Cooperation of friends

Link to Github: **HERE**

Introduction

We ran a study about how well do friends cooperate and if this cooperation is better than two strangers cooperating.

To investigate this we designed an experiment in form of a cooperative board game. The game was based on Codenames Duet (Chvatil & Scot, 2017). The players were presented with 25 words and their task was for each of them to connect 9 words with a linkword and their partners had to guess which words their partner referred to. There were either two friends playing together or two strangers.

The hypotheses were the following:

H1: Pairs of friends overestimate their common knowledge, therefore make more mistakes.

H2: Pairs of friends will try to connect more words in one round. However in later rounds the number of connected words will decrease. This pattern should be present for strangers as well but to a lesser extent.

H3: Pairs of friends will take more rounds to complete the game.

Methods

All analysis was done in R 3.5.0 'Joy in Playing' (R Core Team, 2018). For constructing models R packages "rethinking" (McElreath, 2016) . For plotting we used ggplot2 (Wickham, 2009) and bayesplot (Gabry & Mahr, 2017).

Participants

10 groups of two participated in the experiment, overall 20 people, 10 males and 10 females. In the first part of the analysis one group was excluded due to not being able to finish the game, so in the whole analysis only 8 females were included. 8 Czechs, 3 Danes, 2 Hungarians, 2 Slovaks, 1 Chinese, 1 Estonian, 1 German, 1 Finn and one person from the USA participated. The excluded group included 2 Danes. Their mean age is 21.37, ranging from 19 to 26.

Experiment

The game was coded in oTree (Chen, Schonger and Wickens, 2016). For better orientation the words were also presented physically on the board in front of the participants. The participants were seated at a desk opposite to each other and each had a computer. The instructions were presented on the computers. Then one of the players enters into the computer how many words to connect in the round, which words these are and a link word. The link word and number of words to guess is then displayed to the other player. In the following round the guessing player is linking the words. The game was played until all words were successfully linked and guessed.

Data

The collected data were split into two datasets. The aggregated dataset has one row per game/group. The variables in this dataset are presented in Table 1. The long dataset consists of one row per round. The variables in this dataset are presented in Table 2.

The variables ID, a unique number for each group, and condition, group of friends or strangers, is present in both datasets.

Aggregated Dataset	
mistakes	Number of words guessed wrong in the whole game
n_rounds	How many rounds did the game last

Table 1: Variables in the aggregated dataset

Long Dataset	
linked_n	Number of linked words in the round
performance	% of linked_ guessed correctly
trial	Number of the current round

Table 2: Variables in the long dataset

Aggregated data analysis

In the aggregated dataset there are 2 variables of our interest: mistakes and n_rounds.

Models

We built 6 bayesian binomial regression models as reported below. For each variable: mistakes and n_rounds; two models were built, without random effects and with random intercept for each group (ID). Then 2 models were built with both variables as predictors, again one without random effects and one with random intercept for each group. All models were estimated with MCMC. We ran 2 chains with 10 000 iterations.

Mistakes: Condition ~ Binomial (1, p) logit(p) ~ α + ßM * mistakes α ~ Normal (0, 0.05), ßM ~ Normal (0, 0.2)	**Mistakes_random:** Condition ~ Binomial (1, p) logit(p) ~ α[ID] + ßM * mistakes α[ID] ~ Normal (0, 0.05), ßM ~ Normal (0, 0.2)
n_rounds: Condition ~ Binomial (1, p) logit(p) ~ α + ßR * n_rounds α ~ Normal (0, 0.05), ßR ~ Normal (0, 0.2)	**n_rounds_random:** Condition ~ Binomial (1, p) logit(p) ~ α[ID] + ßR * n_rounds α[ID] ~ Normal (0, 0.05), ßR ~ Normal (0, 0.2)
Full: Condition ~ Binomial (1, p) logit(p) ~ α + ßM * mistakes + ßR * n_rounds α ~ Normal (0, 0.05), ßM ~ Normal (0, 0.2), ßR ~ Normal (0, 0.2)	**Full_random**: Condition ~ Binomial (1, p) logit(p) ~ α[ID] + ßM * mistakes + ßR * n_rounds α[ID] ~ Normal (0, 0.05), ßM ~ Normal (0, 0.2), ßR ~ Normal (0, 0.2)

Priors

The same prior was used for all intercepts: normal distribution with mean 0 and standard deviation of 0.05. We also used a common prior for all ß parameters: normal distribution with mean 0 and standard deviation of 0.1. The priors are shown in Figure 1 for better visualization of the effects of these priors.

The prior for the intercepts reflects the chance level. So without any predictors the probability of classifying the group correctly is 0.5 and because the model has binomial likelihood the mean of the prior has to be logit of 0.5 which equals to 0. The

