Child Mind® Institute

QUESTIONNAIRE RESPONSE CORRELATIONS TO IMPROVE EFFICIENCY: PRELIMINARY EVIDENCE FROM THE HEALTHY BRAIN NETWORK

Jon Clucas[1], Jake Son[1], Anirudh Krishnakumar[1,2], Michael P. Milham[3,4], Arno Klein[1]
jon.clucas@childmind.org

[1]MATTER Lab, Child Mind Institute [2]Centre de Recherches Interdisciplinaires, IIFR, Paris, France [3]Center for the Developing Brain, Child Mind Institute [4]Nathan Kline Institute

INTRODUCTION

The Healthy Brain Network, a multimodal pediatric psychiatric biobank [1], includes dozens of questionnaires [3]. In labs and in practice, questionnaires can be burdensome to participants and to administrators. While a response to any individual question is informative, the informative value of each subsequent question will vary. With hundreds of (eventually ten thousand) individuals' responses to many overlapping questionnaires, we are well-positioned to measure the relative information of pairs of questions. Knowing these relative values can afford more efficient questionnaires, allowing administrators to automatically prioritize the most informative questions.

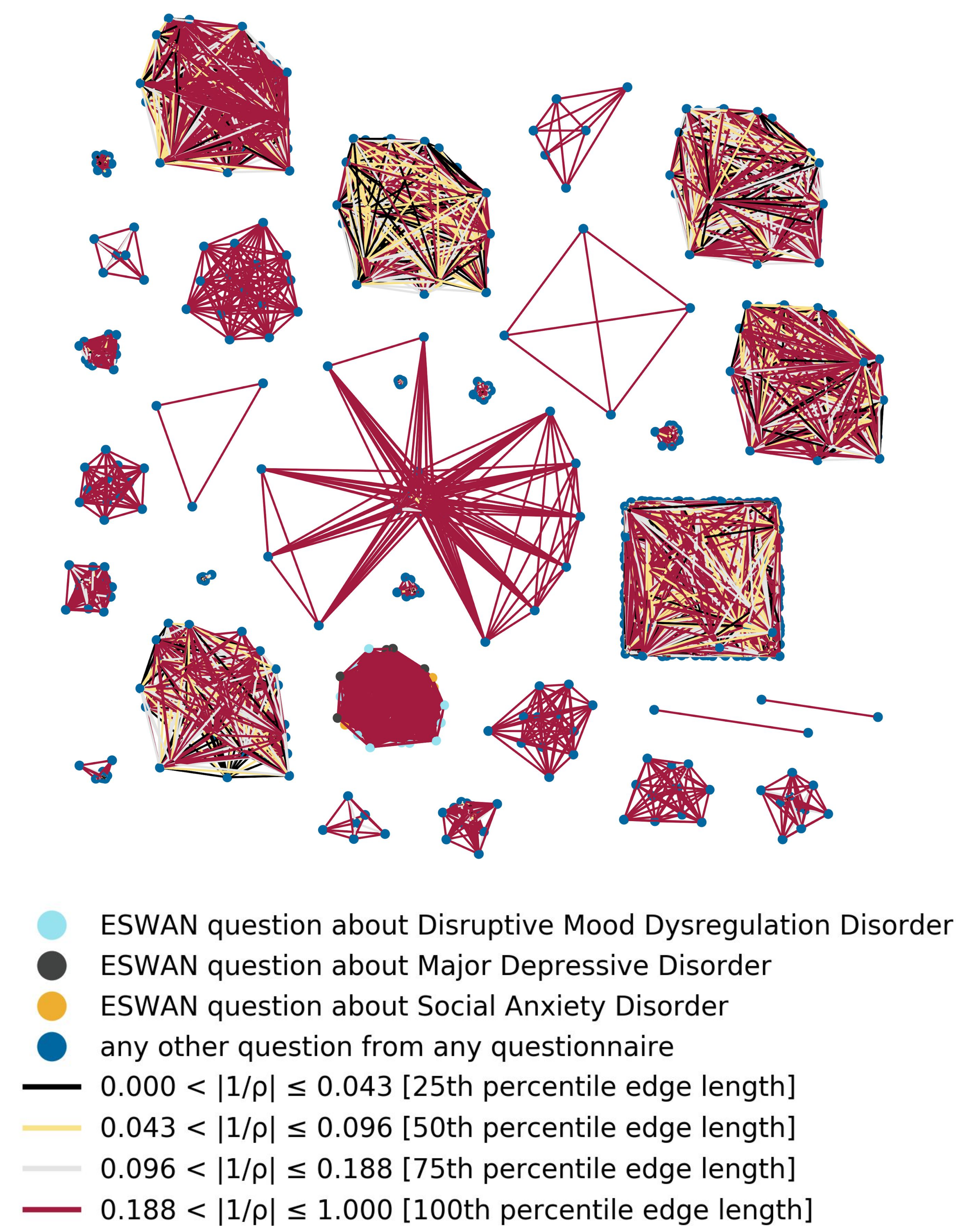
METHODS

We analyzed questionnaire responses from the first two Healthy Brain Network releases (n=881 subjects, 79 questionnaires, 2,630 questions, available at Link 1. For each pair of question response vectors, we calculated and inverted