this proposed method is that we can quantitatively compare the activity levels of two clusters using the mean to-idle probability. Previously, we relied on $x_{engagement}$ and x_{word}. The second advantage is that this method can be generalized in two aspects. First, currently we quantify the activity level of a cluster. However, we can easily quantify the activity level of a player using the same method. Thus, we can compare activity levels among different players/agents. Second, we use the mean to-idle probability as the criterion for activity level. We can easily use other criterion based on our needs or goals. For example, if we are interested in the activity level of forming words, we can use the mean to-word probability (π_{i4}) as the criterion.

Note that there can be data scarcity in some clusters, in which case the distribution of $\boldsymbol{B}_r$ would have very large variance. For example, if there is only one to-request transition in the game data, then the estimated parameter for to-request in $\boldsymbol{B}_r$ can be extremely large. Then the probability of the to-request transition could become close to 1 or 0. If the initial state is request and the probability of to-request is close to 1, then the model will fall into a request-to-request "infinite" loop. This potential issue of data scarcity is avoided as follows. First, if the minimum $\bar{\boldsymbol{\pi}}^r(\text{idle})$ is very small ($\bar{\boldsymbol{\pi}}^r(\text{idle}) < 0.01$), then we choose the $\boldsymbol{B}_r$ in which $\bar{\boldsymbol{\pi}}^r(\text{idle})$ is at the 10% percentile, instead of the minimum. Second, the worst and best $\boldsymbol{B}$ matrix matrices are replaced by the mean $\boldsymbol{B}$ matrix if any of these criteria are met: (i) all of the numbers of to-reply, to-request, and to-word transitions are less than 5, or (ii) extremely large values appear in $\boldsymbol{B}$.

5 Model Evaluation

In this section, the variabilities within clusters will be investigated and presented in terms of mean transition probabilities. For each group, cluster, and initial state, 1000 random $\boldsymbol{B}_r, r = 1, \ldots, 1000$ matrices are draw from the 95% confidence contour $S_{\hat{\boldsymbol{\beta}}}$. Then the mean transition probabilities are calculated using the training data in which initial states are the same as those of the $\boldsymbol{B}_r$ matrices. The histogram of mean transition probabilties are presented in Figs. 2 and 3.

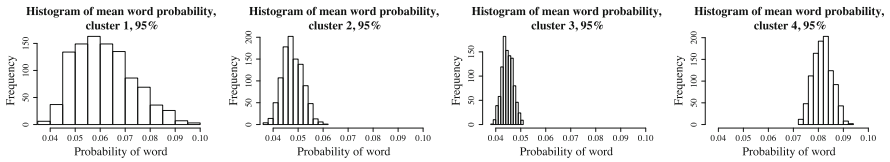


Fig. 2. Histograms of mean **to-word** probabilities of random $\boldsymbol{B}$ matrices for the four clusters in **group 1**. The **initial state is idle** and the $\boldsymbol{B}$ matrices samples are drawn on the contour of the **95%** confidence region. The plots from left to right are for cluster 1, 2, 3, and 4.

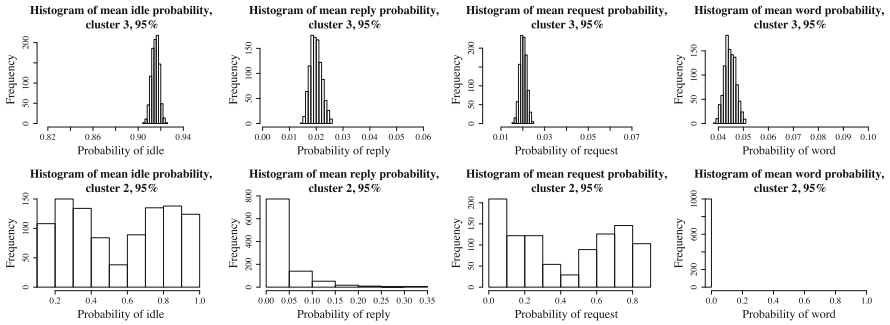


Fig. 3. The top four histograms are for **group 1, cluster 3** where the initial state is **idle** and the bottom four histograms are for **group 1, cluster 2** where the initial state is **request**. The $\boldsymbol{B}$ matrices samples are drawn on the contour of the **95%** confidence region. The plots in the first column are **to-idle**, the plots in the second column are **to-reply**, the plots in the third column are **to-request**, and the plots in the fourth column are **to-word**.