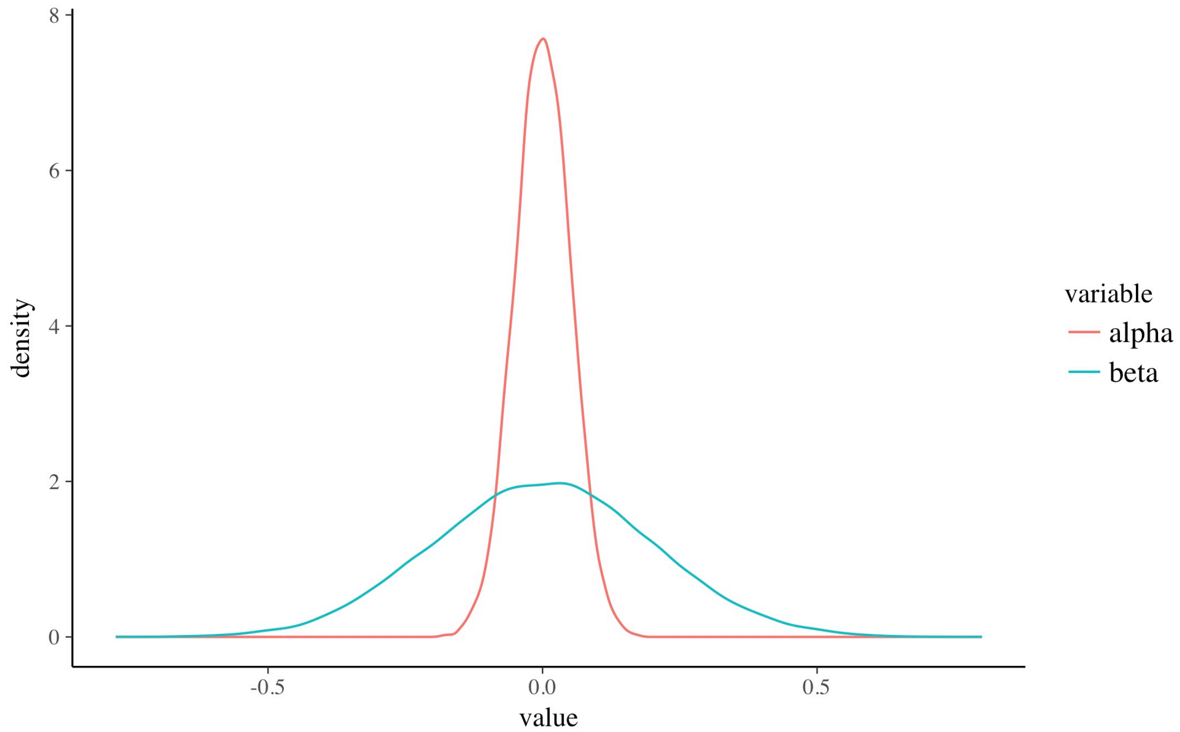

Figure 1: Priors for intercepts and ß parameters

prior is this narrow because the estimates have to be very close to chance level because there is no information about the group in its intercept.

The prior for ß parameters is centered at 0 because we do not have any prior information about this kind of effects as our experimental design has not been used previously as far as we know. Also we want to be skeptical of any effects so the mean at 0 is ideal for that purpose. Similarly we express our skepticism with the standard deviation of 0.2. This prior would allow the effect size to be estimated up to 0.4. If we take a logistic of 0.4 and add the maximum value of an intercept a change of the predictor variable by 1 unit it would increase the absolute odds of being friends by 9.7%, which is a reasonable effect size to expect with a strong evidence in the data.

Long data analysis

In the long dataset there are 3 variables of our interest: linked_n, performance and trial.

Models

We built 10 bayesian binomial regression models as reported below. For each variable: linked_n and performance; 2 models were built: with variable as predictor and with interaction of the variable with trial. Then 2 models were built that were a combination of the previous models. So one with the 2 variables as predictors and one that included both variables and their interaction with trial.

For each of the models described above a version with random intercept for group was built.

All models were estimated with MCMC. We ran 2 chains with 10 000 iterations.

Linked_n: Condition ~ Binomial(1 , p) , logit(p) ~ α + ßL * linked_n , α ~ Normal (0, 0.05), ßL ~ Normal(0, 0.2)	**Linked_n_random:** Linked_n with α[ID]
Linked_n_round: Condition ~ Binomial(1 , p), logit(p) ~ α +ßL * linked_n + ßT * trial +ßLT * linked_n * trial, α ~ Normal (0, 0.05), ßL ~ Normal (0, 0.2), ßT ~ Normal (0, 0.2), ßLT ~ Normal (0, 0.2)	**Linked_n_round_random:** Linked_n_round with α[ID]
Performance: Condition ~ Binomial (1 , p) , logit(p) ~ α + ßP * performance, α ~ Normal (0, 0.05), ßP ~ Normal (0, 0.2)	**Performance_random:** Performance with α[ID]
Performance_round: Condition ~ Binomial (1 , p) , logit(p) ~ α + ßP * performance + ßT * trial + ßPT * performance * trial , α ~ Normal (0, 0.05), ßP ~ Normal (0, 0.2), ßT ~ Normal (0, 0.2), ßPT ~ Normal (0, 0.2)	**Performance_round_random:** Performance_round with α[ID]

Priors

The same priors that were described for the aggregated data analysis above were used for these models too.

Results

Model selection and model quality for aggregated data analysis

To choose models to report the effects from we compared the models with widely applicable information criterion (WAIC). The results are reported in Table 3 where the models are ordered from the best model to the worst one.

	WAIC	pWAIC	dWAIC	weight	SE	dSE
m_mistakes_random	11.088	0.567	0	0.25	2.782	NA
m_mistakes	11.143	0.593	0.055	0.243	2.807	0.038
m_all_random	11.547	1.397	0.46	0.199	2.95	2.008
m_all	11.674	1.455	0.586	0.187	2.986	2.054
m_rounds_random	13.875	0.958	2.788	0.062	1.543	1.734
m_rounds	13.972	1.001	2.885	0.059	1.505	1.75

Table 3: Model comparison of models from aggregated data analysis

From the model comparison we decided to report only two. The m_mistakes_random, for interpreting the effect of mistakes, and the m_all_random to interpret the effect of rounds.

For these two models we performed also a model quality check by constructing a predictive posterior check (PPC) and plotted the predicted proportions of friend groups and the proportion in the training data. We also checked the Markov chains and they were all stationary and mixing well. Finally we checked the correlation between the parameters of the models but did not find any alarming values.

The PPC plots for m_mistakes_random and m_all_random are reported in Figures 2 and 3 respectively. Both models are performing reasonably well but of course not perfectly.