Pearson's ρ, dropping any pairs for which abs(ρ)>0. Figure 1 shows each question as a node connected by edges of length 1/ρ. The code used to generate the figures is available in a Jupyter notebook at Link 2.

RESULTS

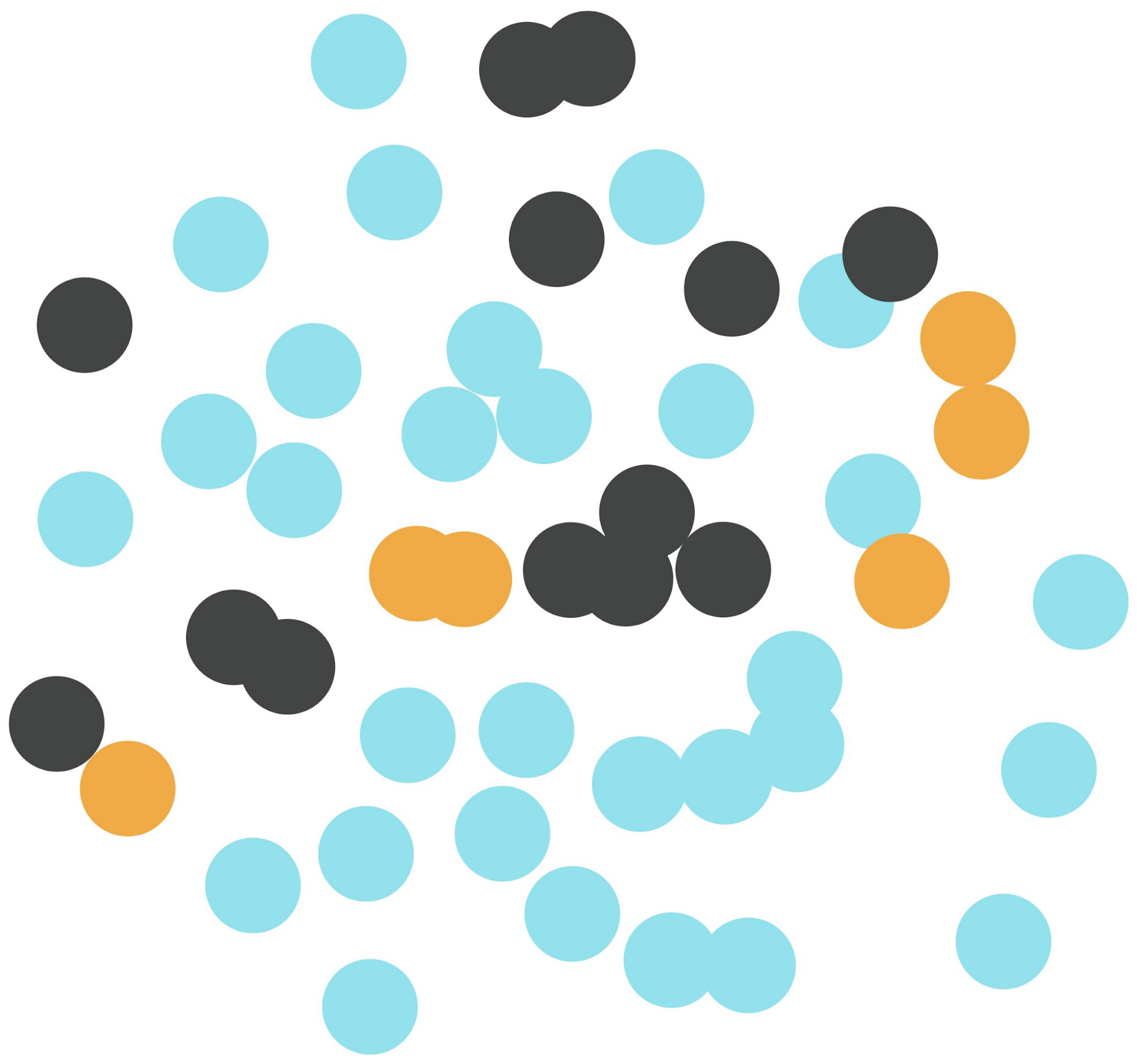
Our initial visual exploration indicated 30 groupings of correlated responses (based on inverted Pearson's ρ; see Figure 1), often linking questions within a single questionnaire. Two of these clusters contain only two questions each (the Fagerstrom Test for Nicotine Dependence [5] questions "Are you currently a smoker?" and "Have you been a smoker within the past two years?" clustered only with one another; the Goldman-Fristoe Test of Articulation [4] sounds-in-sentences completion clustered only with accuracy from the same test). One cluster contains 1,876 questions. The second-largest cluster contains 66 questions (excluding the 1,876-question cluster: mean=26, standard deviation=19.5). Most of the clusters contain questions from only one questionnaire each, indicating a sensitivity of this comparison method to artifacts of questionnaire administration. Figure 2 shows a cluster containing only questions from the Extended Strengths and Weaknesses Assessment of Normal Behavior questionnaire [2], but questions about three disorders: Disruptive Mood Dysregulation Disorder, Major Depressive Disorder and Social Anxiety Disorder.

