Interpersonal Coordination in Perception and Memory in an Online Experiment: Using Networked Crowdsourcing for Experiments on Human Social Interaction

Alexandra Paxton
paxton.alexandra@gmail.com
Institute of Cognitive and Brain Sciences
Berkeley Institute for Data Science
University of California, Berkeley

Thomas J. H. Morgan
thomas.j.h.morgan@asu.edu
School of Human Evolution and Social Change
Arizona State University

Jordan W. Suchow
suchow@berkeley.edu
Social Science Matrix
University of California, Berkeley

Thomas L. Griffiths
tom_griffiths@berkeley.edu
Department of Psychology
University of California, Berkeley

Abstract

Recent advances in crowdsourcing have helped many cognitive scientists reach out beyond traditional undergraduate subject pools to run a range of experimental paradigms with a wider audience. To date, however, many of these opportunities for online experiments on crowdsourcing platforms have been closed to researchers interested in capturing the dynamics of human social interaction. We argue that an important next step for increasing the adoption and utility of online experiments will lie in using *networked crowdsourcing*—moving beyond providing individual participants separate tasks to support more complex interactive or interdependent configurations. Networked crowdsourcing allows researchers to capture real-time and transmission-chain interaction between participants to study social cognition and behavior. Here, we use networked crowdsourcing to move the study of real-time interpersonal coordination from the lab and onto Amazon Mechanical Turk, examining how people grow similar over time in their perception and memory.

Keywords: interpersonal coordination; networked crowdsourcing; human communication; online experiments; social interaction

Introduction

Over the past several decades, a sizable body of literature has documented the ways in which interacting individuals become more similar over time as a result of their interaction. Intuitively, people often recognize this: We recognize that we'll start to synchronize our steps with a friend while walking, and we'll comment on how longtime couples have similar gestures and manners of speech.

Interpersonal Coordination

We here focus on the phenomenon of *interpersonal coordination*, an interdisciplinary research area that focuses on the ways in which individuals affect one another over time as a result of their interaction (also known as interactive alignment, interpersonal synchrony, mimicry, and more; see Paxton, Dale, & Richardson, 2016). This area is increasingly marked by principles of dynamical systems theory, conceptualizing interaction as a complex adaptive system (Riley, Richardson, Shockley, & Ramenzoni, 2011). A fundamental principle of this is that the emergent behavior—in this case, interpersonal coordination—is not static but changes over time.

Crowdsourcing for Social Experiments

Recent developments in how data can be collected and analyzed are transforming cognitive science. This is reflected in an increased interest in big data and naturally occurring datasets (Goldstone & Lupyan, 2016), such as social media activity and video game logs, which hold the promise of capturing behavior in the wild and providing a testing ground for key scientific theories (Paxton & Griffiths, 2017). While these data can provide a window into observational data about human behavior at a massive scale, technological advances are quickly expandinsg to accommodate new *experimental* paradigms as well.

Crowdsourcing platforms like Amazon Mechanical Turk (`http://www.mturk.com`) have been extensively used as a means to collect data with relatively simple but robust experimental paradigms, like surveys (Buhrmester, Kwang, & Gosling, 2011) and mouse-tracking (Freeman, Dale, & Farmer, 2011). At first, work in this domain required researchers to use established survey creation tools, which were quick to do but constrained experimental designs, or to program bespoke experiments, which is more open ended but far more time-consuming. More recently, cognitive scientists have worked to create solutions to support the efficient creation of a wider range of experiments (e.g., Gureckis et al., 2016). As the community around online psychology experiments has grown, it has done so with the intent to broaden its reach (especially to researchers with less programming experience) and to continue to provide more powerful experimental tools.

To date, many of these experiments have focused on individuals, making it difficult for researchers to study social processes through online experiments. We believe that the next step in online experimentation, then, is to move to *networked crowdsourcing*, creating interactive or interdependent experimental paradigms that construct interacting networks of people to understand social processes and phenomena. In doing so, networked crowdsourcing can provide researchers interested in social behavior the opportunity to expand their experimental capabilities beyond the lab while not compromising on the richness and complexity of true interactive contexts.