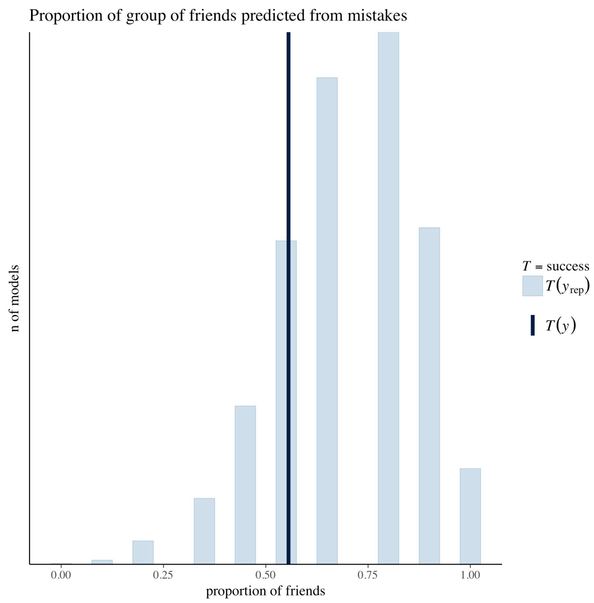

Figure 2: PPC plot of m_mistakes_random

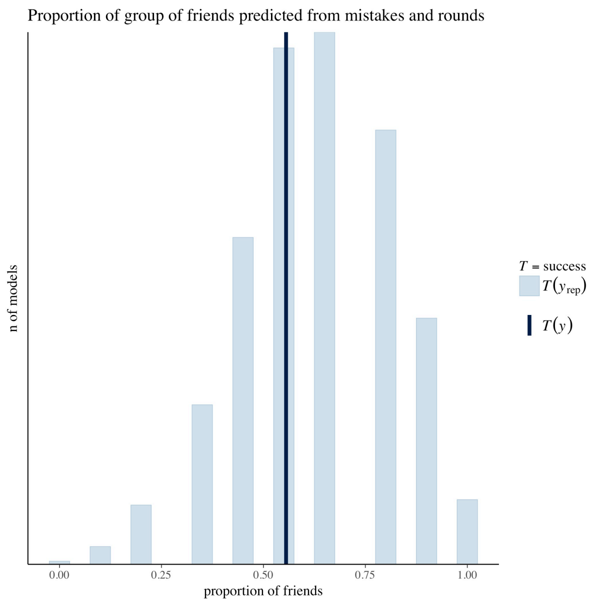

Figure 3: PPC plot of m_all_random

Results of aggregated data

For better understanding of the estimates all parameters are plotted in Figure 4 and Figure 5 for m_mistakes_random and m_all_random respectively. The circle is the mean and the lines represent the 89% credibility intervals.

The estimates of m_mistakes_random are reported in Table 4.

	Mean	StdDev	lower 0.89	upper 0.89
α[1]	0	0.049	-0.075	0.081
α[2]	-0.002	0.05	-0.085	0.075
α[3]	-0.002	0.05	-0.082	0.078
α[4]	-0.001	0.049	-0.078	0.08
α[5]	0.001	0.051	-0.081	0.082
α[6]	-0.002	0.051	-0.083	0.078
α[7]	0.001	0.05	-0.08	0.08
α[8]	0	0.05	-0.078	0.082
α[9]	0.001	0.05	-0.077	0.08
ßM	0.12	0.078	0.001	0.242

Table 4: Estimates of model: condition ~ mistakes

The estimates of m_all_random are reported in Table 5.

	Mean	StdDev	lower 0.89	upper 0.89
α[1]	0	0.05	-0.078	0.085
α[2]	-0.001	0.05	-0.083	0.076
α[3]	-0.001	0.05	-0.08	0.078
α[4]	-0.001	0.051	-0.081	0.08
α[5]	0	0.05	-0.081	0.076
α[6]	-0.002	0.05	-0.083	0.074
α[7]	0	0.051	-0.08	0.082
α[8]	0.001	0.05	-0.082	0.078
α[9]	0.002	0.05	-0.079	0.079
ßR	-0.077	0.094	-0.228	0.066
ßM	0.18	0.106	0.004	0.339