Figure 2 reports the histograms of mean to-word probabilities of random $\boldsymbol{B}$ matrices for clusters in group 1. From the histograms in Fig. 2, it is clear that there are within-cluster variabilities in terms of forming words. It further confirms the need of VUQ for quantifying the contrastive behaviors of players. Figure 3 shows the histograms of mean transition probabilities, where the top panel is for the group 1, cluster 3 with initial state being idle and the bottom panel is for group 1, cluster 2 with initial state being request. It is seen that the top four histograms show small variability because sufficient data are available. However, the data can be insufficient in some other cases as reported in the seventh row in Table 1, where there are very few data points for the player actions of reply, request, word. As shown in the bottom four plots of Fig. 3, the variability becomes very large and even ranges from 0 to 1, which is not realistic. Therefore, the mean $\boldsymbol{B}$ matrices are used in this case for both the worst and best behaviors. Such a limitation of the proposed method is due to data scarcity in the numbers of some actions, where the variability of the model parameters becomes unrealistically large. These issues illustrate the importance of model transparency. A summary of actions for some clusters are shown in the Table 1. It is seen that the majority of clusters have sufficient data while some encounter data scarcity in some actions.

Table 1. Summary of actions in selected group, cluster, and initial state triples. The first column is the group ID, with value 1 or 2. The second column is the cluster ID, which ranges from 1 to 4. The third column is the initial state, which ranges from 1 to 4 (idle, reply, request, word). The next four columns are the number of actions by players in games. For example, the number in the idle column is the number of to-idle actions. The last column shows the data sufficiency.

group	cluster	initial_state	idle	reply	request	word	data sufficiency
1	1	1	8311	28	70	230	sufficient data
1	3	1	17399	259	366	802	sufficient data
1	4	1	8593	110	185	993	sufficient data
2	2	1	17310	282	413	902	sufficient data
2	3	4	801	16	0	44	sufficient data
2	4	1	1572	33	71	344	sufficient data
1	2	3	306	2	3	0	data scarcity
2	1	4	199	1	0	0	data scarcity

6 Simulations and Results

6.1 Simulation Parameters and Process

We use the models from Sect. 4 to develop ABMs for players in the NGAG. We confine this work to the networked game configuration of Fig. 4, which is a circle graph on six players. All players have behaviors in group 1 because all players have degree $k = 2$. The symmetry of the setup enables us to assess variability in simulation results. Table 2 contains the parameters in simulations. We limit our simulation conditions owing to space limitations; the simulation system can handle agents of any group and cluster. Also, we use a graph structure that is similar to the network configurations in the NGAG.

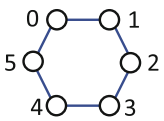


Fig. 4. Graph of six anagram game players, each with two neighbors (Circ_6).

Table 2. Summary of the parameters and their values used in the simulations.

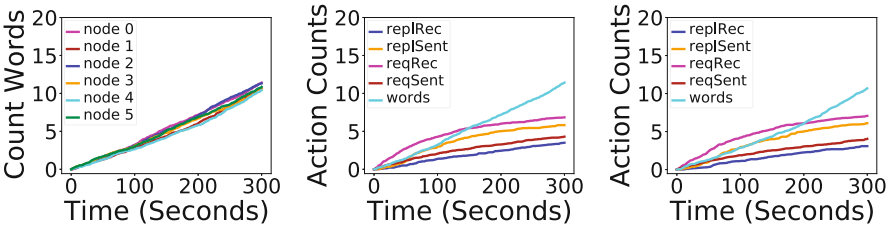
Parameter	Description
Network	Circ_6 (each of six players has degree 2)
Num. of initial letters n_ℓ	Four per player
Number of groups	One. Group $g = 1$ is for agents with degree ≤ 2
Number of clusters	There are four clusters within group $g = 1$
Number of different performances	For each cluster, there are three models of game player performance: worst, average, and best. These are the contrastive behaviors

A simulation consists of 50 instances. Each instance is a computation from time $t = 0$ to $t = 300$ s, consistent with conditions in experiments. That is, each

instance is a simulation of one NGAG. From experimental NGAG data, players do not take successive actions in less than one second. Thus, we set one time increment in a simulation to one second. All players are assigned $n_\ell = 4$ letters and model parameters based on group number (always group 1 in these experiments), cluster number, and performance level. Players request letters from their neighbors, reply to letter requests, form words, and think (idle). A simulation outputs all player actions at all times, similar to the data shown in Fig. 1. Average values and median and error bars in boxplots below are produced from all player data, at each time $t \in [0, 300]$ over all 50 instances. Since players are paid in the game in direct proportion to the number of words that they form, increasing numbers of actions (particularly in forming words) means increasing performance.