Networked Crowdsourcing

Online experimental platforms have been capable of serving individual paradigms for the past several years, but what marks this idea as different from these paradigms is our focus on *networks*. Instead of handling participants as individuals, networked crowsourcing allows researchers to connect participants with one another—whether through direct, real-time interaction or through sequential transmission chains—to directly manipulate social dynamics online in the same way that is possible within the lab. As such, networked crowdsourcing is uniquely positioned to support experimental research into human social behavior at scale.

A recurring concern for using online experiments lies in its participant population. Like all convenience samples, there can be questions about the degree to which the participants reflect the broader population dynamics—including the use of undergraduate students at Western universities as participants in return for course credit, who often do not reflect global demographics (Henrich, Heine, & Norenzayan, 2010). Considerations of sampling and population representativeness are vital for any study, and researchers should carefully consider their sampling choices at the outset of their work. For those interested in using online participants (especially from Amazon Mechanical Turk), recent surveys suggest that U.S.-based MTurk workers are more diverse in a variety of ways than typical college students but not entirely reflective of the general U.S. population (e.g., Buhrmester et al., 2011; Paolacci & Chandler, 2014).

The Present Study

The current study focuses on understanding how interacting individuals become entrained in perception and memory over time, becoming a sort of "line estimation system"—just as two people in an lab experiment become a "tangram recognition system" (Dale, Richardson, & Kirkham, 2011). To do that, we isolate the dyad to a minimally interactive context, allowing participants to engage with one another solely by communicating line lengths

Method

All research activities were completed in compliance with oversight from Committee for the Protection of Human Subjects at the University of California, Berkeley.

Participants

Participants (n = 148) were individually recruited from Amazon Mechanical Turk to participate as dyads (n = 74). Participants were paired with one another according to the order in which they began the experiment. All participants were over 18 years of age and fluent English speakers (self-reported); participation was restricted to only recruit from participants located within the U.S. with a 95% HIT approval rate.[1]

The experiment lasted an average of 11.69 minutes (range: 7.98—21.34 minutes). All participants were paid $1.33 as base pay for finishing the experiment and earned a bonus of up to $2 for the entire experiment based on mean accuracy over all trials (mean = $1.80; range: $0.00—$1.95). Participants were not informed about the amount of performance-based bonus that they earned during the experiment.

Procedure

All data collection procedures were run on Amazon Mechanical Turk (`http://mturk.com`) using the experiment platform Dallinger (v3.4.1; `http://github.com/dallinger/Dallinger`). Code for the experiment is available on GitHub (`http://github.com/thomasmorgan/joint-estimation-game`), and the resulting experiment data are available on the OSF repository for the project (`https://osf.io/8fu7x/`).

Each participant was individually recruited on Amazon Mechanical Turk to play a "Line Estimation Memory Game" (advertisement: "Test your memory skills!"). Upon completing informed consent, participants were told that they would be playing a game in which they would be required to remember and re-create line lengths. Participants were informed that they would be complete their training trials individually and would then begin playing with a partner. Participants were given no information about their partner other than the guess that their partner made; no information about the partner's identity was shared.

In each trial, participants were shown 3 red lines, each of a different length (see figure; **NB**: add figure), and were asked to remember all three of them.[2] The 3 stimulus lines were displayed for 2 seconds then removed, providing participants with a blank screen for 0.5 seconds. Participants were then told which line to re-create (#1, #2, or #3) and were then given 1 second to submit their guess at how long the target line had been. To do so, participants were given a blank box and used their cursor to fill in the box with a blue line. All lines were presented within bounded boxes of 500 pixels (wide) by 25 pixels (high).

During training, participants were then shown the correct length of the target line (as a grey bar above their own guess) for 2 seconds. This was accompanied by a message telling the participant that they had guessed correctly ("Your guess was correct!") or incorrectly ("Your guess was incorrect") or that they had not submitted a guess within the 1-second time limit ("You didn't respond in time").

During testing, participants' stimulus viewing, waiting, and recreation times remained the same as during testing, but they no longer received information about whether their guess was correct or incorrect. Instead, after both participants

[1] A measure of MTurk worker quality, capturing how often their work is rejected by a requester. A 95% HIT approval rate means that only 5% of all of their submitted HITs have been rejected.