Table 5: Estimates of model: condition ~ n_rounds + mistakes

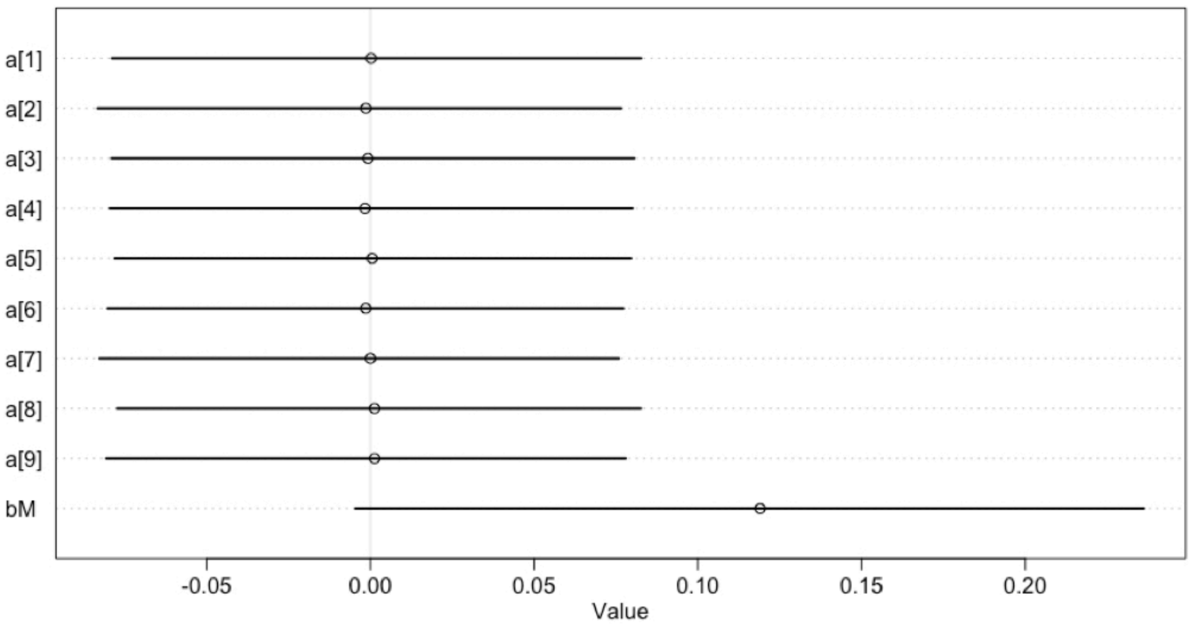

Figure 4: Plotted estimates of m_mistakes_random

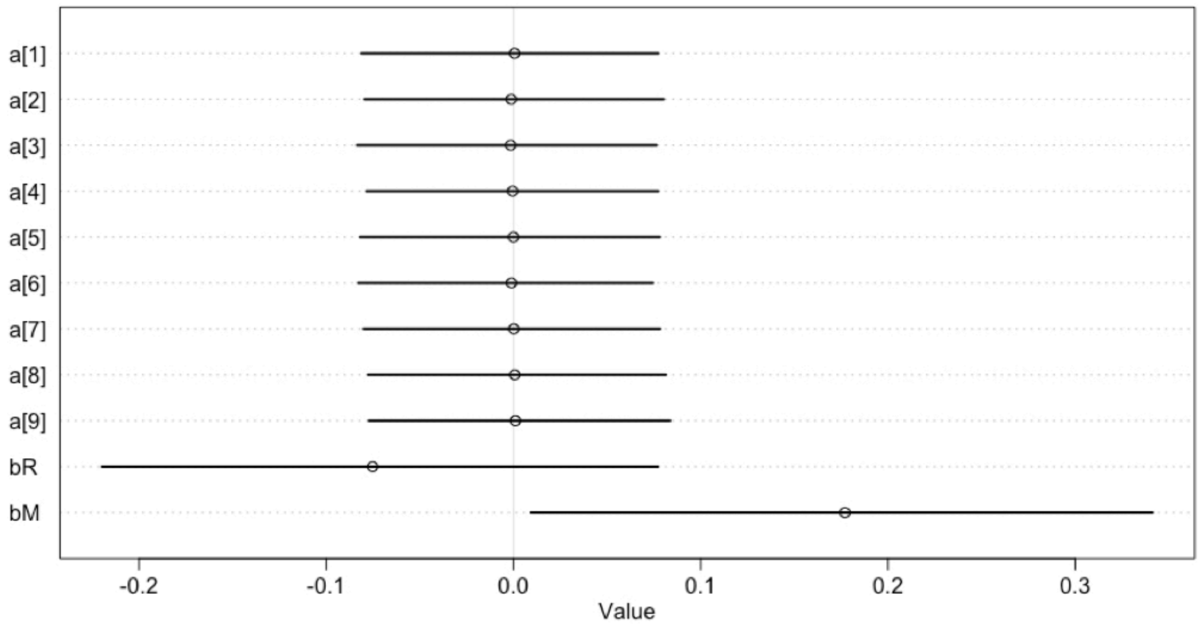

Figure 5: Plotted estimates of m_all_random

Model selection and model quality for long data analysis

To choose models to report the effects from we compared the models with widely applicable information criterion (WAIC). The results are reported in Table 6 where the models are ordered from the best model to the worst one.

	WAIC	pWAIC	dWAIC	weight	SE	dSE
m_full	130.879	4.077	0	0.323	7.82	NA
m_linked_n_round_random	131.349	2.336	0.47	0.255	5.736	5.144
m_performance_round_random	132.757	3.144	1.878	0.126	7.011	2.547
m_linked_n_random	133.61	0.788	2.732	0.082	4.06	6.868
m_performance_round	133.764	3.136	2.885	0.076	7.088	2.528
m_performance_linked_n	134.684	1.522	3.805	0.048	3.919	6.782

m_linked_n	134.768	0.764	3.889	0.046	4.151	6.866
m_linked_n_round	135.455	2.157	4.576	0.033	5.981	5.138
m_performance_random	138.765	1.143	7.886	0.006	3.06	7.381
m_performance	139.716	1.047	8.837	0.004	3.101	7.36

Table 6: Model comparison of models from long data

From the model comparison we decided to report only the best model. The m_full, to interpret the effects of linked_n and performance.

For these this model we performed also a model quality check by constructing a predictive posterior check (PPC) and plotted the predicted proportions of friend groups and the proportion in the training data. We also checked the Markov chains and they were all stationary and mixing well. Finally we checked the correlation between the parameters of the models but did not find any alarming values.

The PPC plot is reported in Figure 6. The model is performing reasonably well but of course not perfectly.

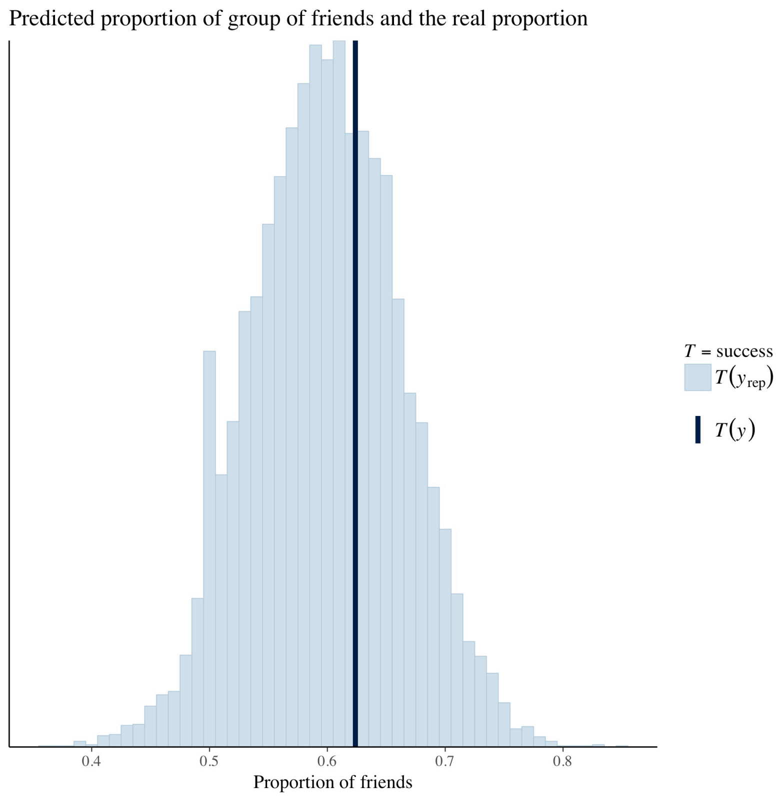

Figure 6: PPC plot of m_full

Results of long data

The estimates of m_full are reported in Table 7. For better understanding of the estimates all parameters are plotted in Figure 7. The circle is the mean and the lines represent the 89% credibility intervals.