Figure 1: 30 clusters of questions with correlated responses.



● ESWAN question about Disruptive Mood Dysregulation Disorder
● ESWAN question about Major Depressive Disorder
● ESWAN question about Social Anxiety Disorder
● any other question from any questionnaire
— 0.000 < |1/ρ| ≤ 0.043 [25th percentile edge length]
— 0.043 < |1/ρ| ≤ 0.096 [50th percentile edge length]
— 0.096 < |1/ρ| ≤ 0.188 [75th percentile edge length]
— 0.188 < |1/ρ| ≤ 1.000 [100th percentile edge length]

Figure 2: One of the 30 clusters, enlarged, with edges hidden.



FUTURE WORK

By employing a variety of methods, we can simultaneously assess the appropriateness of each method and the degree of correspondence between these methods. We are pursuing analyses with random forests [7][8], randomer forests [9] and probabilistic metamodeling [6], to estimate the most informative questions for predicting ADHD subtype consensus diagnosis and Autism Spectrum Disorder consensus diagnosis. The code for these analyses is available online at Link 3.

REFERENCES

1. Lindsay M. Alexander, Jasmine Escalera, Lei Ai, et al. 2017a. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data* 4 (Dec. 2017), 170181. DOI:10.1038/sdata.2017.181
2. Lindsay M. Alexander, Giovanni Salum, James M. Swanson, and Michael P. Milham. 2017b. Balancing Strengths and Weaknesses in Dimensional Psychiatry. *bioRxiv* (Oct. 2017), 207019. DOI:10.1101/207019
3. Child Mind Institute. 2016. Complete List of Assessments. (2016). http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/assessments/master-list.html
4. Ronald Goldman and Macalyne Fristoe. 2015. *Goldman-Fristoe Test of Articulation 3*. American Guidance Service, Inc., Circle Pines, MN. https://www.pearsonclinical.com/language/products/100001202/goldman-fristoe-test-of-articulation-3-gfta-3.html
5. T. F. Heatherton, L. T. Kozlowski, R. C. Frecker, and K. O. Fagerström. 1991. The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction* 86, 9 (Sept. 1991), 1119–1127.
6. Vikash Mansinghka. 2016. The MIT Probabilistic Computing Project. (Sept. 2016). http://probcomp.csail.mit.edu/
7. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (Oct. 2011), 2825−2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
8. scikit-learn developers. 2017. Random Forests. In *scikit-learn User Guide*. 1.11.2.1. http://scikit-learn.org/stable/modules/ensemble.html#random-forests
9. Tyler M. Tomita, Mauro Maggioni, and Joshua T. Vogelstein. 2015. Randomer Forests. *arXiv:1506.03410 [cs, stat]* (June 2015). http://arxiv.org/abs/1506.03410

LINKS

1. http://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network/sharing_phenotypic.html
2. https://github.com/ChildMindInstitute/questionnaire-correlations/releases/tag/v0.1.0
3. https://github.com/ChildMindInstitute/questionnaire-diagnosis