6.2 Simulation Results

Figure 5 contains data for group 1, cluster 3, and worst performance. The first plot provides average word history curves for each of the six players. The next two plots of the figure show time histories for all actions for players 0 and 1, respectively. These actions are replies received (replRec), replies sent (replSent), requests received (reqRec), requests sent (reqSent), and words formed (words). Requests sent and replies received are the lesser curves because they are both bounded by $n_\ell = 4$ letters for each player. Requests received and replies sent are greater curves because their numbers are bounded above by $k \cdot n_\ell = 2 \cdot 4 = 8$.

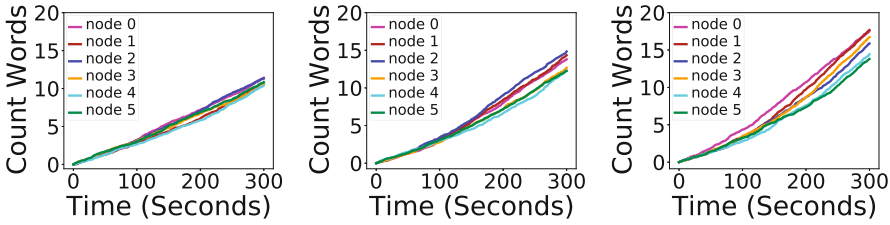


(a) word counts all agents (b) behavior agent 0 (c) behavior agent 1

Fig. 5. Results of anagram simulations with six players forming a Circ_6 graph. All players have behaviors assigned based on group 1, cluster 3, and worst performance. (a) word count histories for all six agents, (b) action histories for agent 0, and (c) action histories for agent 1.

Figure 6 contains word count histories for all six players, for cluster 3, and moving left to right, for worst, average, and best performance, respectively. (Figures 5a and 6a are the same plot.) It is clear that the numbers of words formed by players increase in going from worst to best behavior models by 25%.

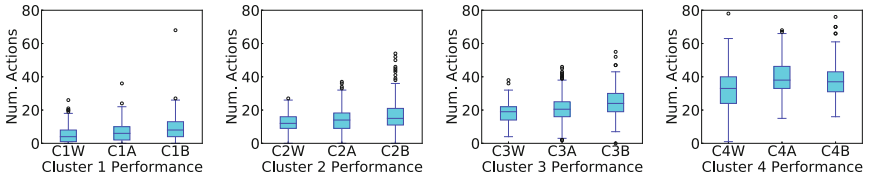
Figure 7 contains boxplots for each of the four clusters for group 1. Each box is for the total number of actions (which is the sum of the total number



(a) worst performance (b) average behavior (c) best behavior

Fig. 6. Results of anagram simulations with six players forming a Circ_6 graph. All players have behaviors assigned based on group 1, cluster 3. Plots are word count histories for all players in a simulation for: (a) worst, (b) average, and (c) best behavior.

formed words, requested letters, and replies to letter requests), on a per-player basis, across all six players in a simulation. Thus, each box is comprised of 300 data points (=6 players · 50 simulation instances). For each of the first three clusters, the counts of actions increases from worst (W) to average (A) to best (B) performance models. The numbers of actions, for a given performance value, also increases across the first three clusters, consistent with the experimental data. The fourth cluster is interesting and different. The worst, average, and best performance models do not generate monotonic results. The behavior appears to saturate with cluster 4, the cluster that produces the greatest numbers of player actions.



(a) Cluster 1 (b) Cluster-2 (c) Cluster 3 (d) Cluster 4

Fig. 7. Results across all clusters and all performance values, where boxplots are given for performance types worst (W), average (A), and best (B) behavior for each cluster. Boxes are per-player numbers of total actions in a simulated game, and represent the sum of form words, request letters, and reply to letter requests. The clusters are: (a) 1, (b) 2, (c) 3, and (d) 4. Labels on x-axis are "C", cluster number, and performance type.

7 Conclusion

We provide motivation and novelty of our work, along with contributions, in Sect. 1. A key aspect of the models, that enable explainability of results, is that

we map model parameters to player actions in a game. For our game, this is not straight-forward and hence may serve as a template of how this may be done in other game settings. We believe that our approach can be used with other human subject game data, including complicated experiments like ours with different actions types and the ability of players to repeat action types over time. Our current method uses asymptotic distributions to infer the uncertainty of model parameters, which may not be appropriate for some problems. Alternatively, a Bayesian approach for quantifying uncertainty can be useful.