	Mean	StdDev	lower 0.89	upper 0.89
α[1]	0.009	0.049	-0.07	0.085
α[2]	-0.014	0.05	-0.093	0.067
α[3]	-0.014	0.049	-0.089	0.067
α[4]	-0.009	0.049	-0.087	0.069
α[5]	0.01	0.05	-0.072	0.089
α[6]	-0.016	0.05	-0.093	0.066
α[7]	0.011	0.051	-0.068	0.095
α[8]	0.004	0.05	-0.077	0.082
α[9]	0.011	0.05	-0.067	0.09
α[10]	0.014	0.05	-0.063	0.095
ßP	-0.014	0.007	-0.025	-0.003
ßL	0.241	0.163	-0.018	0.503
ßT	0.107	0.149	-0.118	0.355
ßPT	0	0.002	-0.002	0.003
ßLT	0.034	0.062	-0.06	0.137

Table 7: Estimates of m_full

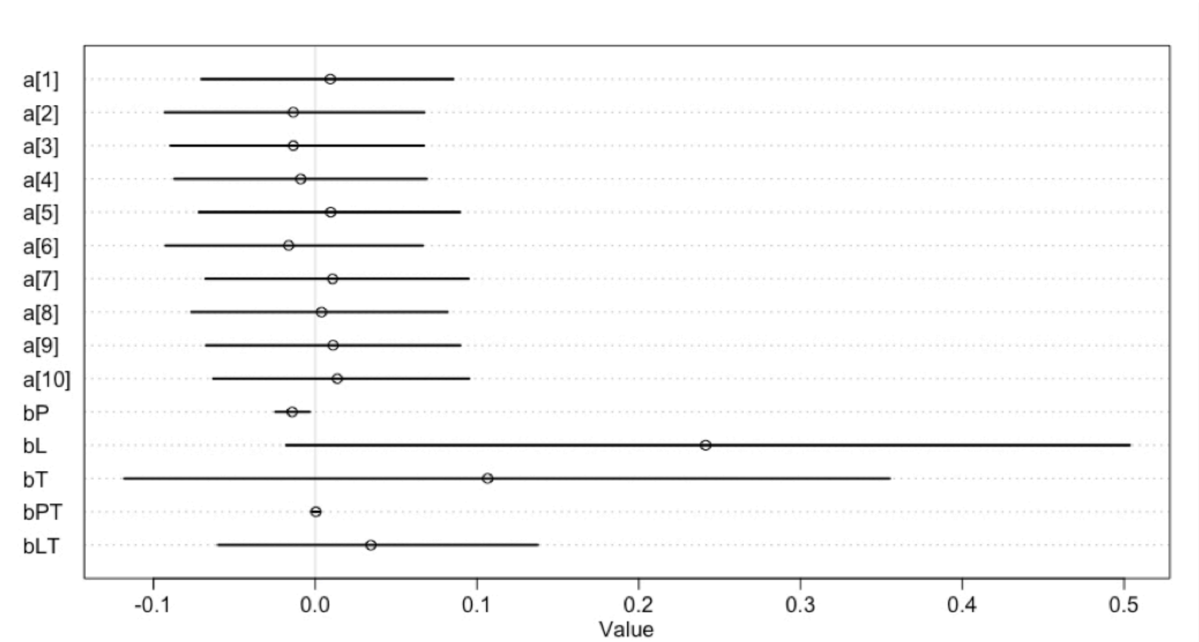

Figure 7: Plotted estimates of m_full

Discussion

Interpretation of the aggregated dataset

Results show a positive effect of number of mistakes, 0.12. Which means that with number of mistakes, the probability of being friends increases, which supports H1. This pattern is plotted in Figure 8. Even though the credibility intervals are in the positive range, the lower 89% interval lies very close to zero (0.001). For illustration we include a plot (Figure 9) of how the posterior moves away from the prior. In the plot we can see that the posterior crosses zero. However the probability that the effect goes in opposite direction is only 0.039. It might be a good idea to collect more data to resolve this uncertainty.

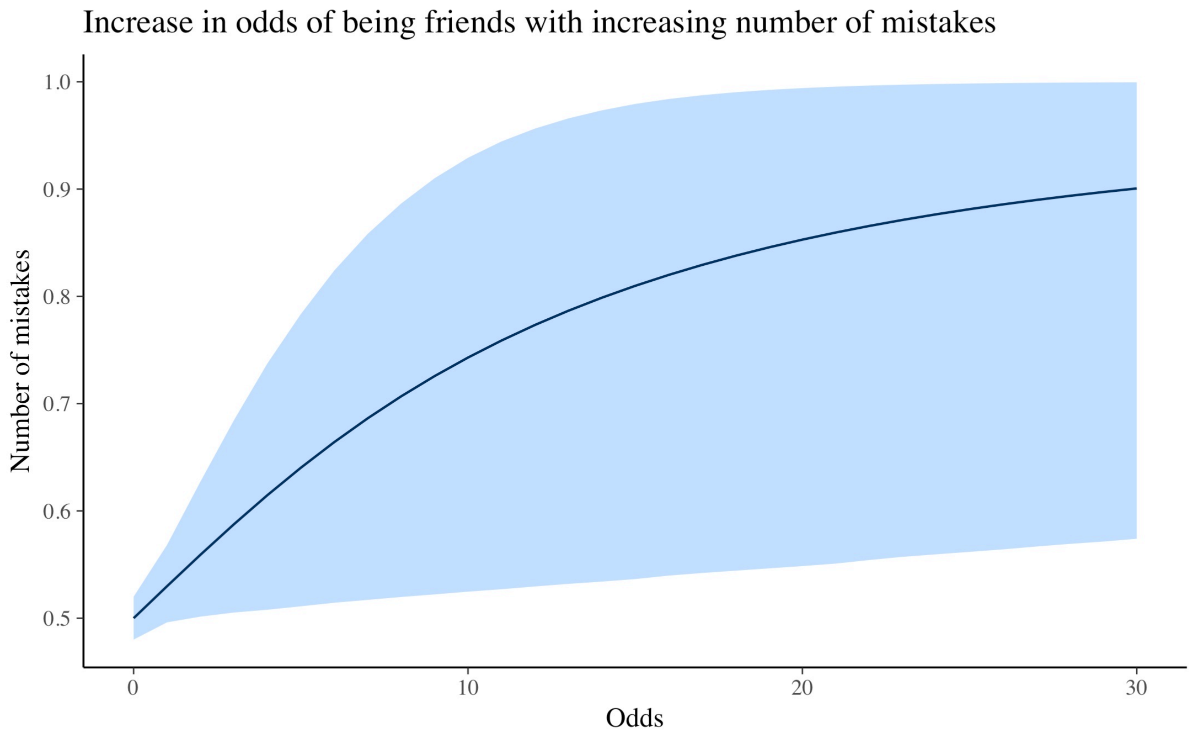

Figure 8: Absolute effect of number of mistakes (summed with intercept)