[2] A pilot version of this study showed that participants performed at ceiling when given only 1 line to remember and recreate. The additional 2 lines were added to strictly increase the memory load, as opposed to adding difficulty in other ways (e.g., creating a moving stimulus).

had submitted their first guess, participants were shown their guess (in blue) above their partner's guess (in green). Both participants were then asked whether they wanted to change their own guess or to keep the guess they had submitted. If either participant in the dyad indicated that they wanted to change their guess, that participant was then allowed to change their guess (again with a 1-second time limit) *while* still being able to view their partner's guess. Participants who chose to keep their previously submitted guess was informed that their partner chose to submit a new guess and waited for the other participant to finish. At that point, participants were again allowed to change or keep their guess. This process continued until both participants chose to keep their guess.

Participants were informed that their final accuracy would only be calculated for their final guess. However, because they had no means to communicate with their partner about whether each would be accepting or changing their guesses, each participant could not have known whether their decision to keep the guess would have been their final guess for the trial.

For clarity, we will refer to each new stimulus set as a *trial* and to each submitted line length estimate within each trial as a *guess*. This means that some participants may have submitted multiple guesses per trial. The last submitted estimate—the one by which trial-level accuracy is calculated—will be referred to as the *final guess*.

All dyads completed 10 training trials (alone) and 15 test trials (with their partner). All training and test stimuli were randomly generated for each dyad, but both participants within the dyad were given the same stimuli. After participating, each individual participant was asked to complete a series of questionnaires about the game on a series of 1-10 Likert-style scales, including the perceived difficulty of the task, how engaged they were in the task, and questions about their own and their parner's cooperativeness and trustworthiness.

Analyses

Similarity To measure how participants' perceptual and memory systems became more similar over time, we calculated the cross-correlation coefficient of participants' guess errors across trials (Paxton & Dale, 2013), within a window of +/-5 guesses. Although cross-correlation produces information about leading and following behavior, we have no *a priori* expectations about which of the two participants would emerge as a leader (given they have no information about their partner nor any assigned roles), so our first-pass analyses ignore any directionality by averaging across the each incremental lags (i.e., leading/following in both participants' directions).

Accuracy Accuracy was measured as a ratio relative to the total possible error on a given target stimulus trial. That is, rather than taking a given guess's error relative to the total line length, error was calculated as the maximum *possible* error. For example, if the target stimulus was 60 units long, the maximum possible error for that trial would be 60, and the participant's error would be calculated relative to that maximum possible error.

Results

All analyses were performed in R (R Core Team, 2016). Linear mixed-effects models were performed using the `lme4` package (Bates, Mchler, Bolker, & Walker, 2015) using the maximal random slope structure for each random intercept to achieve model convergence (Barr, Levy, Scheepers, & Tily, 2013). All main and interaction terms were centered and standardized prior to entry in the model, allowing the model estimates to be interpretable as effect sizes (Keith, 2005).

Discussion

Future Directions

Injecting social dynamics in experimentally and/or as much as is desired

Conclusion

Acknowledgements

Thanks go to the Dallinger development team for their assistance in debugging the experiment.

This work was funded in part by the **DALLINGER GRANT INFO GOES HERE**.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, *63*(3), 255–278.

Bates, D., Mchler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?, *6*, 3–5. http://doi.org/10.2307/41613414

Dale, R., Richardson, D. C., & Kirkham, N. K. (2011). *How two people become a tangram recognition system*. Aarhus University, Aarhus, Denmark.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59. http://doi.org/10.3389/fpsyg.2011.00059

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*, 548–568. http://doi.org/10.1111/tops.12212

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). PsiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Henrich, J., Heine, S., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Keith, T. (2005). *Multiple regression and beyond.* Boston, MA: Pearson.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. http://doi.org/10.1177/0963721414531598

Paxton, A., & Dale, R. (2013). Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology*, *66*(11). http://doi.org/10.1080/17470218.2013.853089

Paxton, A., & Griffiths, T. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638. http://doi.org/10.3758/s13428-017-0874-x

Paxton, A., Dale, R., & Richardson, D. (2016). Social coordination of verbal and nonverbal behaviors. In P. Passos, K. Davids, & C. J. Yi (Eds.), *Interpersonal coordination and performance in social systems* (pp. 259–273). Abington, Oxon; New York, NY: Routledge.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Riley, M. A., Richardson, M., Shockley, K., & Ramenzoni, V. C. (2011). Interpersonal synergies. *Frontiers in Psychology*, *2*, 38.