Dear Authors,

Thank you for submitting this interesting Stage 1 protocol. The RR builds upon Stiller et al. (2003) and aims to investigate whether the structure of social networks in all of Shakespeare's plays mimics real-world social networks. Before providing my comments, I'd like to disclose that I'm a psychology researcher focused on the methodology of behavioral research. I've experience applying the network approach to psychological phenomena, which is not necessarily the same as social network analysis, Additionally, I possess an average non-native-English-speaker level of knowledge about Shakespeare's plays. The review will thus be based on my overall impressions of this RR and questions I had while reading it.

Open science practices

I'd like to start by praising the authors for their level of transparency and adherence to open science practices. The publicly available materials could indeed be a valuable resource for other researchers pursuing similar research questions.

Abstract and introduction

The abstract is succinct and outlines the study's rationale. Even though it's Stage 1, the abstract would benefit from specifying the research questions.

In the introduction, the authors reference the upper bound of the size of social groups. Can the authors provide a more detailed explanation of this limit, especially in the context of today's society (e.g., usage of social networks) and its heterogeneity across cultures?

Can the authors elaborate on why Shakespeare's plays (besides the ten investigated by Stiller et al.) were chosen for this analysis. What are the characteristics that make these plays unique/suitable for such a study? Furthermore, can the authors summarize (methodological) differences between the present study and the study by Stiller et al.? From a layperson's perspective, I also wonder about the rationale of examining the research questions on all 37 plays. Do the authors assume that this mirroring of social interactions is universal across the plays and can the heterogeneity in the genres, contents, and complexity of character interactions be disregarded? While I very much appreciate the authors' efforts to create methodologically and technically rigorous workflow, I'd like these substantial questions clarified.

A minor note: Although I mostly agree with authors' definitions of reproducibility, replicability, robustness, and generalizability, I'd suggest adding a reference (e.g., https://doi.org/10.1146/annurev-psych-020821-114157 or NASEM's report https://www.nationalacademies.org/our-work/reproducibility-and-replicability-in-science).

Reproducibility, generalizability, and robustness testing

The authors state that there are three major pathways (i.e., the choice of plays to be included in the analysis, how the play is segmented into time slices, and the criterion for tie-formation) that could determine the results. I fully agree. These researcher degrees of freedom cover the selection of plays and, in essence, data pre-processing. Based on my experience with network analysis of psychological phenomena, the resulting parameters are often sensitive to analytical choices (e.g., estimator selection, setting tuning parameter/s, etc.). If these analysis-related choices could determine the results (i.e., no single optimal network construction algorithm exists), would it be possible (and make sense) to incorporate this into the code and multiverse the results?

Non-registered analyses

When comparing the number of speaking characters, the authors propose to use paired Wilcoxon tests. I'd suggest using Welch's t-tests instead, set SESOI, and conduct equivalence testing (or a Bayesian analysis) in addition to NHST. I've never seen Weber fraction used outside cognitive psychology research – it looks interesting and promising.

Registered analyses

Overall, as far as my expertise goes, I find the registered analyses technically sound. For registered analysis no. 2, I suggest reporting not only the frequency of three- or four-character configuration but extending it to a distribution of all character configurations. Perhaps a formal test (e.g., chi-square) can be a useful addition to determine if the observed frequency of three or four characters per time slice differ from what one would expect by chance. This can also be useful to examine if the distributions of characters per time slice differ across the analytic variants. For registered analysis no 3, I'd suggest the authors take a look at the NetworkComparisonTest (van Borkulo et al., 2017) package for R. The permutation-based approach introduced in the package can be helpful to answer the pursued research question in a more precise manner.

Final remarks

From a technical and methodological perspective, the present RR is rigorous and can be impactful in a way that it has the potential to greatly help other researchers who pursue similar research questions. From a substantial viewpoint, I think that the RR would benefit from several clarifications. These would greatly help non-specialized readership to get a better understanding of the paper and its rationale.

Best,

Matúš Adamkovič