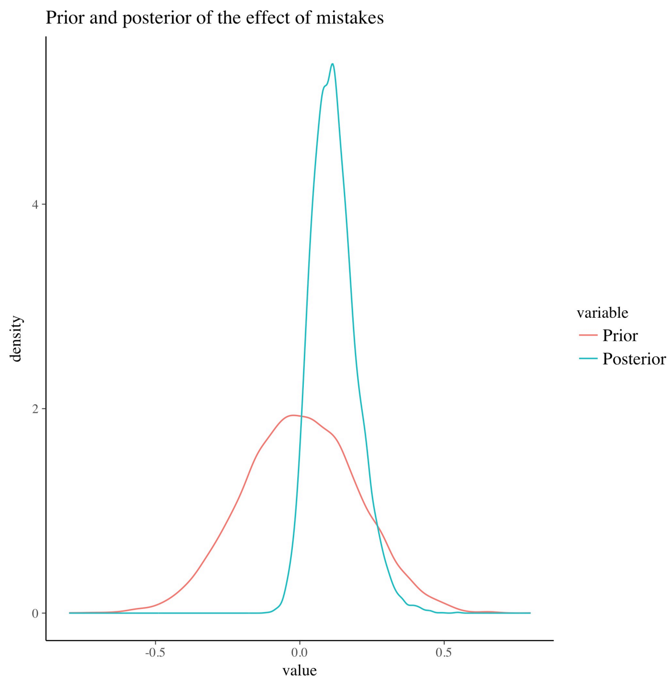

Figure 9: Prior and posterior distributions of effect of number of mistakes

The results shows a mean negative effect of the number of rounds, -0.077. However the credibility intervals cross zero substantially which makes the effect very uncertain. It would be reasonable to collect more data to resolve the uncertainty. This effect is plotted in Figure 10. From the plot we can see that because the estimate is very uncertain the upper edge of the shaded area shows a positive effect but the lower boundary shows a negative effect. Since the mean of the posterior shows a weak negative effect we must conclude that it contradicts our hypothesis 2.

We also include a plot (Figure 11) to show how the posterior moved away from prior. In this case we can see that although the posterior moves away from the prior it remains very broad which represents the uncertainty of the effect described above.

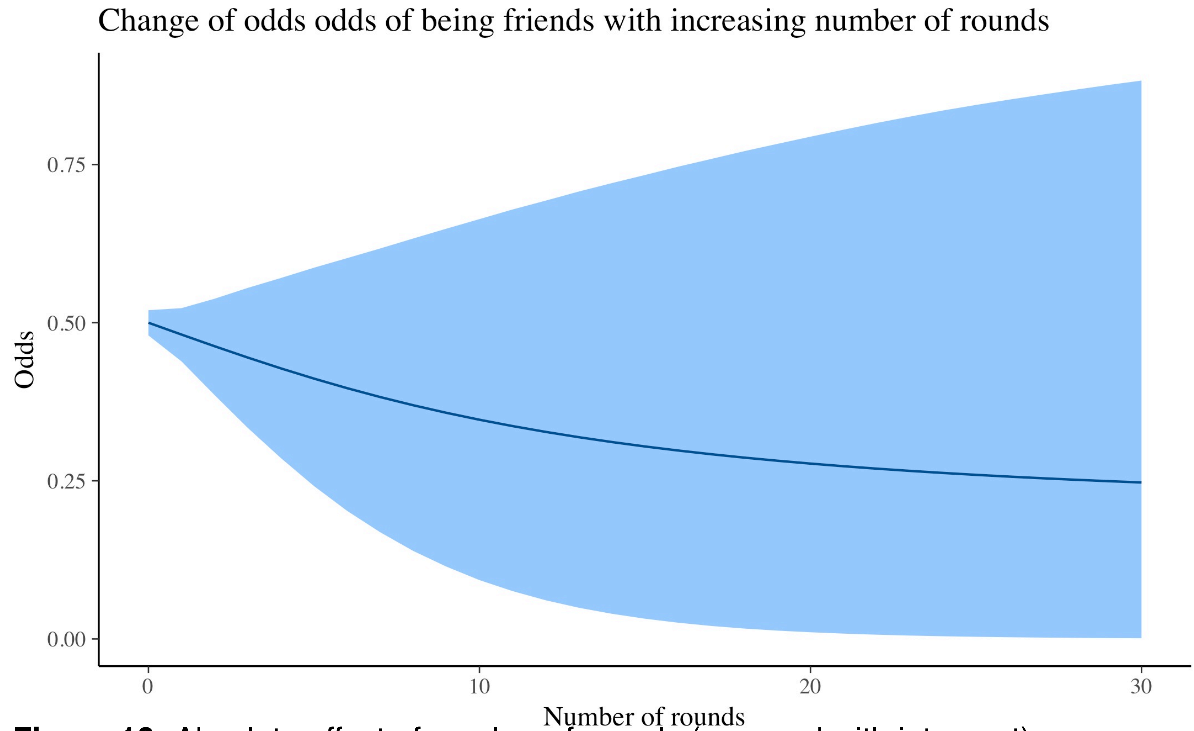

Figure 10: Absolute effect of number of rounds (summed with intercept)