Acknowledgment. We thank the anonymous reviewers for their helpful feedback. This work has been partially supported by NSF CRISP 2.0 (CMMI Grant 1916670) and NSF CISE Expeditions (CCF-1918770).

References

1. Alam, M., Deng, X., et al.: Sensitivity analysis of an enteric immunity simulator (ENISI)-based model of immune responses to helicobacter pylori infection. PLoS ONE **10**(9), e0136139 (2015)
2. Arendt, P.D., Apley, D.W., Chen, W.: Quantification of model uncertainty: calibration, model discrepancy, and identifiability. J. Mech. Design **134**(10), 1–12 (2012)
3. Aronson, E., Aronson, J.: The Social Animal, 12th edn. Worth Publishers, New York (2018)
4. Broere, J., Buskens, V., Stoof, H., et al.: An experimental study of network effects on coordination in asymmetric games. Sci. Rep. **9**, 1–9 (2019)
5. Cedeno, V., Hu, Z., et al.: Networked experiments and modeling for producing collective identity in a group of human subjects using an iterative abduction framework. Soc. Netw. Anal. Min. (SNAM) **10**, 11 (2020). https://doi.org/10.1007/s13278-019-0620-8
6. Charness, G., Feri, F., et al.: Experimental games on networks: underpinnings of behavior and equilibrium selection. Econometrica **82**(5), 1615–1670 (2014)
7. Duvivier, L., Cazabet, R., Robardet, C.: Edge based stochastic block model statistical inference. In: Benito, R.M., Cherifi, C., Cherifi, H., Moro, E., Rocha, L.M., Sales-Pardo, M. (eds.) COMPLEX NETWORKS 2020 2020. SCI, vol. 944, pp. 462–473. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-65351-4_37
8. Gerstein, D.R., Luce, R.D., et al.: The behavioral and social sciences: achievements and opportunities. Technical report, National Research Council (1988)
9. Hartigan, J.A., Wong, M.A.: Algorithm as 136: a k-means clustering algorithm. J. R. Stat. Soc. Ser. C **28**, 100–108 (1979)
10. Hu, Z., Deng, X., Kuhlman, C.J.: An uncertainty quantification approach for agent-based modeling of human behavior in networked anagram games. In: WSC (2021)
11. Judd, S., Kearns, M., Vorobeychik, Y.: Behavioral dynamics and influence in networked coloring and consensus. PNAS **107**, 14978–14982 (2010)
12. Kearns, M., Judd, S., Tan, J., Wortman, J.: Behavioral experiments on biased voting in networks. PNAS **106**, 1347–1352 (2009)
13. Mayzner, M.S., Tresselt, M.E.: Anagram solution times: a function of letter order and word frequency. J. Exp. Psychol. **56**(4), 376 (1958)
14. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books, New York (2018)

15. Ren, Y., Cedeno-Mieles, V., et al.: Generative modeling of human behavior and social interactions using abductive analysis. In: ASONAM, pp. 413–420 (2018)
16. Riley, M.E.: Evidence-based quantification of uncertainties induced via simulation-based modeling. Reliab. Eng. Syst. Saf. **133**, 79–86 (2015)
17. Shields, M.D., Au, S.K., Sudret, B.: Advances in simulation-based uncertainty quantification and reliability analysis. ASCE-ASME J. Risk Uncertainty Eng. Syst. Part A-Civ. Eng. **5**(4), 02019003-1–02019003-2 (2019)
18. Stones, C.R.: Self-determination and attribution of responsibility: another look. Psychol. Rep. **53**, 391–394 (1983)
19. Usui, T., Macleod, M.R., McCann, S.K., Senior, A.M., Nakagawa, S.: Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. PLoS Biol. **19**, e3001009 (2021)
20. Wuebben, P.L.: Experimental design, measurement, and human subjects: a neglected problem of control. Sociometry **31**(1), 89–101 (1968)
21. Yan, Y., Toriumi, F., Sugawara, T.: Influence of retweeting on the behaviors of social networking service users. In: Benito, R.M., Cherifi, C., Cherifi, H., Moro, E., Rocha, L.M., Sales-Pardo, M. (eds.) Complex Networks. SCI, vol. 943, pp. 671–682. Springer, Heidelberg (2020). https://doi.org/10.1007/978-3-030-65347-7_56