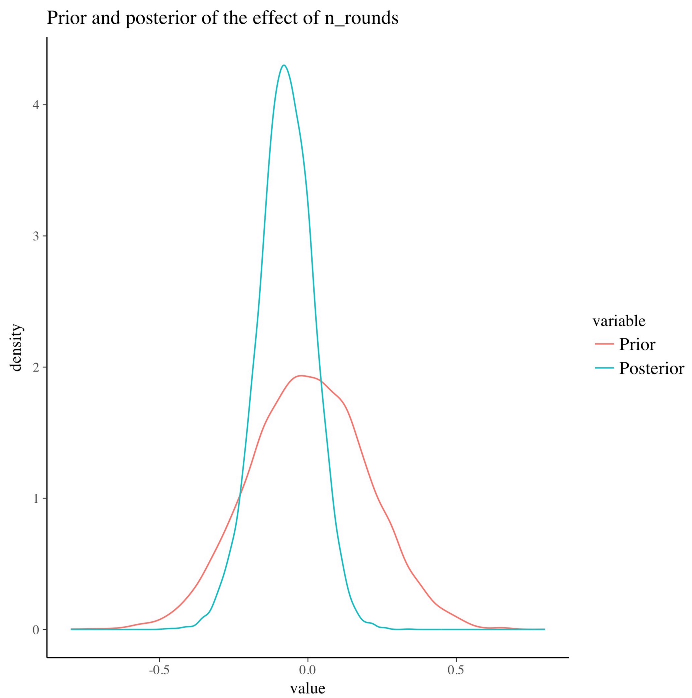

Figure 11: Prior and posterior distributions of effect of number of rounds.

Interpretation of the long dataset

The results show no effect or a very unreliable one for the interaction of performance with trial, interaction of linked_n with trial and trial. Their credibility intervals overlap with zero to such an extent that no real effect can be expected even if we were to collect additional data.

The effect for performance is -0.014, which means that with every percent of increased performance, the probability of being friends decreases. This is in agreement with Hypothesis 1. The effect is plotted in Figure 12. The credibility intervals do not overlap with zero, but the upper bound lies very close to zero.

To visualize how much new information the data brought we plotted the prior and posterior (Figure 13). We can see that the posterior moved away from prior a lot and is also very narrow compared to the prior.

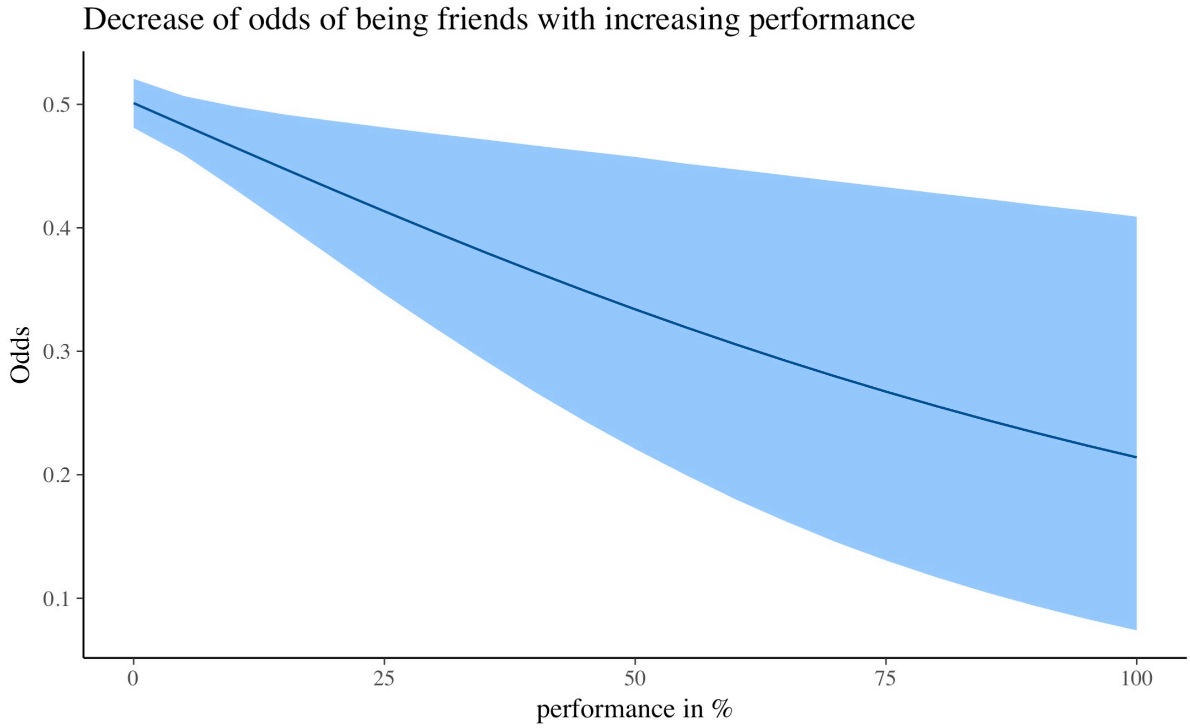

Figure 12: Absolute effect of performance (summed with intercept)

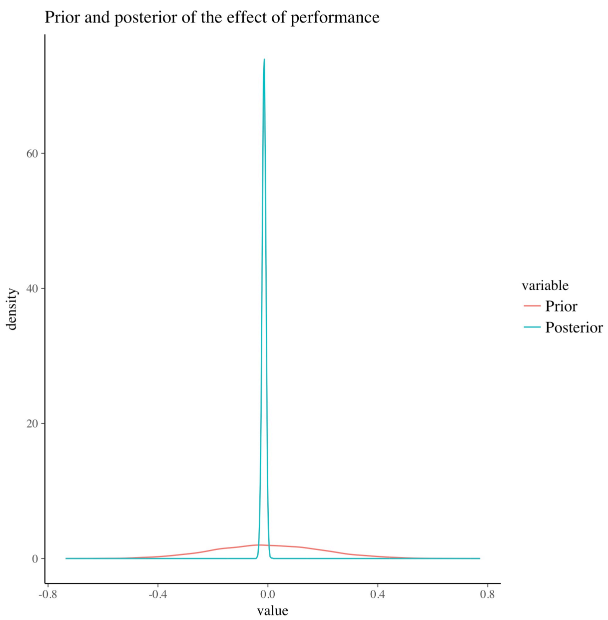

Figure 13: Prior and posterior distributions of effect of performance

The effect of linked_n is 0.241, meaning that based on the predictions of the model, with every increasing number of linked words, the probability of being friends increases, which supports Hypothesis 2. Even though the credibility intervals overlap with zero, which makes the effect unreliable, the lower bound lies close to zero (-0.018). The probability of the effect pointing only in the positive direction is 0.96. To visualize this very small overlap we plotted the prior and posterior of the effect (Figure 14). The plot also shows that the data brought new information since the posterior moved away from the prior quite a lot. We also plotted the meaning of the effect in Figure 15 where we can see that the effects goes in the positive direction for the most part but in the lower parts of the shaded area the effect turns into a negative one.

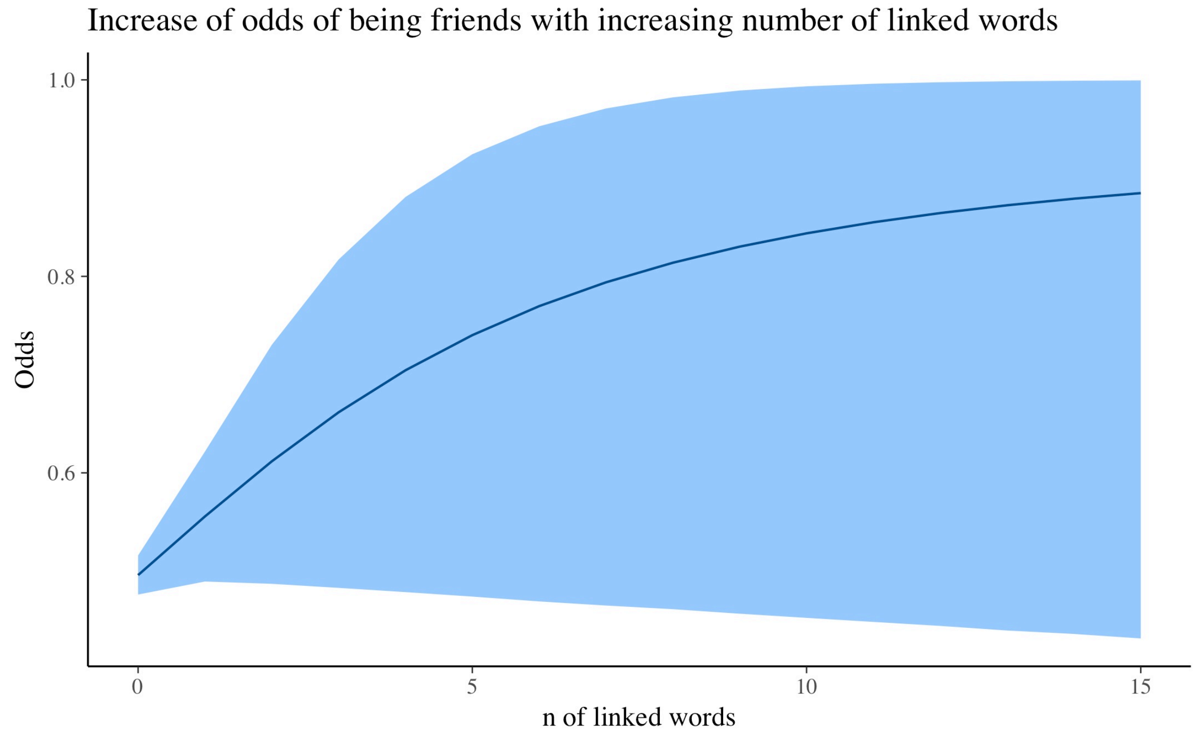

Figure 15: Absolute effect of linked words (summed with intercept)

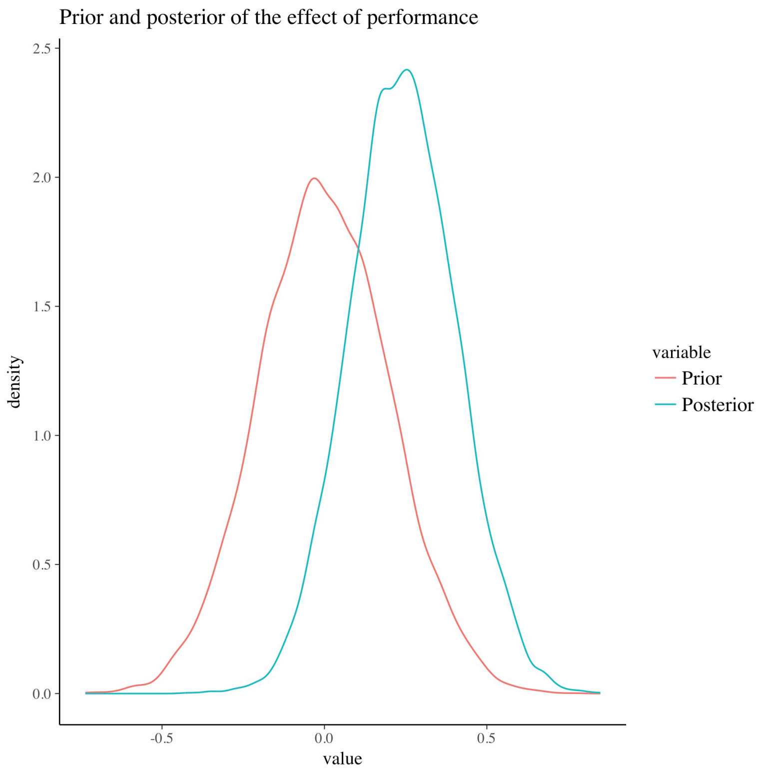

Figure 14: Prior and posterior distributions of effect of linked words

Conclusion

The analysis of the experiment seems to support our hypothesis number 1 that friends do indeed make more mistakes when collaborating. For hypotheses 2 and 3 there is not enough evidence in the data. However some of the effects point in the predicted direction but remain uncertain and it would be reasonable to collect more data to resolve this conflict.

References

Chen, D.L., Schonger, M., Wickens, C., 2016. oTree - An open-source platform for laboratory, online and field experiments. Journal of Behavioral and Experimental Finance, vol 9: 88-97

Chvatil, V. & Eaton, S., (2017). Codenames Duet. Kladno: Czech Games Edition. Board Game.

Gabry, J. & Mahr, T., (2017). bayesplot: Plotting for Bayesian Models. R package version 1.4.0. https://CRAN.R-project.org/package=bayesplot

McElreath, R. (2016). rethinking: Statistical Rethinking book package. R package version 1.59.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wickham H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.